University of Ljubljana

Faculty of Computer and Information Science

Aleks Jakulin

Machine Learning Based on Attribute Interactions

PhD Dissertation

Advisor: Acad. Prof. Dr. Ivan Bratko

Sežana, June 13, 2005

Univerza v Ljubljani

Fakulteta za računalništvo in informatiko

Aleks Jakulin

Strojno učenje na osnovi interakcij med atributi

DOKTORSKA DISERTACIJA

Mentor: Akad. Prof. Dr. Ivan Bratko

Sežana, 13. junij 2005

Abstract

Two attributes A and B are said to interact when it helps to observe the attribute values of both attributes together. This is an example of a 2-way interaction. In general, a group of attributes $\mathcal{X}$ is involved in a k-way interaction when we cannot reconstruct their relationship merely with ℓ-way interactions, $\ell < k$. These two definitions formalize the notion of an interaction in a nutshell.

An additional notion is the one of context. We interpret context as just another attribute. There are two ways in which we can consider context. Context can be something that specifies our focus: we may examine interactions only in a given context, only for the instances that are in the context. Alternatively, context can be something that we are interested in: if we seek to predict weather, only the interactions involving the weather will be interesting to us. This is especially relevant for classification: we only want to examine the interactions involving the labelled class attribute and other attributes (unless there are missing or uncertain attribute values).

But the definitions are not complete. We need to specify the model that assumes the interaction: how to we represent the pattern of co-appearance of several attributes? We also need to specify a model that does not assume the interaction: how do we reconstruct the pattern of co-appearance of several attributes without actually observing them all simultaneously? We need to specify a loss function that measures how good a particular model is, with respect to another model or with respect to the data. We need an algorithm that builds both models from the data. Finally, we need the data in order to assess whether it supports the hypothesis of interaction.

The present work shows that mutual information, information gain, correlation, attribute importance, association and many other concepts, are all merely special cases of the above principle. Furthermore, the analysis of interactions generalizes the notions of analysis of variance, variable clustering, structure learning of Bayesian networks, and several other problems. There is an intriguing history of reinvention in the area of information theory on the topic of interactions.

In our work, we focus on models founded on probability theory, and employ entropy and Kullback-Leibler divergence as our loss functions. Generally, whether an interaction exists or not, and to what extent, depends on what kind of models we are working with. The concept of McGill's interaction information in information theory, for example, is based upon Kullback-Leibler divergence as the loss function, and non-normalized Kirkwood superposition approximation models. Pearson's correlation coefficient is based on the proportion of explained standard deviation as the loss function, and on the multivariate Gaussian model. Most applications of mutual information are based on Kullback-Leibler divergence and the multinomial model.

When there is a limited amount of data, it becomes unclear what model can be used to interpret it. Even if we fix the family of models, we remain uncertain about what would be the best choice of a model in the family. In all, uncertainty pervades the choice of the model. The underlying idea of Bayesian

statistics is that the uncertainty about the model is to be handled in the same was as the uncertainty about the correct prediction in nondeterministic domains. The uncertainty, however, implies that we know neither if is an interaction with complete certainty, nor how important is the interaction.

We propose a Bayesian approach to performing significance tests: an interaction is significant if it is very unlikely that a model assuming the interaction would suffer a greater loss than a model not assuming it, even if the interaction truly exists, among all the foreseeable posterior models. We also propose Bayesian confidence intervals to assess the probability distribution of the expected loss of assuming that an interaction does not exist. We compare significance tests based on permutations, bootstrapping, cross-validation, Bayesian statistics and asymptotic theory, and find that they often disagree. It is important, therefore, to understand the assumptions that underlie the tests.

Interactions are a natural way of understanding the regularities in the data. We propose interaction analysis, a methodology for analyzing the data. It has a long history, but our novel contribution is a series of diagrams that illustrate the discovered interactions in data. The diagrams include information graphs, interaction graphs and dendrograms. We use interactions to identify concept drift and ignorability of missing data. We use interactions to cluster attribute values and build taxonomies automatically.

When we say that there is an interaction, we still need to explain what it looks like. Generally, the interaction can be explained by inferring a higher-order construction. For that purpose, we provide visualizations for several models that allow for interactions. We also provide a probabilistic account of rule inference: a rule can be interpreted as a constructed attribute. We also describe interactions involving individual attribute values with other attributes: this can help us break complex attributes down into simpler components. We also provide an approach to handling the curse of dimensionality: we dynamically maintain a structure of attributes as individual attributes are entering our model one by one.

We conclude this work by presenting two practical algorithms: an efficient heuristic for selecting attributes within the naïve Bayesian classifier, and a complete approach to prediction with interaction models, the Kikuchi-Bayes model. Kikuchi-Bayes combines Bayesian model averaging, a parsimonious prior, and search for interactions that determine the model. Kikuchi-Bayes outperforms most popular machine learning methods, such as classification trees, logistic regression, the naïve Bayesian classifier, and sometimes even the support vector machines. However, Kikuchi-Bayes models are highly interpretable and can be easily visualized as interaction graphs.

Keywords

- machine learning, data mining, information visualization
- interaction, dependence, dependency, correlation
- independence, independence assumption, factorization
- information theory, entropy, mutual information, maximum entropy
- Bayesian statistics, significance testing, confidence intervals

Acknowledgements

I wish to express gratitude to my advisor, Ivan Bratko, who gave me the complete freedom to explore, yet kept reminding me to focus on well-defined problems appreciated in the research community. The study of interaction information developed through adapting the HINT constructive induction approach by B. Zupan in a more statistical way, inspired by the work of J. Demšar. My work was supported by a junior researcher grant from the Slovenian Research Agency.

I particularly appreciate advice and help from Wray Buntine, who patiently helped me understand a number of issues in statistics. Some of the work that is reported in this dissertation was performed jointly with my colleagues. The discrete latent variable analysis of the US Senate was performed by W. Buntine, I report his findings, and I developed the visualization. I. Rish originally told me about the Kikuchi approximation, and we collaborated in developing the Kikuchi-Bayes method. B. Zupan's efforts were crucial for the work on Harris hip score: I report from his account of physician's opinions of the identified interactions. The idea of expressing the Venn diagram in the form of a line is due to G. Leban. I am grateful to S. Salthe who helped me understand numerous issues in the philosophy of science.

Several people provided feedback that importantly shaped my priorities. V. Batagelj pointed me to Rajski's work. T. Minka's questions helped shape my understanding of interaction information. N. Peek, V. Rupnik and M. Pukl provided the incentive for dealing with continuous variables, along with much feedback.

I have been helped by several colleagues in my search for literature. S. Ravett Brown provided the original Psychometrika publication. Te Sun Han kindly sent me his papers by mail. The advice of H. Harpending, his population genetics data, and his pointers to literature had a large impact on my work. I am grateful to A. Pajala, A. Gelman and B. Lawson for getting me up to date with political science research: all of them were very kind and helpful. The discussions with Y. Bulatov about the CRF and information geometry are appreciated, along with his suggestions about the presentation.

Having good empirical data at hand was important. D. Smrke provided the hip score data. M. Žnidaršič provided the ecological data that led towards the development of the contour plot visualization of 3-way interactions. I collaborated with U. Brefeld, J. Dobša and T. Scheffer when analyzing the interactions in text. J. McCrone helped me refine and understand the organic developmental perspective.

The community of colleagues and friends always provided support. I am most grateful to M. Grobelnik and D. Mladenić, who kindly supported me in the final stages of the work, and provided me with the excellent opportunity to remain in the Slovenian AI research community. The work of J. Demšar, G. Leban, T. Curk, M. Možina and B. Zupan on the Orange framework helped me save much time programming. D. Vladušič, J. Žabkar, P. Juvan, A. Sadikov, D. Šuc, M. Robnik-Šikonja and M. Kukar were also kind and friendly colleagues. Without the encouragement given by C. Faloutsos, the US Senate research would not happen. I am grateful to I. Pitchford, P. Marijuán, R. Ulrich and Y. Bar-Yam for providing unusual learning environments: the newsgroups and mailing lists.

But most of all, I must express gratitude to my family for the patience and support in spite of the stress and toil over these two years: Tina, Angela, Boris and Darja.

Sežana, April 2005. Aleks Jakulin

CONTENTS

CHAPTER 1

Introduction

1.1 Interactions

Interaction is a fundamental concept we often use, but rarely specify with precision. There are two senses to the word. The first, ontic sense is 'influence or a mutual or reciprocal action in the world'. For example, when a ball and a wall meet, they interact. When a seller and a buyer exchange money and goods, they interact. When two people hold a conversation, they interact. When a user types on a computer, the computer and the user interact. By using the term 'interaction' rather than 'cause' we stress that the causal direction is ambiguous or bidirectional. This sense of *ontic interaction* will not be discussed in this work.

The second, epistemic sense of interaction implies some sort of association, correlation, entanglement. This is an aspect of the mind. For example, the height and the weight of a man are involved in an interaction: the greater the height, the greater the weight. There is an interaction between smoking and lung cancer: the incidence of lung cancer is greater among the smokers than among the non-smokers. The bridge between these two notions is that we infer action from association, and that action may cause association. This dissertation will concern itself with the sense of *epistemic interaction*.

Many similar interactions are suggested in newspapers daily: "CO_2 interacts with global warming." "Meditation interacts with stress." "Gender interacts with the amount of earnings." "Nazism interacts with evil." "Eating vegetables interact with long lifespan." How do we know? A priori, we do not. All these conclusions are derived from the data in proper empirical science. For all these conclusions, we can only claim an interaction. People often interpret these claims as causal claims, but they are not. For example, not being under stress might well allow one to meditate, not that meditation would relieve stress. Or, there might be a factor that interacts both with CO_2 and with global warming, but there is no strong direct connection between CO_2 and global warming. Moreover, global warming might even encourage metabolism that in turn generates CO_2. In nature simple causes rarely exist.

Leaving aside these pitfalls, we will formalize the notion of interaction in the epistemic sense. Namely, it is the device of analyzing interactions that underlies a large part of

modern empirical science. Yet, the assumptions that go into this are rarely investigated in detail. For example, the analysis of correlations is based on the assumption of a particular model (Gaussian), a particular fitting method (maximum joint likelihood) and a particular loss function (explained variance).

We can generalize interaction to multiple factors. For example, do the three concepts health, vegetable-eating and health-conscious attitude 3-interact? Yes, they do: health-conscious people eat vegetables, so it is hard to separate other things that health-conscious people do from vegetable-eating in the interaction with health. We can also treat the existence and the worth of an interaction as a random quantity, and estimate the level of faith in the interaction we are justified to have. We can see how important the interactions are to understand the complex phenomena in the world.

But interaction analysis can provide completely novel ways of analysis. For example, we can segment attributes into groups based on whether the attributes interact or not. We can build predictive models purely on the basis of interactions. We can examine how an individual interaction can be structured. Interactions provide a unifying paradigm for machine learning. Certain methods such as cross-tabulation, recursive partitioning, curve fitting, subgroup discovery, and constructive induction are used inside an interaction to capture its structure. On the other hand, the methods of voting and weighting are used to fuse multiple interactions together.

1.2 Overview of the Text

Philosophically, interactions are a particular element of the language we use to build models. The language of models allows interpreting the data. The data are specified in a more primitive language of instances and attributes. A model can be seen as the 'meaning' of the data. We infer models by superimposing the elements of the language onto the data. Because there are several ways of performing such connections, we need algorithms to guide the search. Loss functions judge the quality of a model. We have adopted probability theory to tackle the uncertainty about the predictions, and also the uncertainty about the models in Bayesian statistics. The underlying philosophy is described in Chapter 2. These ideas have been previously published as (Jakulin, 2004, 2005).

A reader less interested in the philosophy of machine learning might prefer to skip directly to Chapter 3. There we choose the particular loss functions and model families that will be used in the remainder of the text. Our models are factorizations of probability distributions. They have convenient properties and form the basis of both information theory (Shannon, 1948) and the recent developments in graphical models (Pearl, 1988, Buntine, 1996). Our loss function will be logarithmic, most commonly used in information theory. We survey the literature on interactions within information theory, and present a thorough historical review. We provide a novel way of illustrating entropy decompositions, and show its meaningfulness on examples. We offer an information-theoretic explanation of the problem of polysemy (a word with multiple meanings) and synonymy (a concept with multiple names). The chapter builds on earlier unpublished work (Jakulin, 2003, Jakulin and Bratko, 2004a).

The probabilistic model is often postulated irrespective of the data, or unquestioningly estimated from the data. The resulting model P is then fixed and perfect. It is a gold standard when we perform entropy decompositions in the context of P. In Chapter 4 we

switch the perspective by having a different gold standard, the data. The model P is not known in advance, but must be learned. For these tasks, we need to apply the tools of statistical modelling. Neither the model nor the information-theoretic quantities tied to it are fixed any longer.

If we were certain that the mutual information takes a specific value in Ch. 3, mutual information becomes a quantity with an uncertain distribution in Ch. 4. We survey various techniques for examining these distributions of information-theoretic quantities, both Bayesian and frequentist. While we have already noticed that information-theoretic quantities are both decompositions and model comparisons, we now distinguish these two ways of interpreting information-theoretic quantities. For entropy decompositions the same gold standard varies together for two models. For model comparison the gold standard varies independently for two models. Based on this dichotomy, we introduce several novel significance tests, and compare them empirically. We also describe a Bayesian treatment of the problem of multiple testing. Another contribution of this chapter is a rigorous definition of the 'amount' of interaction: it is the decrease in loss achieved by modelling the data allowing that the values of all the attributes are observed simultaneously. Interaction information is a practical implementation of this quantity by using Kullback-Leibler divergence as a loss function and the Kirkwood superposition approximation as the no-interaction model. Some of the material of the chapter has previously appeared in (Jakulin and Bratko, 2004b).

In Chapter 5 we show that information-theoretic analysis can also be performed for continuous attributes. While differential entropy can be negative, mutual and interaction information behave in sensible ways. We suggest that it is better to work with mutual information than with Pearson's correlation coefficient. Namely, the correlation coefficient carries the underlying assumption of the bivariate Gaussian model and implies a particular definition of loss. To generalize it, we discuss Gaussian mixture models that handle both continuous and discrete attributes, and can be estimated with the EM algorithm. These distinctly non-linear models can then be used as a ground for computing mutual and interaction information.

Chapter 6 lists numerous novel ways of exploratory data analysis using the notions of interaction and mutual information. Rajski's distance allows transforming information-theoretic quantities into distances, and this allows us to cluster both attributes and their individual values. Interaction graphs and matrices allow us to illustrate interactions among specific pairs of attributes. We employ these tools to address specific problems, such as the concept drift and non-random patterns of missing values. In the process of such analysis we identified some errors in the UCI repository of machine learning benchmarks. Furthermore, it turned out that the statistical significance of interactions was associated with whether an expert found them meaningful on a medical domain; it seems that our assessment of significance is close to human intuition.

In the second half of Ch. 6 we examine ('drill') an individual interaction. Drilling is composed of two kinds of analysis: first, we compare the interaction-assuming model with the no-interaction model; second, we examine what structures are used by the interaction-assuming model. There are many ways in which we can structure the interaction-assuming model: a Cartesian product, recursive subdivision, rules, mixture models, and so on. Several of these structures can be understood as new attributes that were constructed to explain the interaction. Such attributes can be either continuous or discrete, and we

show both approaches on an example of capturing the interactions among the votes of senators of the US Senate. This example is interesting because the problem is very far from being sparse: a very large number of attributes (individual senators' votes) interacts simultaneously. We also discuss that a rule can be understood as a binary attribute: when the attribute takes the value of 1, the rule holds, if the attribute takes the value of 0, the rule does not hold. We conclude the chapter with some ideas on how to perform interaction analysis when there are several thousand attributes, and report on experiments in the domain of text mining. Several of the visualizations in this chapter have been described earlier (Jakulin and Leban, 2003).

The final part of this dissertation deals with the applications of interactions to classification and regression. In Ch. 7 we survey various probabilistic models. We discuss how interactions can be represented in the notation of Bayesian networks. We then focus on the applicability of interaction information as a heuristic for guiding attribute selection, and for deciding what attributes to join. We show that Bayesian networks can represent non-disjunct decompositions of attributes in certain cases. We present a predictive model of the success of hip arthroplasty that was constructed using interaction analysis. We find that interaction information is an efficient heuristic for attribute selection in the naïve Bayesian classifier, and that it outperforms similar approaches. Some of the heuristics were originally discussed in (Jakulin and Bratko, 2003, Jakulin et al., 2003), and the material is based on those two publications.

In Chapter 8 we synthesize the findings and concepts of the previous chapters in the Kikuchi-Bayes model. This model specifies the joint probability model in terms of the constraints that act upon it. Models specified in such a way have been known for some time, but obtaining the explicit parameterization is considered to be time-consuming, and is usually handled with optimization methods. We combine the Kirkwood superposition approximation of Ch. 4 with the recent Kikuchi approximation (Yedidia et al., 2004). Kikuchi approximation yields a generalization of the chain rule, expressible in closed form. We use it for fusing multiple interactions in order to make a prediction.

The second part of Ch. 8 describes the algorithm that underlies the Kikuchi-Bayes classifier. The Kikuchi-Bayes classifier is based on approximate fusion of marginals with the Kikuchi method. It performs a greedy search for the best structure expressed in terms of interactions. It employs a parsimonious Bayesian prior and model averaging (BMA). It is a versatile algorithm that can be used also in tasks other than classification. Furthermore, it competes favorably with other algorithms, even outperforming some support vector machine implementations.

The present text is a journey through several areas of machine learning and some areas outside machine learning. It is impossible to combine breadth with depth; we do not hide the fact that this work is primarily of breadth. A deeper coverage of most terms and expressions, along with their derivations, properties and proofs, appears in the original references. We have attempted to analyze our own theoretical contributions in depth, especially in Chs. 4, 7 and 8. However, our analysis, although careful, is primarily experimental. We find a graph, a picture, an example or a scheme far easier to comprehend and appreciate than an unwieldy sequence of assumptions, arguments, and derivations. All such sequences were omitted from the text, but the claims, assumptions and the equations were checked, double-checked, implemented and tested. We hereby apologize to those readers who are guided more by logic than by intuition.

A Summary of the Contributions

- A machine learning method, Kikuchi-Bayes, for efficient structure learning in interaction models. The method combines approximate Bayesian model averaging by integrating along a single trajectory. Kikuchi-Bayes outperforms a number of popular machine learning algorithms, and is competitive with the state-of-the-art.
- A family of parsimonious priors for Bayesian modelling, motivated in terms of approximating the expected loss on an independent sample of data from the model. Kikuchi-Bayes avoids overfitting the training data by using degrees of freedom to discount the complexity of the model. No time-consuming internal cross-validation is required for capacity control: Kikuchi-Bayes does not overfit.
- A heuristic for fast context-dependent attribute selection for the naïve Bayesian classifier. It is based on summing the interaction information between all the attributes already in the model, the candidate attribute, and the label.
- A number of visualization methods based on the notion of an interaction: information graphs, interaction graphs and interaction dendrograms. A method for clustering attribute values based on an information-theoretic metric. An application of information visualization to political science.
- A formal definition of 'an interaction' through the notion of a part-to-whole model and a model comparison. This definition allows the assessment of nonlinear correlation both for discrete, continuous attributes. Furthermore, the definition is modular and can be adjusted for different loss functions and different hypothesis spaces. Interaction information, mutual information, and correlation coefficient are all particular ways of quantifying interaction.
- A number of significance tests for model comparisons. A Bayesian treatment of the multiple comparisons problem.
- An interdisciplinary historical survey of interaction information across the literature.

A Guide to Notation

$\lvert \cdot \rvert$	absolute value / cardinality of a set / determinant
$\times$	Cartesian product
$\langle \cdot \rangle$	a tuple
$[\cdot]^T$	a vector
$\hat{z}$	an approximation to z
$\triangleq$	'is defined as'
$\mathbb{I}\{C\}$	indicator function: has value 1 when C is true and 0 otherwise
X	an unlabelled attribute; an independent variable
Y	a labelled attribute; the class; a dependent variable
x	the value of attribute X
$\Re_X$	the range of values of attribute X
$\mathbf{X}$	an unlabelled attribute vector
$\mathbf{x}$	the value of an attribute vector $\mathbf{X}$
$\mathbf{x}^{(i)}, \langle \mathbf{x}, \mathbf{y} \rangle^{(i)}$	an instance
$\mathcal{D}$	a data set
Θ	a parameter determining a probabilistic model
θ	a parameter value
$\boldsymbol{\Theta}$	a parameter vector
$P(X, Y\|\theta)$	probability function on X and Y when $\Theta = \theta$
$P(X\|y)$	conditional probability distribution of X given the value of $Y = y$
$p(x\|\Theta)$	marginal probability density of X at $X = x$ parameterized by Θ
H	Shannon entropy
h	differential entropy
$L(y, P)$	the loss incurred by the model P on an outcome y
$D(P\|\|Q)$	Kullback-Leibler divergence between two probability distributions
$D\left(P(Y\|X)\|\|Q(Y\|X)\right)$	KL-divergence between two conditional probability distributions
$\mathbb{E}_{x\sim P}\{L(x, Q)\}$	the expected (average) loss of model Q when x is sampled from P
$H(A\|B)$	entropy of A controlling for B
$H(A, B), H(AB)$	joint entropy of A and B together
$I(A; B)$	mutual information between attributes A and B
$I(A; B, C)$	mutual information between A and the tuple $\langle B, C \rangle$
$I(A; B\|C)$	conditional mutual information between A and B controlling for C
$I(A; B\|c)$	conditional mutual information between A and B when $C = c$
$I(A; B; C)$	interaction information between A, B and C
$C(A, B, C)$	total correlation (multiinformation) between A, B and C

CHAPTER 2

Modelling

The present chapter will present a review of the fields of machine learning, probability and statistics. The chain of thought of the present text begins in Chapter 3, so the reader may wish to skim the following pages. It is impossible to capture the three disciplines in a single chapter without oversimplification, and it involves some subjective discussion of more philosophical issues. First, we introduce the notion of a 'model' that forms under the constraints of the data, assumptions, goals and means of expression. Then we examine several issues in learning, such as validation, consistency, supervised and unsupervised learning. Finally, we examine the foundations of probability theory as a particular means of expression.

2.1 Models

Terms such as data, attribute, probability, algorithm and model are often taken for granted. This section will formulate the problem of machine learning under uncertainty as interplay between four factors: the *data*, the *hypothesis space*, the *utility* and the *algorithm*. All these notions are illustrated in Table 2.1.

	stage	*subject*	*particulars*	*formalization*
1	**percepts**			/
2	representation	**data**	instances, examples	Instance Space
3	learning	**algorithm**	priors, heuristics, procedures	Program Space
4	knowledge	**model**	hypotheses, concepts	Hypothesis Space
5	decision-making	**actions**	utility, preferences	Policy Space
6	**consequences**			/

Table 2.1: The process of learning.

Learning is only a particular phase in the flow of information from percepts to actions and their consequences. Machine learning largely concerns itself with developing algorithms that construct models from a data set. It starts with the product of phase 2

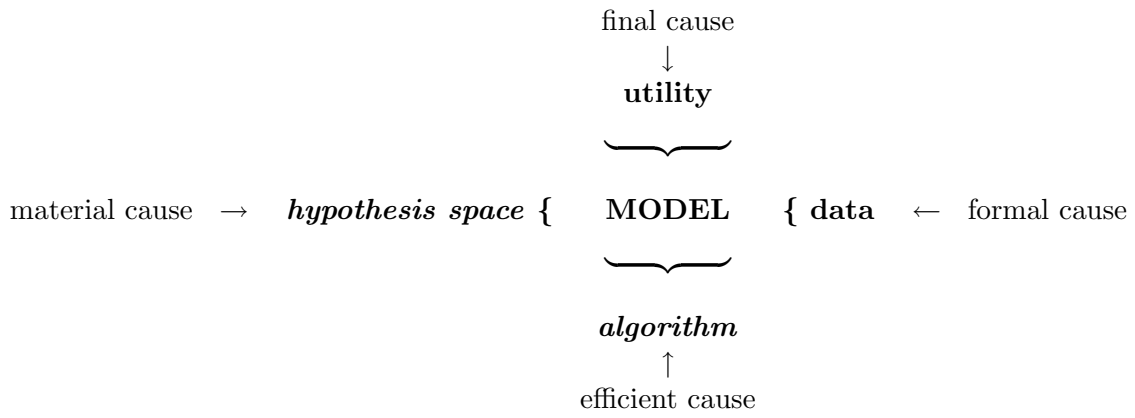


Figure 2.1: The four Aristotelian causes of a model. The symbol $a\{b$ indicates that a is more general than b. The causes are considered to be fixed and constrain the model.

and creates the product of phase 4. It can be seen that the objective truth (before phase 1) and the subjective optimality (after phase 6) are out of its reach. Furthermore, the input to a learner may be corrupted by perception, which is not formalized in the learning problem. The optimality of the actions made with the knowledge can only be judged by the consequences of the actions. These consequences are rarely formalized.

One has to be humble when using terms such as 'true' and 'optimal'. The instance space, program space, hypothesis space and policy space are externally imposed as postulates prior to examining the data. But once they are formally represented, they can be studied with mathematical rigor. The instance space defines what is the data and what can be distinguished. An instance space for the coin toss is that the coin can fall either heads or tails: these two values are all we distinguish. The hypothesis space defines which models are possible. The models of the linear regression hypothesis space are of the form $y = ax + b$, so a particular model is a pair of parameters $\langle a, b \rangle$. The program and policy spaces restrict the kinds of policies and learning procedures that can be developed. For example, the program space is commonly the language of Turing machines, meaning that each learning algorithm can be expressed as a particular Turing machine. The same is valid for policy spaces, and this is a concern of reinforcement learning (Sutton and Barto, 1998).

Our focus in this text will be on the four Aristotelian causes of a model, as shown in Fig. 2.1. Model will be the focal element of our investigation, and the model cannot be seen as independent of these four causes. The model arises from an interaction between the internal hypothesis space and the data, and the external utility and algorithm. The model is expressed in terms of the hypothesis space, conforms to the data, is generated by procedures, priors and rules of the algorithm, and is judged by the utility.

		the label ↓
attributes →	X	Y
an instance →	rain	H
an instance →	sunny	H
an instance →	sunny	T
.	rain	H
.	sunny	T
.	sunny	T
.	rain	H
.	rain	H

Table 2.2: A non-deterministic data set: for the same combination of attributes, there are multiple labels. The weather is described with an unlabelled attribute X with the range $\Re_X = \{\text{rain}, \text{sunny}\}$. The coin is modelled as a labelled attribute Y with the range $\Re_Y = \{\text{H}, \text{T}\}$.

2.2 Dichotomies in Learning

2.2.1 Obtaining a Model from Data

We will now discuss different theories and views of machine learning. We will present the notion of truth versus approximation: identification is based on the idea that the hypothesis space is assumed to be complete and universal, so the one true model is sought. On the other hand, approximation merely seeks to minimize the loss and maximize the utility of the final hypothesis, but also assumes that the loss function is true and universal. The identification view of learning trusts that the choice of the hypothesis space is correct (but is agnostic about the utility function), whereas the approximation view of learning trusts that the utility function is correct (but is agnostic about the choice hypothesis space).

Most data sets in practical machine learning are structured with instances and attributes. The data consists of a number of *instances* (or experiments or examples), and each instance is described with a number of *attributes* (or variables). Each attribute X can take upon a number of values, its *range* $\Re_X$. Some of the attributes are *labelled*. The objective of learning is to predict labelled attributes $\mathbf{Y}$, using the information provided by the unlabelled ones $\mathbf{X}$. Ideally, we would denote the hypothesis as a function $\mathbf{Y} = f(\mathbf{X})$: the domain of the function are the unlabelled attributes' range, and the codomain is the labelled attributes' range. The value of the attribute X for the instance (i) is $x^{(i)} \in \Re_X$. If there are several attributes, we may represent them together in an attribute vector $\mathbf{X} = [X_1, X_2, \ldots, X_M]$, and we refer to $\Re_{\mathbf{X}}$ as the attribute space. These concepts are illustrated in Tab. 2.2.

The hypothesis space is the internal 'language' in which the models are described. Individual hypotheses or models are points in the hypothesis space. For example, a statement in the logical hypothesis space would be **if** $X = a$ **then** $Y = b$, where a is a particular value of the attribute X and b is a particular value of the labelled attribute Y. A specific model would be **if** $X = rain$ **then** $Y = H$. Mathematical expressions are also a

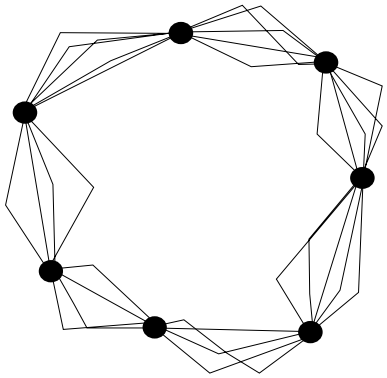


Figure 2.2: Multiple consistent hypotheses form the version space. We assume a hypothesis space of line segment sequences: each model is a particular sequence of line segments. Many hypotheses in this hypothesis space are consistent with the data set consisting of 7 points in space.

hypothesis space ($y = a \times x + b$), as are case-based inferences (**if** X like a **then** Y like b) and non-causal inferences ($X = a$ **with** $Y = b$).

Learning as Identification

The original learning theory (Valiant, 1984) was concerned with problems of deductive identification. We assume that there exists some true deterministic concept. The data is sampled from the world, and identified whether it belongs to the concept or not. Consequently, we try to identify that concept from this data. The concept corresponds to a model within a particular hypothesis space. Valiant proved that the learning problem is tractable for several non-trivial hypothesis spaces, such as expressions in conjunctive and disjunctive normal forms with a bounded number of literals.

More generally, the learning procedure can thus be phrased a search for a model in an arbitrary hypothesis space that is consistent with all the data. But sometimes several different models may be consistent with the data. Imagine a hypothesis space in which a hypothesis consists of segmented lines: there can be many such segmented lines that successfully pass through a sample of 5 points on a sphere, as shown in Fig. 2.2. These hypotheses form a space of their own, the *version space* (Mitchell, 1997). The algorithm usually chooses a single hypothesis in the version space that has the highest utility. The utility in this case can simply be the simplicity of the hypothesis. In the case of line segments, we could seek the set with the fewest line segments, or the smallest set of line segments that are all of equal length. Occam's razor is a common utility function that favors simplicity as an inherent quality for choosing one among several consistent hypotheses.

Learning as Approximation

For some hypothesis spaces, finding a consistent hypothesis may be impossible: we cannot find any hypothesis $\langle a, b \rangle$ in the hypothesis space of linear functions $y = ax + b$ that would be consistent with the data that consists of points on a sphere, as in Fig. 2.2. In the PAC (probably approximately correct) view of learning there is an inherent assumption that

the instances are indeed consistent with some hypothesis allowed by the hypothesis space. This problem is usually not apparent as hypothesis spaces such as CNF are usually able to capture any consistent set of data. There are situations, however, when the data is not consistent with any hypothesis, as in Table 2.2.

An agnostic learner (Haussler, 1992) does not attempt to find a function that would be fully consistent with the data. Instead, a function is sought that results in minimum loss compared to other functions of the hypothesis space, even if it is not fully consistent. The loss or negative utility is quantified by a loss function that measures the difference between the true outcome y and the predicted one $\hat{y}$. A popular choice of a loss function is the classification error or 0-1 loss:

$$L_{01}(y, \hat{y}) = \begin{cases} 0 & ; \quad y = \hat{y} \\ 1 & ; \quad y \neq \hat{y} \end{cases} \tag{2.1}$$

In the data set of Table 2.2, the prediction that the outcome of the coin toss will be 'H' results in a lower classification error than the prediction that it is 'T', so indeed one of the two possible functions would be selected. However, the concept of the coin toss is not learnable in the sense that the function's error will not become arbitrarily small with an increasing number of instances. However, it is agnostically learnable if it is possible to learn a function arbitrarily close to the best of all feasible functions. Thus, even if zero loss is unattainable, we can still achieve zero *regret* R (Bernardo and Smith, 2000) by picking an appropriate prediction $\hat{y}$:

$$R(y, \hat{y}) = L(y, \hat{y}) - \inf_{y'} L(y, y') \tag{2.2}$$

2.2.2 Noise vs Conditional Probability

The second dichotomy concerns the interpretation of non-deterministic phenomena. Most theories represent uncertainty with probability but differ with respect to what the probability refers to. One view is that probability arises because of the inherent 'noise' in data. Another view is that the data is a sample from a probabilistic model; this does not mean that the reality is non-deterministic, as probability can also arise because of an incomplete picture of the reality.

Let us consider a hypothetical data set in Table 2.2. It is quite clear that the attribute X is insufficient to predict the outcome Y of the coin toss. A larger number of instances would not help, nor would the choice of a different hypothesis space. No function is consistent with these observations because the same weather situation may imply two different coin toss outcomes: sunny weather appeared both with heads and with tails. It is known that a controlled coin toss is deterministic and predictable (Diaconis et al., 2004), but because most actual coin tosses are not fully controlled, probability arises.

Noise

One interpretation of the conflict is to consider the coin toss to be corrupted by *noise*. Specifically, we model the process as $\mathbf{Y} = f(\mathbf{X}) + \epsilon$, where ϵ is a model of noise or error. In classification $\mathbf{Y}$ is not a number, so the addition of noise is often interpreted as an occasional random change of the label. The aim of learning is to minimize the amount of

error. When the range of $\mathbf{Y}$ is a discrete set, this problem is referred to as *classification*, and when $\Re_{\mathbf{Y}} = \mathbb{R}$, as *regression*.

The learning itself can be formulated as an agnostic learning problem of $f(\mathbf{X})$, with the objective of minimizing the error. Whatever error there may be, it is interpreted as noise ϵ. This way, the model phrased as $f(\mathbf{X}) + \epsilon$ is consistent with the data, but the utility function favors a smaller ϵ. Additive noise is just one of the choices, multiplicative noise is also used sometimes. The core idea, however, is that f is a deterministic function.

Conditional Probability

A different strategy is to reformulate the learning problem. Instead of trying to learn explicit functions of the type $\mathbf{Y} = f(\mathbf{X}) + \epsilon$, we learn conditional probability models $\hat{P}(\mathbf{Y}|\mathbf{X}, \mathbf{\Theta})$, where $\mathbf{\Theta}$ is a vector of *parameters* that describe the model. In this case, the hypothesis space is the parameter space. This is no longer classification or regression, but conditional *probability modelling*. Conditional probability models already subsume the uncertainty of the model without separating the predictions from their error. The predictions themselves capture the error. Here, the best model would be the one that would correctly assess the probabilities of all outcomes, rather than attempt to predict the single most likely labelled value. For example, if a patient has a 90% chance of surviving a surgery, the model should predict that the probability is $p = 0.9$ rather than attempt to guess the most likely outcome. Probability is a formalization of chance that physicians understand well. They make use of it when making decisions in a complex set of circumstances. Hypothesis spaces making use of probability are therefore a good choice for medical models.

It is possible to assume that the data was indeed generated by a particular model, and then identify its parameters. An agnostic approach is also possible. There we do not assume that the data were indeed generated by the model, but we seek a model that will achieve the greatest utility through probabilistic predictions of the data. Thus we can define probabilistic loss or *p-loss* functions (Grünwald and Dawid, 2004) that evaluate the loss made by a probabilistic prediction of events in the face of the event that actually took place.

Examples of p-loss functions are the logarithmic loss $L_{LL}(y, \hat{P}) = -\log \hat{P}(y)$ and the Brier loss. By minimizing probability loss functions we can formulate the agnostic probabilistic learning. Because of randomness, even the correct models generally do not achieve zero p-loss. We usually require the p-loss functions to be *proper* (Bernardo and Smith, 2000). A proper p-loss function guarantees that for data y generated from P:

$$\inf_{\hat{P}} \left(\sum_{y \in \Re_Y} P(y) L(y, \hat{P}) \right) = \sum_{y \in \Re_Y} P(y) L(y, P), \tag{2.3}$$

where the infimum over all the possible $\hat{P}$ is attained if and only if $\hat{P} = P$. In essence, we compute the expectation of p-loss over all the data y in the data set. Both of the above p-loss functions are proper, and can be used to assess the probabilistic calibration of a model. Without a proper and discriminate p-loss function, the resulting minimum loss 'probability' may no longer be meaningful.

However, some p-loss functions are not proper, such as the information score (Kononenko and Bratko, 1991), the 0-1 loss, and the classification error. Information

score and 0-1 loss reach the optimum for $\hat{P} \neq P$. Classification error is improper because it is indiscriminate: although the loss of $\hat{P}$ may reach a minimum where $\hat{P} = P$, there are $P' \neq P$ that get the same loss. For example, the indiscriminate classification error will not distinguish between a model that confidently predicted the event with the probability of 1.0 and another one that predicted it timidly with the probability of 0.501. Also, classification accuracy will not penalize a model that predicted an event that happened with the probability of 0.0 any more than a model that predicted the same event with the probability of 0.49.

To demonstrate the proper, improper and indiscriminate p-loss functions, we have performed an experiment. We took a twisted coin that falls heads with the probability of 0.6 when thrown on a table. We tossed the coin infinitely many times to obtain a series of outcomes. The probabilities do not exist in the data, which in this case is merely a sequence of H and T, but must be inferred. We have tested all the predictions ranging from 0 to 1. For each prediction, we have evaluated the mean p-loss on the data set. The results are shown in Fig. 2.3. We can see examples of proper p-loss functions (logarithmic loss and Brier loss) that indeed achieve the minimum at 0.6. The improper p-loss functions zero-one loss and information score have no meaningful minima or maxima at 0.6. In this case, both 0-1 loss and information score would favor always predicting heads with 100% certainty. Furthermore, classification error would not distinguish between models predicting the probability of heads to be larger than 0.5. Information score is actually a utility function, as we seek to maximize it, so it was negated in the chart. Furthermore, information score is expressed relative to the prior distribution, for which we employed the true probability itself.

The probability estimation view is especially favorable in situations when the choice of an explicit function f would be misleading. For example, if we have a bag with a mix of 50 apples and 50 cherries, we can expect the weight of a randomly chosen object from the bag to be either approximately 150 ± 30 grams (an apple) or 15 ± 2 grams (a cherry). It is quite unlikely that any object would have the mean weight of 82 grams. The conditional probability would be a bimodal mixture of two normal distributions $\frac{1}{2}(\mathcal{N}(150, 30) + \mathcal{N}(15, 2))$. Such a situation is not realistically captured by an explicit function that presupposes a mean value of 82.5 with the normally distributed noise with the variance of 70.

2.2.3 Generative vs Discriminative Learning

In the previous section we discussed a *discriminative* learning problem where one labelled attribute is predicted based on the knowledge of an unlabelled one. In general, when we predict the labelled attributes, not the unlabelled ones, we speak of supervised or discriminative learning. On the other hand, when we do not distinguish the labelled from the unlabelled attributes, but predict them all, we speak of a *generative* (Jebara, 2003), an *informative* (Rubinstein and Hastie, 1997), or an unsupervised learning problem.

Discriminative Learning

In discriminative or supervised learning, we have a labelled vector of attributes $\mathbf{Y}$, and an unlabelled vector of attributes $\mathbf{X}$. Thereby, discriminative learning is achieved either probabilistically with a conditional probability model $\hat{P}(\mathbf{Y}|\mathbf{X}, \boldsymbol{\Theta})$, or using a deterministic model of a function plus noise $\mathbf{Y} = f(\mathbf{X}) + \epsilon$. Examples of practical discriminative methods

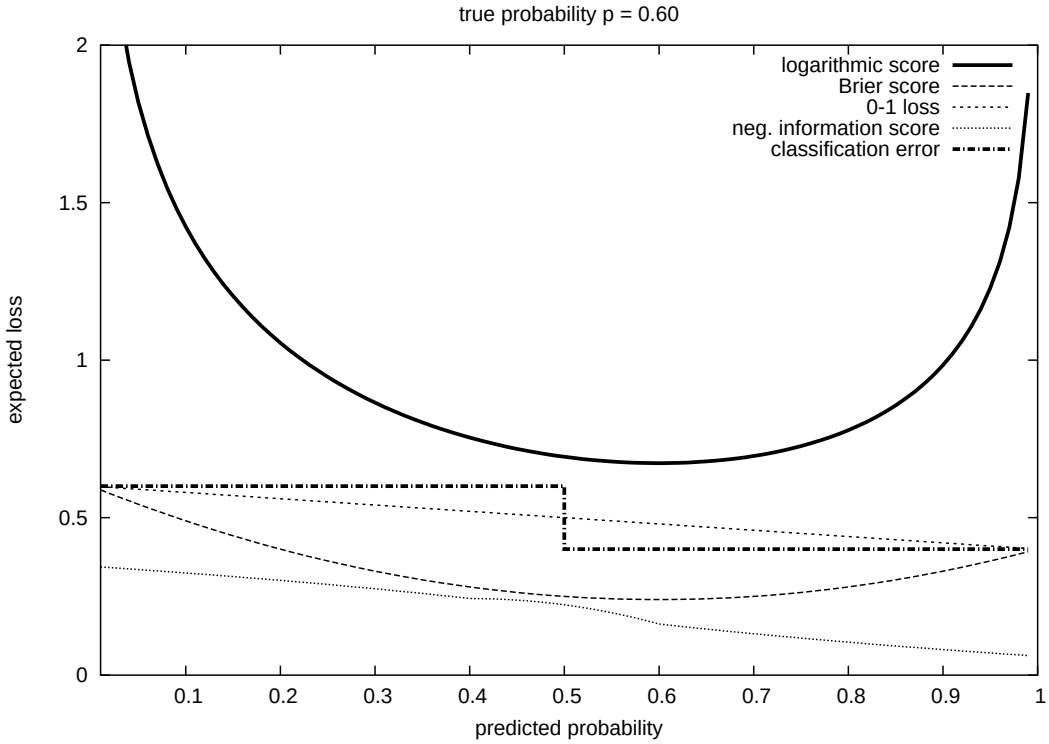


Figure 2.3: Proper and improper p-loss functions. Improper p-loss functions attain the minimum at the incorrect probability. For example, predicting the majority class for a sample is a win when evaluating with the 0-1 loss or the information score.

are logistic regression, Gaussian processes and support vector machines.

The justification for discriminative learning is that it is simpler, and usually sufficient for most needs. Many practical applications involve discriminative prediction: we predict the outcome of a surgery, the survival of a patient, the validity of a transaction. There are, however, disadvantages to the discriminative view that will become apparent in Sect. 2.2.4.

Generative Learning

In generative or unsupervised learning, there is no difference between the labelled and unlabelled attributes in the representation of the model. What is built is a *joint probability model*: $\hat{P}(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})$. This means that all attributes, both labelled or unlabelled, can be predicted from the values of the model's parameters $\mathbf{\Theta}$. A particularly simple model is the COBWEB approach to conceptual clustering (Fisher, 1987), where the $\mathbf{\Theta}$ consists of a single nominal attribute (category), that captures the dependencies between attributes well.

It is also possible to employ the joint model for classification. By conditioning the joint probability model, we can obtain the conditional model for the labelled attributes $\mathbf{Y}$ given the unlabelled ones $\mathbf{X}$:

$$\hat{P}(\mathbf{Y}|\mathbf{x}, \mathbf{\Theta}) = \frac{\hat{P}(\mathbf{Y}|\mathbf{x}, \mathbf{\Theta})}{\sum_{\mathbf{x}' \in \Re_{\mathbf{X}}} \hat{P}(\mathbf{Y}|\mathbf{x}', \mathbf{\Theta})} \tag{2.4}$$

While discriminative learning coexists with probabilistic models (Rubinstein and Hastie, 1997), it is difficult to see how the noise model would be applicable to generative models.

2.2.4 Generalization

The model that achieves the lowest loss on the training data may not achieve the lowest loss in later applications of the model. The specific problem that usually arises is referred to as *overfitting*: the model is overconfident about the validity of its predictions. On the other hand, *underfitting* describes the phenomenon of a model not being confident as much as it could be. While overfitting is easily detected, underfitting is identified through a superior model that does not overfit.

There are numerous methodologies for dealing with overfitting, and they are sometimes mentioned as *capacity control*. Some operate in the domain of utility, such as regularization, and penalize complexity. Validation approaches separate the training and testing data. Bayesian priors explicate the assumptions about feasible models. These approaches are not mutually exclusive and can be combined. Prequential learning is based upon treating the data set in a sequence, not as a batch.

Regularization or Penalty Methods

Regularization (Golub et al., 1999, Tikhonov and Arsenin, 1977) is a modification of the loss function that takes the complexity or some other undesirable aspect of the model into consideration. This aspect is then interpreted as additional cost of the model. For example, if we are learning a function $\hat{y} = w_0 + w_1 x_1 + w_2 x_2$, we might want the weights w_0, w_1, w_2 to be as low as possible. This is achieved by introducing a regularized loss function

$$L(\mathbf{y}, \hat{\mathbf{y}}) = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 + \lambda \|\mathbf{w}\|^2. \tag{2.5}$$

Thereby, we would add $\lambda(w_0^2+w_1^2+w_2^2)$ to the ordinary quadratic loss, and so large weights would be penalized. There are other choices of regularization functions, for example, lasso employs $\|\mathbf{w}\|$ (Tibshirani, 1996). A regularized loss function also prevents ambiguity when multiple solutions satisfy an ill-defined problem. Regularization will penalize certain solutions. Corrupting the data with noise can be seen as a particular kind of regularization (Bishop, 1995), albeit an indirect and a less efficient one.

Certain model selection heuristics, such as AIC (Akaike information criterion) and BIC (Bayesian information criterion) can be interpreted as penalized logarithmic p-loss functions. With AIC, the regularization term is $2k$ where k is the effective number of parameters. With BIC, the regularization term is $2k \log n$, where n is the number of instances. Therefore, both AIC and BIC penalize the models by the number of their parameters, with the assumption of parameter independence. BIC also takes the data set size into consideration, increasing the penalty with the size of the data set. The aim of AIC is to minimize the expected loss, whereas BIC attempts to maximize the probability of identifying the correct model.

Partitioning and Resampling

The idea behind validation is to separate the training from the test set: the training set is used for building the model, and the test set for evaluating the resulting model's utility; this way we prevent the model from simply memorizing the instances and 'peeking at the correct answers'. The resulting validated utility will reflect the mistakes in generalization. The idea underlying the validation is that a reliable model will be able to show a consistent gain in utility with incomplete data. By induction, if a model achieved reliable performance with a part of the given data, we then expect that it will also perform well on future truly unseen data.

It is important to note that the loss depends on both the test/training proportion and on the particular choice of the partition. K-fold cross-validation (Stone, 1974) performs multiple experiments using the same size test/training proportion in order to reduce the influence of the partition choice. The data is split into K subsets of equal size, and from these, K training/test splits are made, so that each subset is once the test set and $K-1$ times the training set. The loss estimate obtained with cross-validation depends both on the initial partitioning, and on the number of subsets used. To further eliminate the influence of the initial partitioning, Kohavi (1995) recommends multiple replications of cross-validation. The replications are repeated until the standard error across the folds falls to a certain level, implying that smaller data sets will involve more replications than large ones. This does not, however, remove the dependence on the choice of K, which will be illustrated in the next section.

Leave-one-out is a special case of K-fold cross-validation where $K = n$: the subsets are of size 1. Leave-one-out fully removes the dependence on the particular choice of the partitioning, as all possible choices of the training/test split are used. Still, leave-one-out does not remove the dependence on the test set being of size 1 and the training set being of size $n-1$. Furthermore, the loss is computed for each instance individually, so the leave-one-out estimate of loss for each instance is a scalar value with zero variance.

To remedy these overconfident estimates of loss in leave-one-out, Hastie et al. (2001) recommend performing a bootstrap and leave-one-out simultaneously. First, a number of bootstrap resamples of the original data are created. Each *resample* has the same size n

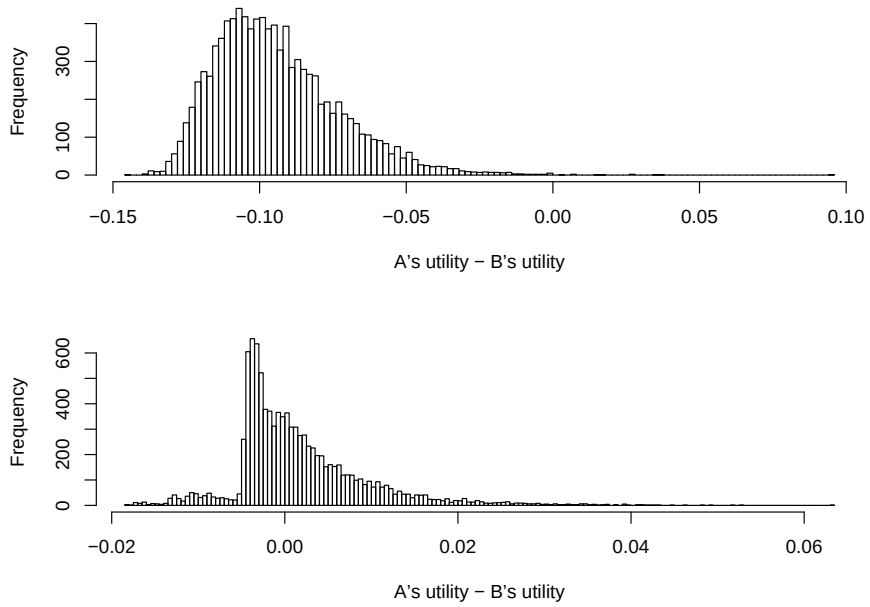


Figure 2.4: Replicated comparisons. We can compare the utility of two models over several experiments. Sometimes it is easy (top), and sometimes hard (bottom) to decide which model is better, A or B.

as the original data set, but because sampling with replacement was used, a particular instance can appear more than once, and some instances may not appear. For each resample, a model is estimated. In the leave-one-out phase, the distribution of prediction loss is assessed using all those models that were trained without the particular instance under consideration.

Sometimes the choice of the model is clearly ambiguous: the problem is illustrated in Fig. 2.4: two models A and B were tested over a large number of experiments in two contexts. For each experiment the utility of model B was subtracted from the utility of model A. In the first case (top) the model B achieved a predominantly higher utility than model A. Of course, there is a small number of situations when A was better. In the second case (bottom), deciding which model is better becomes a very difficult problem: in the most frequent case (mode), B was better; for the average utility over all experiments, A was better; in the average case (median), B was better; in the best case, A was better; at the worst, B was not as bad. What to do? Deciding between two models may be ambiguous even when the consistent and quantitative utilities are given in full detail. Of course, such a dilemma only arises when the methods are similar in performance: then we could argue that *any* choice would be fine.

Estimator Bias and Variance

Assume that a generative model $P(\mathbf{X}, \mathbf{Y}|\mathbf{\Theta})$ is estimated from a data set $\mathcal{D}$, using an estimator T: $\boldsymbol{\theta} = T(\mathcal{D})$. We randomly generate instances from $P(\mathbf{X}, \mathbf{Y}|\boldsymbol{\theta})$, and form data sets $\mathcal{D}_1^*, \mathcal{D}_2^*, \ldots$ of the same size n as the original one $|\mathcal{D}| = |\mathcal{D}_\cdot^*| = n$. For each $\mathcal{D}_i^*$ we estimate $\hat{\boldsymbol{\theta}}_i = T(\mathcal{D}_i^*)$. The estimator T is unbiased iff $\mathbb{E}_i\{\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\} = 0$ for all $\boldsymbol{\theta}$ and all n. If not, $\mathbb{E}_i\{\hat{\boldsymbol{\theta}}_i - \boldsymbol{\theta}\}$ is the *bias* of T for sample size n at point $\boldsymbol{\theta}$ (Dietterich and Kong, 1995).

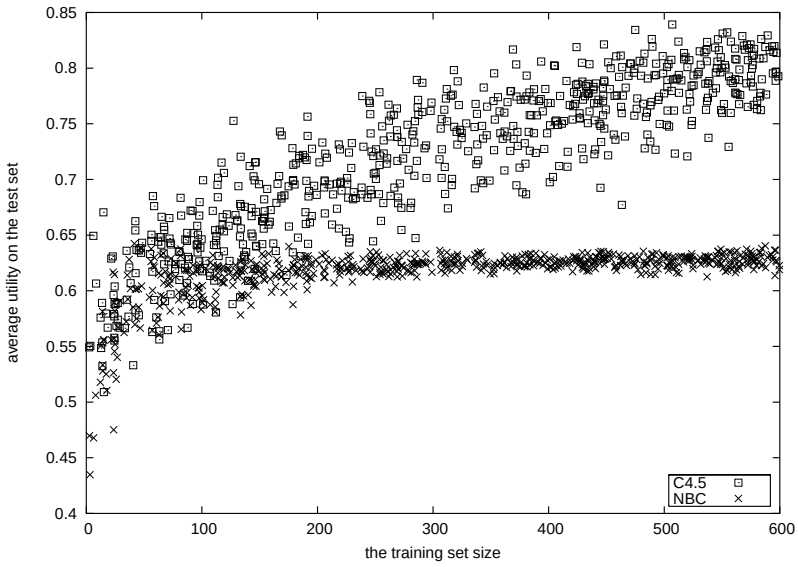


Figure 2.5: The learning curves. Most models become better with an increasing number of instances. Some of them quickly reach a plateau and result in reliable utility. Others take more chances, and reach greater levels of utility, but pay a cost in reliability.

We often pursue unbiased estimators and learning procedures. If the estimator is biased, we may employ the parametric bootstrap in order to estimate the bias and correct it (Davison and Hinkley, 1997). A second-order bootstrap may be needed if the bias correction is itself biased. Once the estimator is unbiased, we may seek to minimize the variance of the estimator. Note that the above formulation of bias involves the assumption that $\boldsymbol{\theta}$ is indeed the true model.

This definition of bias and variance referred to the model's parameters. On the other hand, the bias and variance in the model's loss underlies the *machine learning bias and loss* (Dietterich and Kong, 1995). It is possible to separate the bias from the variance in the model's loss using the bias/variance decomposition (Geman et al., 1992, Murphy, 1973). Models that achieve consistent loss across data sets have low variance, while models that achieve low average loss have low bias. There is often a trade-off involved between choosing a hypothesis space with low bias but possibly high variance (such as a classification tree), versus a hypothesis space with higher bias but low variance (such as the naïve Bayesian classifier).

Just as cross-validation, bias and variance too depend on the amount of data used for training, and this dependency can be analyzed using *learning curves* (Kadie, 1995). A learning curve shows the relationship between the performance of a model on unseen data depending on how much data was used for training. If the utility no longer changes, the model has converged and additional data is less likely to affect the model. In Fig. 2.5 we compare two commonly used algorithms in machine learning, the naive Bayesian classifier (NBC), and the C4.5 classification tree induction algorithm (Quinlan, 1993), on the 'tic-tac-toe' data set using zero-one p-utility.

The utility is not simple to characterize when there is little data (less than 50 instances), but NBC is less robust than C4.5. When there is more data (50-150), it is still

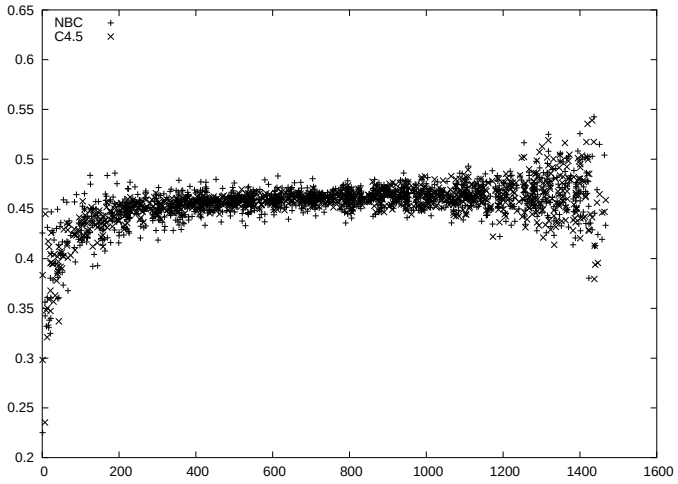


Figure 2.6: The test set needs to be large enough to reliably evaluate the expected utility and its variance for a learning curve. If there are fewer than 400 instances in the test set, the variance in the expected utility increases.

difficult to compare both methods. Beyond 150 instances, NBC becomes reliable: we know that the NBC model requires approximately 150 instances to be characterized almost unambiguously. On the other hand, C4.5 keeps gaining utility indefinitely. Therefore, two conclusions can be made: the NBC model is simple enough to be identified unambiguously with 300 instances: this is good, as there are 960 instances in that data set. And, when there are 250 instances, the C4.5 model has not yet fully converged, but it is already clear that it is consistently better than the NBC model. It is important to note that the test set also has to be large enough, otherwise the estimate of the expected utility becomes unreliable, as shown in Fig. 2.6. Although we can average multiple replications, we can no longer evaluate the variance of the utility in such a case.

There is an important connection between simplicity and variance. It is often thought that simple models have lower variance, but it would be mistaken to assume that this connection is causal or rigid. Whether a complex hypothesis space will yield models with high variance depends upon the prior assumptions and on the algorithm. Seemingly complex models often have low variance (Breiman, 1996). This kind of low-variance models are obtained by a frequentist equivalent of Bayesian model averaging: we average fitted models over a number of data set perturbations.

Bayesian Priors

In Bayesian statistics the models are not estimated. The model $\boldsymbol{\Theta}$ is no longer seen as a specific parameter value, but as another attribute to be modelled probabilistically. Of course, the model is not a random attribute in the stochastic sense. Instead, we have different *degrees of belief* in a particular parameter that specifies the model. Technically, degrees of belief and probabilities are very similar, but they have different semantics: degrees of belief refer to models, whereas probabilities refer to the data. The earlier notions of cross-validation and bootstrap may also be phrased in terms of belief; for example, in bootstrap we believe in a number of possible data sets, not just the given one,

even though we only believe in a single model for each data set.

We use $P(\mathbf{\Theta}|\mathcal{D},\mathcal{H})$ to indicate the posterior belief of a particular model given the data $\mathcal{D}$ and the choice of the hypothesis space or the hypothesis space $\mathcal{H}$ (MacKay, 2003). In most practical circumstances, there are many values of $\mathbf{\Theta}$ that all have non-zero belief given the data. Practical Bayesian inference requires the *prior* $P(\mathbf{\Theta}|\mathcal{H})$ to be defined. The prior captures the expectations about the probability of each of the possible models before seeing the data, or when the data set is of size zero. In our Aristotelian scheme of Fig. 2.1, priors correspond to the algorithms. The posterior belief in a particular model is then developed through the *likelihood* of the model $P(\mathcal{D}|\mathbf{\Theta},\mathcal{H})$:

$$P(\mathbf{\Theta}|\mathcal{D},\mathcal{H}) = \frac{P(\mathbf{\Theta}|\mathcal{H})P(\mathcal{D}|\mathbf{\Theta},\mathcal{H})}{P(\mathcal{D}|\mathcal{H})}. \tag{2.6}$$

The likelihood is usually expressed through a likelihood function, which is one of the assumptions. Generally, the likelihood function weights how well the interesting aspects of the data matches the model. If we assume that all instances in the data set $\mathcal{D}$ are conditionally independent given the parameter $\boldsymbol{\theta}$, the likelihood function is:

$$P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H}) = \prod_{\mathbf{x}\in\mathcal{D}} P(\mathbf{x}|\boldsymbol{\theta}). \tag{2.7}$$

Because the *evidence* $P(\mathcal{D}|\mathcal{H})$ is usually not modelled, we can get rid of it and normalize the posterior by summing or integrating over the hypothesis space:

$$P(\mathbf{\Theta}|\mathcal{D},\mathcal{H}) = \frac{P(\mathbf{\Theta}|\mathcal{H})P(\mathcal{D}|\mathbf{\Theta},\mathcal{H})}{\sum_{\boldsymbol{\theta}\in\Re_{\mathbf{\Theta}}} P(\boldsymbol{\theta}|\mathcal{H})P(\mathcal{D}|\boldsymbol{\theta},\mathcal{H})}. \tag{2.8}$$

In some cases, however, we can use the evidence as a measure of how well our hypothesis space copes with the data.

Observe that every statement in Bayesian inference is conditional upon the hypothesis space $\mathcal{H}$. Because the assumption of the hypothesis space $\mathcal{H}$ is always implied, we will use $\hat{P}(\cdot)$ as an abbreviation for $P(\cdot|\mathcal{H})$.

The prior sets the preferences among various models in the hypothesis space. The prior may be *subjective* and thus indicate which models are likelier, from experience on other problems or domains, for example. *Objective* priors do not include such preferences, but instead consider all models to be equally likely, or that the predictions are independent of the parametrization. There are further types of priors: hierarchical priors are priors on priors, conjugate priors are computationally convenient, and empirical 'priors' are estimated from the data. The concern about priors is important in Bayesian statistics, as modelling is essentially about assuming the hypothesis space, the likelihood function and the prior over the models in the hypothesis space: the golden standard algorithm simply considers all the possible models.

The posterior belief in a model $\hat{P}(\mathbf{\Theta}|\mathcal{D})$ can be interpreted as a kind of a proto-utility function. In this view, the *maximum a posteriori* (MAP) model $\hat{\boldsymbol{\theta}}$ is best, and can be determined through:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \hat{P}(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \hat{P}(\boldsymbol{\theta})\hat{P}(\mathcal{D}|\boldsymbol{\theta}) \tag{2.9}$$

Picking the model with the minimum description length (MDL) (Rissanen, 1986) is identical to picking the MAP model. With uninformative priors, the MAP models may be

equal to the maximum likelihood models. Furthermore, both the description length of the model in MDL and the prior belief in a model for MAP can be seen as a penalty in a regularization framework. Finally, regularization can too be seen as the imposing of a prior with MAP, assuming that the probability is associated with utility, as in the pursuit of highly probable models.

For conditional modelling (such as regression), one can also use a conditional likelihood function. If $\mathbf{y}$ is predicted using $\mathbf{x}$, the conditional likelihood function can be defined under the assumption of instance independence as:

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) = \prod_{(\mathbf{x},\mathbf{y})\in\mathcal{D}} P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}). \tag{2.10}$$

The underlying assumption of Bayesian conditional modelling is that we assume a prior composed of two independent parameters $P(\boldsymbol{\theta}, \boldsymbol{\psi}) = P(\boldsymbol{\theta})P(\boldsymbol{\psi})$. Furthermore, we assume a factoring of the likelihood in $P(\boldsymbol{\theta}, \boldsymbol{\psi}) = P(\boldsymbol{\psi}|\mathbf{X})P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y})$. We then focus just on $\boldsymbol{\theta}$:

$$P(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Y}) \propto P(\boldsymbol{\theta})P(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) \tag{2.11}$$

For a justification and a discussion, see (Gelman et al., 2004a). With this conditional likelihood function we can obtain both conditional posterior distributions and maximum a posteriori conditional probability models. The conditional approach is preferable when we want to avoid spending parameters for modelling the distribution $P(\boldsymbol{\psi}|\mathbf{X})$.

Prequential Analysis

The basic idea of prequential analysis (Dawid, 1984) is that the data is not to be considered a single piece of information, but instead a sequence of instances. We start from the initial prior assumptions, and transmit one instance after another into the model. For each instance, the model suffers a prediction loss, but also updates itself to better predict the successive instances. In that sense, prequential analysis is reminiscent of sequential data compression, or of on-line learning. However, if the data is initially not an ordered sequence, assuming a particular ordering would be misleading.

Entropy-Based Analysis

There are many models $\boldsymbol{\theta}$ that satisfy a particular set of constraints we might impose on it: Π is the set of satisfactory models. For example, we might want to enforce the constraint that the model should have a particular value of the mean and a particular value of the standard deviation. Or, we could enforce the constraint that a attribute is bounded within $[a, b]$. The question is which specific model of many $\boldsymbol{\theta} \in \Pi$ for the attributes $\mathbf{X}$ is preferable. The maximum entropy principle (Jaynes, 2003) states that one should pick the model that results in maximum entropy. For a definition of entropy, refer to Sect. 3.1. There are three main interpretations and generalizations of the MaxEnt approach (Topsøe, 2004, Grünwald, 1998):

- **Entropic loss** (Jaynes, 2003). If entropy is interpreted as a loss function, the maximum entropy model will be the worst of all the satisfactory models:

$$\boldsymbol{\theta}_{ME} = \arg\max_{\boldsymbol{\theta}\in\Pi} H(\mathbf{X}|\boldsymbol{\theta}) \tag{2.12}$$

In that sense, it will not provide any more information than what is provided by the constraints that form Π.

- **Entropic projection** (Csiszár, 1991). If relative entropy is interpreted as a distance from a particular 'prior' model $\boldsymbol{\theta}_0$, we seek the model in Π that is closest to $\boldsymbol{\theta}_0$, or the *I-projection* of $\boldsymbol{\theta}_0$ on Π:

$$\boldsymbol{\theta}_{MD} = \arg\min_{\boldsymbol{\theta}\in\Pi} D(P(\mathbf{X}|\boldsymbol{\theta})\|P(\mathbf{X}|\boldsymbol{\theta}_0)) \tag{2.13}$$

  The same idea was suggested earlier by Kullback (1968), and is currently referred to as *variational inference* (Wainwright and Jordan, 2003). The justification for this approach is that while $\boldsymbol{\theta}_0$ might be unknown or intractable, the hypothesis space Π is tractable and can be used to represent $\boldsymbol{\theta}_0$. Also note that this approach closely resembles minimization of expected logarithmic p-loss, just that the expectation is computed using the approximation, not using the 'truth'.

- **Entropic equilibrium** (Harremoës and Topsøe, 2001). Let us imagine a game played by nature and the statistician, where the nature picks distributions from $\mathcal{P}$ and the statistician the predictions from Π. What the statistician seeks is the minimum risk model:

$$\boldsymbol{\theta}_{MR} = \arg\min_{\boldsymbol{\theta}\in\Pi} \left(\sup_{\boldsymbol{\vartheta}\in\mathcal{P}} D(P(\mathbf{X}|\boldsymbol{\vartheta}))\|P(\mathbf{X}|\boldsymbol{\theta}) \right) \tag{2.14}$$

  There is an equilibrium in this game if the risk equals the maximum entropy $H_{max} = \sup_{\boldsymbol{\vartheta}\in\mathcal{P}} H(\mathbf{X}|\boldsymbol{\vartheta})$.

Some of the concepts we refer to are described in Sect. 3.1. It turns out that many of the probability distributions in statistics are truly maximum entropy distributions given a particular type of constraint (Kapur, 1990). For example, the Gaussian is the maximum entropy distribution given the constraints upon the first two momenta: the mean and the standard deviation. The uniform distribution is the maximum entropy distribution given the constraint of the boundaries.

Therefore, non-Bayesian statistics seeks to minimize the loss with a model based on maximum entropy distributions. The distributions are hence the carriers of timidity and uncertainty. If the distributions are too flexible, however, the minimum loss model may overfit the data. On the other hand, MaxEnt seeks to maximize the loss with a model based on constraints. Thereby, the constraints are the carriers of boldness that connect the model with the data. It is possible to overfit the data by providing powerful constraints. Statistical and MaxEnt approaches both seek to balance boldness and timidity, but approach the problem from opposite directions.

2.2.5 Uncertainty about the Model

It is commonly assumed that the result of learning is a single model. However, several of the approaches in Sect. 2.2.4 actually result in models that are not singular and crisp, but vague. Vagueness does not imply a set of models without any organization. Instead, vagueness is a well-defined coexistence of models, where each of them has a certain probability.

Fiducial Vagueness

In the fiducial perspective (Fisher, 1935), a model $\boldsymbol{\theta}$ is given, along with a particular sample $\mathcal{D}$. The model $\boldsymbol{\theta}$ can be used to generate other samples of the same size as $\mathcal{D}$. If the parameters were estimated from these samples, they would not all be equal to $\boldsymbol{\theta}$ but would be distributed around it. A sample of size $n = \mathcal{D}$ is drawn from a normally-distributed population with an unknown mean μ, while $\bar{x}$ and s are estimated from the sample as:

$$\bar{x} = \frac{1}{|\mathcal{D}|} \sum_{x \in \mathcal{D}} x \tag{2.15}$$

$$s^2 = \frac{1}{|\mathcal{D}| - 1} \sum_{x \in \mathcal{D}} (x - \bar{x})^2 \tag{2.16}$$

$$t = \frac{(\bar{x} - \mu)\sqrt{|\mathcal{D}|}}{s} \tag{2.17}$$

It is known that t has a specific distribution (the Student distribution or t-distribution) which in this case depends only on a single parameter, $n = |\mathcal{D}|$. The implication is that *the estimates of a parameter based on a finite sample will be distributed around the true value under a specific distribution.* The underlying assumption is that the sample is indeed from that specific distribution, something that might be unfounded.

Significance testing is thus performed by postulating a particular null model, for example, that attributes X and Y are uncorrelated. If random samples are created from such a null model, and the correlation coefficient is estimated, the correlation coefficient will rarely be zero! Even the correlation coefficient of 1 or -1 could be obtained from the null model, but increasingly rarely with increasing sample size. Hence, we have to decide between the unlikeliness of the sample or the inappropriateness of the null model. It is always possible that the large difference between the estimate and the null model is purely due to a coincidence. In fact, we are biased towards explaining the difference as a coincidence, except when the probability of such or greater a coincidence is reasonably low, such as 5% or 1%.

Fiducial vagueness can also be used to obtain confidence intervals. Confidence intervals for parameters can be derived from these notions. For example, the 95% range of μ would be would be bounded by the μ at the 2.5%-th and the μ at the 97.5%-th percentile of the t-distribution. However, this characterization is logically dangerous because the μ is unknown: often $\bar{x}$ is used in the place of μ by trusting the unbiasedness and low variance of the estimator. The original approach to fiducial inference (Fisher, 1930) was based on asking questions such as: For which μ_1 the 97.5% of samples of such size would result in an estimate of mean lower than $\bar{x}$? For which μ_2 the 97.5% of samples of such size would result in an estimate of mean larger than $\bar{x}$? This way, we can bound the value of μ, even though the probabilities involved do not carry a consistent meaning.

Frequentist Vagueness

The Neyman-Pearson approach to hypothesis testing (Berger, 2003) seeks to fulfill the *frequentist principle*: "In repeated use of a statistical procedure, the long-run average error should not be greater than (and ideally should equal) the long-run average reported

error." This notion corresponds to calibration of probabilities involved in hypothesis testing.

In the frequentist approach, there are two models, the *null* $\boldsymbol{\theta}_0$ and the *alternative* $\boldsymbol{\theta}_1$. There is also a sample $\mathcal{D}$. The question that frequentist hypothesis testing tries to answer is whether $\mathcal{D}$ came from $\boldsymbol{\theta}_0$ or $\boldsymbol{\theta}_1$. Normally, we compute the loss of the alternative and the null, and classify the sample based on the loss. Often, the decision rule is that $\mathcal{D}$ is from the null if the loss of the null is less or equal to a specified critical level c, and from the alternative otherwise. Such classification of $\mathcal{D}$ might not be correct, and there are two kinds of error:

- Type I error probability α refers to classifying the alternative as true when it is not. This situation is referred to as false rejection, and the corresponding α is the *significance level* of the test;
- Type II error probability β refers to classifying the null as true when it is not. The situation is referred to as false acceptance, and the corresponding β is the *power* of the test.

The significance level α and the power β are computed for a particular pair of hypotheses along with the decision rule. It is not specified what the 'long-run' means, so the frequentist approach forms independent samples of a specified size from either the alternative or the null hypotheses (usually the null), performing the classification with the test. From these samples, the probability of misclassification can be estimated. Usually, the significance level α is fixed, but the power β can be computed.

Bayesian Vagueness

In the fiducial approach, the data is seen as vague: a single model corresponds to a large number of data sets, and the estimates from those data sets can be used to characterize the behavior of the original estimate. The frequentist approach is usually also made tractable by fixing the null and the alternative hypotheses. There are numerous paradoxes that arise from the above assumptions. For example, α and β depend on how the data sets are simulated. If a data set can also be cast from $\boldsymbol{\Theta}_1$, not just from $\boldsymbol{\Theta}_0$, the α and β will be affected. For that reason, the nonparametric resampling approaches such as cross-validation and the nonparametric bootstrap can be used (Davison and Hinkley, 1997) instead. These resampling approaches do not form samples from the hypotheses, but instead form samples that are subsets of the original data set.

The fundamental difference of the Bayesian approach (Gelman et al., 2004b, Bernardo and Smith, 2000) is that while the frequentist and fiducial approach view the data as uncertain but the model as fixed and certain, the Bayesians view the data as certain and fixed, while a number of models can be consistent with the data. The results of both paradigms may be quite similar. The interpretations, however, are subject to various disputes: the Bayesian paradigm requires the prior, which frequentists view as subjective, and the frequentist paradigm requires the prior assumption of a fixed true model, which Bayesians view as pretentious.

Epicurus' principle of indifference states (Kirchherr et al., 1997): *Keep all hypotheses that are consistent with the facts.* These hypotheses form an *ensemble model.* Therefore, instead of making an arbitrary selection, one could perform a combination. Consistency is

not a binary decision: in a probabilistic context several models have a non-zero posterior belief, meaning that they are all consistent, to some extent.

In Bayesian inference, the prior is an essential element of picking a model family or a hypothesis space. Some find the priors arbitrary and subjective, but the choice of the model family is inherently subjective too. The key is in choosing a prior and a model family that the research community will find acceptable. The result of the inference is a posterior belief distribution over the models of the hypothesis space given the data, or the posterior. The Bayesian posterior is Epicurean, as it is actually a belief distribution over the parameter values: all the models with non-zero likelihood and non-zero prior belief are to some extent consistent. It is usually impossible to pinpoint a particular model, but it is possible to characterize this distribution. Only when there is a lot of data, it is possible to use a single model as a reliable characterization of the whole distribution. Such a point-characterization can be the MAP or the expected posterior.

In frequentist inference, the choice of the model family and the estimator corresponds to choosing a prior. The estimators are expected to have certain desirable properties, such as unbiasedness and low variance. In the second step of inference, it is assumed that the estimated model is true, and the fiducial and frequentist vagueness can then be characterized. This step is sometimes problematic, especially in circumstances with insufficient data. For example, a coin is tossed five times, and three times it was a head, while two times a tail. The estimated probability for heads under the binomial model is $\frac{3}{5}$. The very assumption of the truth of this estimate is quite misleading.

Even if the Bayesian posterior distribution is vague, the predictions made with it can be crisp. This is important because we cannot make a series of predictions in practice, but only a single one. We can interpret the model $\boldsymbol{\Theta}$ as a *nuisance attribute*, and integrate it out:

$$\hat{P}(\mathbf{Y}|\mathcal{D}) = \int_{\Re_{\boldsymbol{\Theta}}} \hat{P}(Y|\boldsymbol{\Theta})\hat{P}(\boldsymbol{\Theta}|\mathcal{D})d\boldsymbol{\Theta} \tag{2.18}$$

This is sometimes referred to as the *Bayesian model averaging* (Hoeting et al., 1999). Certain approaches to ensemble learning in machine learning, such as *bagging* (Breiman, 1996) can effectively be seen as frequentist equivalents to Bayesian model averaging: for each resample drawn from the original data set, a crisp model is estimated. All these models are then averaged for making the final prediction. Other ensemble approaches to machine learning, such as boosting (Freund and Schapire, 1997), differ from the idea of model averaging: as iterative reweighting of instances is employed in forming an additive model.

Let us consider the familiar example of the coin toss. We start with some prior belief about the coin's probability: the coin may be biased, or unbiased. We can represent this belief by saying that our *prior* is an ensemble of all possible Bernoulli parameter values, and our belief in each parameter is equal (Fig. 2.7, left panel). Then we toss the coin five times, and the tally is 3 tails and 2 heads. The resulting ensemble reflects this (Fig. 2.7, middle panel): those probabilities that indicate that the coin always falls heads are impossible, and the most likely is the parameter value that claims that the probability of heads is $\frac{2}{5}$. The data has narrowed the range of our beliefs about the probability. It would be improper, however, to claim that this single parameter value is representative of the coin: we have not seen enough data to be so specific. All we can say is that we believe that the probability of heads is in the interval $[0.1, 0.8]$. Performing a few more tosses, we end up with the tally of 9 heads and 10 tails. The distribution of our beliefs over the

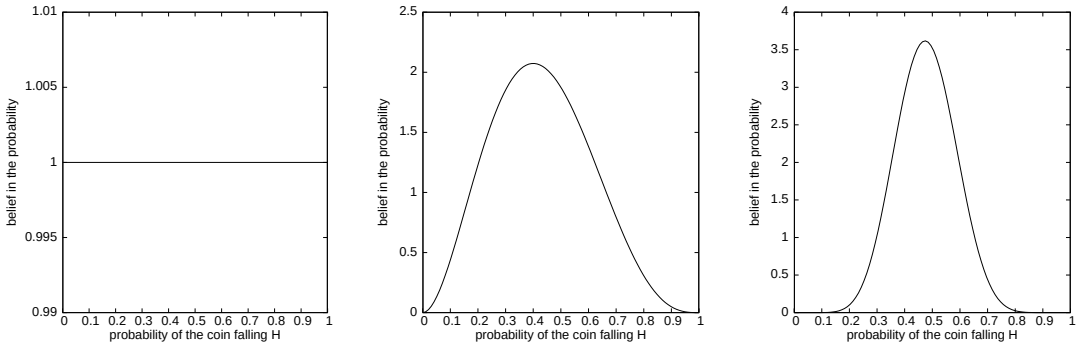


Figure 2.7: A Bayesian ensemble of models. Each probability of the unknown coin falling heads is an individual parameter value, and these parameter value form an ensemble. Each parameter value is assigned a particular belief. Our prior belief is uniform over all the probabilities. Successive observations of coin toss outcomes induce greater and greater precision in our beliefs about the posterior belief (left to right). Still, there is always some uncertainty about the exact probability.

ensemble (Fig. 2.7, right panel) shows that the probability is almost certainly somewhere on $[0.2, 0.8]$, but we cannot yet say anything beyond that with complete certainty.

When such an ensemble is used to make a prediction, each parameter value makes a distinct prediction. This way, we obtain an ensemble of predictions, each of them weighted by the posterior belief in the corresponding parameter value. We can interpret the ensemble as an *imprecise prediction*: not just that the ensemble is not sure about the outcome, it is also unsure about the probability. The other way of interpreting the ensemble is by stating that the identity of the parameter value is a nuisance parameter, a property that exists but we do not want to know. The Bayesian approach for dealing with nuisance parameters is to average the predictions of all parameter values, so that each prediction is weighted by our belief in the parameter value that yielded it.

If we consider the parameter value as a nuisance parameter, we need not treat the model as an ensemble: it is somewhat expensive to lug along all the parameter values and their individual worths. Instead, we may average them together. In this case, we can represent the average as a single model being guided by the following probability:

$$p^H_{BMA} = \frac{n_H + 1}{n_H + n_T + 2}$$

This is referred to as the *Laplace estimate of probability*, because the legend says that Laplace wondered what is the probability of seeing another sunrise after having seen only a single one. Of course, in some applications it is important to keep note of the whole ensemble: p^H_{BMA} is identical for the tally of 1 head and 1 tails and for the tally of 10000 heads and 10000 tails. However, the ensemble is much more distinctly peaked for the latter one. Averaging, therefore, is a way of replacing the Epicurean ensemble with a single parameter value that is closest to the average of the ensemble, but any single parameter value from the ensemble does not faithfully represent the *variation* in the ensemble.

2.2.6 Parametric vs Nonparametric Learning

One way of dividing the hypothesis spaces is into the *parametric* and the *nonparametric*. Some definitions (Bernardo and Smith, 2000) associate nonparametric models with infinite dimensionality of the parameter space, and parametric models with finite dimensionality.

The parameters in some nonparametric models are placed in the same space as the instances: for each possible instance, there is a distinct probability in the model. In the coin toss example, we associate a probability with each of the possible instances or events (heads or tails). This probability is itself the parameter. This approach is extremely flexible on finite instance spaces, but becomes unwieldy when there are infinitely many possible instances.

When the instance space has very high cardinality, we can avoid working with a parameter for each of the possible instances. However, we can assign a parameter to each of the instances from the data set. An example of such parametrization are the support vector machines (Schölkopf and Smola, 2002), where each instance is associated with a parameter (Lagrange multiplier) that identifies the weight of the instance. Support vector machines are usually combined with parametric kernels, and these parameters are not placed in the instance space.

Somewhat differently, nonparametric test procedures are those that are unconcerned about the parameters of the distribution (NIST/SEMATECH, 2002). This means that the test procedure is not trying to answer questions about the parameters of the model, but about other properties that are only indirectly related to the parameters, such as independence, factorization, and so on. A related category of distribution-free test procedures is based on statistics that are not dependent on the form of the underlying model. We will discuss some distribution-free and nonparametric test procedures in Ch. 4.

2.3 Probability

In the previous sections we have described models, hypothesis spaces and data, and identified probability as a way of allowing for uncertainty, both of predictions and of models. We will now provide a more specific account of the assumptions of probability. We will also address some of the criticisms of probability. On the other hand, we will not discuss various interpretations of probability: our applications of probability are compatible with several interpretations, but there may definitely be interpretations incompatible with our applications.

We need to formalize the notion of a 'model'. To do this, we will use two concepts: the universe and the attribute. A universe is a collection of possibilities (sun, clouds, rain), while probability measures the likelihood of each of them (sun: 0.7, clouds: 0.2, rain: 0.1). On the other hand, an attribute wet/not-wet is a shortened projection of the universe (wet:(rain), not-wet:(sun,clouds)). Using attributes, we can condition a universe, split it into separate subuniverses, one for each value of the attribute (wet:(rain:1), non-wet:(sun:0.78, clouds:0.22)). Alternatively, we may marginalize a universe by collapsing all events that cannot be distinguished with the given set of attributes (wet:0.3, non-wet:0.7). The following subsections are intended to be an informal introduction to mathematical probability. A reader who desires a more formal approach should refer to other literature, such as (DeGroot and Schervish, 2002).

2.3.1 Universes

A *probability mass function* (PMF) defines the probability of each event. When we have several PMFs, we assure their cohesion by having them all derived from an underlying *universe*. The universe is a measure space $\langle S, \mathcal{E}, P \rangle$ based on a discrete set of elementary events $\mathcal{E} = \{e_1, e_2, \ldots, e_n\}$. The set of events is sometimes also referred to as sample space in probability theory, or an alphabet. Note that events may be letters, symbols, things, entities, objects in a bag, states of some machine, outcomes of an experiment, or words in a document: events are merely the carriers of distinction. The formal term for a universe along with probability is a *probability space*, but information theory refers to probability spaces with discrete events.

The *probability* P is a measure of each event in the universe. The probabilities for all these elementary events should sum up to 1: $\sum_i P(e_i) = 1$. Therefore, in every circumstance exactly one of the events should happen. The assumption that elementary events be mutually exclusive is sometimes found problematic, but is easily remedied. One frequent example is the case of the 'excluded middle'. Exactly one of a and $\neg a$, where $\neg a$ signifies not-a, is true at the same time. For example, if a signifies a full cup, and $\neg a$ an empty cup, this appears to be a problem. But it is not a problem of assumptions, but of the representation: saying that $\neg a$ marks an empty cup is incorrect, as a cup can be neither full nor empty. More appropriate would be a larger set of four events, based on a signifying a full cup and $\neg a'$ an empty cup: $\{a \wedge \neg a', a \wedge a', \neg a \wedge a', \neg a \wedge \neg a'\}$, here $\wedge$ stands for logical conjunction. It is then the task of the probability to capture the semantic mutual exclusivity of emptiness and fullness: $P(a \wedge a') = 0$, but we could have excluded this joint event when defining the events.

Another problem that may arise with probability are unforeseen circumstances. What happens if we get a broken cup: is it full or empty? Indeed, in some situations we need to create a ghost event e_0 which means 'something else' or 'something unforeseen'. Also, it would be incorrect to use a probability larger than 1 to describe an event that has happened several times: this is achieved by creating multiple events $\{a^1, a^2, a^3, \ldots\}$. As these events are mutually exclusive, we have 'invented' natural numbers.

The events and probabilities are considered to be pure and objective. We do not concern ourselves with the notion of an observer and the observed. If this is necessary, the act of observation should be included among the events. For example, if a signifies the sun rising, and b me observing the sunrise, the universe should be modelled as four events: $\{a \wedge b, \neg a \wedge b, a \wedge \neg b, \neg a \wedge \neg b\}$. If the model should allow for the truth of my claims about the sunrise, a further symbol c would need to be combined with a and b, and would signify what I claimed about the sunrise. Therefore, these three events capture the situation in its full scope: a - truth, b - what I see as true, and c - what I say.

There is a difference between *observing* an event a and *performing* the event a. Therefore, for each event a that is under our control, simply a denotes the event happening, and $a^\dagger$ denotes the intervention: the causing of event a (Pearl, 2000). There are four possible events: $\{a \wedge a^\dagger, \neg a \wedge a^\dagger, \neg a \wedge \neg a^\dagger\}$. If our interventions are sufficient, $P(\neg a \wedge a^\dagger) = 0$. If our interventions are necessary, $P(\neg a \wedge \neg a^\dagger) = 0$.

It is obvious that our model is of limited precision: we cannot break any event into its constituent parts - the events are atomic and internally indistinguishable. Should we want to do that, we would need a new sample space, a new universe. The universe is the embodiment of the notion of an ontology: the list of conceivable events along

with the possibility, impossibility and probability of each one of them. If one prefers the logical hypothesis space, the set of events is the set of possible atomic statements, the probability of each event is their semantics, so that each resulting statement has a probability. The mathematical structure of a sample space generalizes upon this notion by allowing aggregations of elementary events.

The choice of the universe is in the eye of the beholder. The beholder only distinguishes those nuances that matter. There are unimaginably many possible states, but as beholders, we choose not to distinguish all of them. We might distinguish 37.001 and 37.002 as abstract numbers, but we would generally not distinguish them if they indicated the body temperature as one attribute in medical diagnosis. On the other hand, 37.049 and 37.051 would be distinguished in the universe where rounding to the nearest number turned them into 37.0 and 37.1, but not in another universe where all numbers are rounded down. We avoid associated problems by allowing for a number of universes that model the same reality: ultimately the choice of the universe is an event like any other. Furthermore, we may have several probability measures for the same universe: each choice of a probability measure is an event. Finally, all that we truly require is that the probabilities are consistent within a particular universe, and that universes can be coalesced into a single universe which agrees with the above assumption of mutual exclusivity and completeness.

It is also possible to model dynamics with the concept of the universe. Given a static universe $\mathcal{E}$, the dynamic universe is a Cartesian product of the universe before and the universe after: $\mathcal{E}_{before} \times \mathcal{E}_{after}$. The implicit time of the dynamic universe is also discrete: 'before' and 'after' are distinctly separated. At the same time, the model is unable to account for its possible changes through time: it is necessarily invariant with respect to translations in time. The invariance of some kind, with respect to moments, types of cups, translations in time or something else, facilitates the repeatability of a particular event. Multiplicity or repeatability of occurrence of an event, or at least belief in the occurrence of an event is what is needed to speak about probability. A 'thick time' model of the universe would be $\mathcal{E}_0 \times \cdots \times \mathcal{E}_{now}$, but only ignorance or multiplicity of universes (multiverse) would allow probability.

The data $\mathcal{D}$ is represented as a multiset of events, or as a set of *instances* or measurements: a single event may have happened several times and so corresponds to several instances, just the same temperature can be obtained through several acts of measurement. This means that the universe may not distinguish every pair of instances, either due to ignorance or intentional disregard. There is no ordering of instances, unless the order is a part of each event. Many possible probability measures are consistent with a given set of data: the only requirement is that each instance has non-zero probability.

It is possible to *learn* the probability from the data, too: we can seek the probability assignments that make the data as likely as possible (Fisher, 1912). Or, more generally, we can use the Bayes rule to assign probabilities to different probability measures consistent with the data, e.g. (Good, 1965, Jaynes, 2003), thereby creating a universe of probability measures. In some cases it is necessary to interpret the data probabilistically, especially with unreliable sensors or with real-valued measurements. The temperature reading of 37.0 degrees Celsium may be interpreted as an observation that the true temperature has a uniform distribution between 37.05 and 37.15 degrees Celsium: an additional source of uncertainty. Not to get bogged down in this complexity, we will always consider a single universe with a single probability measure. However, if this universe is nested as an event

within another universe, so will every statement or conclusion based on it.

2.3.2 Attributes

We have interpreted the universe as a formalization of everything that can be distinguished. There is no structure in the universe, it is a mere structureless list. We will now consider the notion of an *attribute* A as a construct built on top of the universe. We will start with a binary attribute, the simplest of all, whose *range* $\Re_A$ is $\{0, 1\}$. The binary attribute A is a function $A : \mathcal{E} \to \Re_A$. Thus, for each event, we will know whether the attribute took the value of 0 or 1. The attribute merges all the states of the universe into those that have the value of 0 and those that have the value of 1: the attribute values are mutually exclusive. By summing all the corresponding event probabilities, we can obtain the attribute value probabilities. We can also envision a universal attribute whose range is the universe itself: the universe itself is then the original attribute, the alphabet. More formally, an attribute is a random quantity, and each attribute value corresponds to an element of the event space in probability theory. The attributes whose range is the set of real numbers $\mathbb{R}$ are sometimes referred to as random variables, and that is why we are using the term 'attribute' and not 'random variable'.

There are a few arguments against attributes. First, fuzzy logic (Zadeh, 1965) disagrees with the notion of an attribute which takes a single crisp value for each event. Instead, fuzzy logic recommends using grades of membership of an attribute value for each event. We will attempt to do the same in the existing framework by introducing the notion of a 'perception' or a *sensor*. The sensor is unreliable, and may or may not react to a particular event. But this can easily be handled within our notion of events and attributes. As earlier, we will include the sensor reading $\{s, \neg s\}$ into the universe, obtaining four events: $\{a \wedge \neg s, a \wedge s, \neg a \wedge s, \neg a \wedge \neg s\}$. If the sensor is precise, $P(a \wedge \neg s)$ and $P(\neg a \wedge s)$ will be low. Nevertheless, there is a good reason why sensors should not always be precise: consider a indicating the height of $183.2321\ldots$ centimeters and the s signifying 'tall': there is a good reason for working with clumpier s rather than with a. Of course, if we have several heights and several sensors, the situation of sensors 'tall' and 'very tall' both taking the value of 1 for the same objective height is perfectly possible. When there are k mutually exclusive binary attributes, meaning that for each event in the universe there is exactly one of them taking the value of 1, we may replace them all with a single k-ary attribute with the range $\{1, 2, \ldots, k\}$. This is a major gain in economy, but it is contingent upon mutual exclusivity.

Another common assumption is the invariance of a sensor: it should remain the same for all instances, in the same way as an event is atomic. This assumption is not always realistic: there may be drift through time (Widmer and Kubat, 1996), old sensors may not be the same as new sensors and consequences once upon the time are no longer the same. A systematic deviation from this assumption cannot be captured by the model, and the resulting model will carry some uncertainty because of that. The solution lies in introducing a sensor's sensor, indicating if the sensor is new, old or broken. And one can continue by including the sensor of a sensor's sensor.

In other circumstances, the value of the attribute might be unknown or undefined. Assume patients coming to a physician: each patient is an event. For some patients, the body temperature is known, but for others it is not. Technically, the attribute's range must then include 'not known' as one of its values. Even in the binary case, we can imagine the

range {'1', '0 or unknown'}. Using the methods of *imputation* we can guess the missing value, but this assumption must sometimes be verified. For example, the sentences in a censored document have not been deleted by random, and people do not say "I do not know" purely at random.

Alternatively, we may create a binary attribute {'temperature known', 'temperature unknown'}, and *condition* the universe to contain only those patients whose temperature is known. In that conditional universe, the temperature is always known. This conditioning is always implicit: the patients themselves are conditioned on the binary attribute 'The event is a patient coming to a physician.' in the first place. The probabilities in each branch of a conditional universe sum up to 1.

The second kind of operation is *marginalization*. Here, we take a set of attributes, for example $\{A, B, C\}$, and collapse all events that cannot be distinguished with these attributes into elementary ones. For example, if we marginalize away all colors except for two $A : \{\text{black}, \text{not-black}\}$ and $B : \{\text{white}, \text{not-white}\}$, every color will be mapped either to black, white or gray (not-black and not-white). Furthermore, zero probability is put in effect for each combination $\langle a, b, c\rangle$ of attributes' values, $\langle a, b, c\rangle \in \Re_A \times \Re_B \times \Re_C$, that cannot be found in the universe (such as 'black and white'). In the example of physician's patients, the attribute 'astrological signs' has been marginalized away and is not known or used by the physician (presumably). On the other hand, 'body temperature' is usually marginalized away in the discourse of an astrologer. In all, contemporary medicine generally assumes that all people are equal, and this assumption both allows generalizing from one patient to others, but also prevents distinguishing specific characteristics of patients. Some theories of probability claim that that probability is purely a result of marginalization and a consequence of the fact that the causes are not known. In summary, the *marginal probability distributions* are projections of the joint probability distribution where we disregard all attributes but a subset of them, for example:

$$P(A) = P(A, \cdot, \cdot) = \sum_{b \in \Re_B} \sum_{c \in \Re_C} P(A, b, c).$$

In all, we see that attributes can be seen as projections of the universe, as views of the universe. Marginalization serves as integration, as merging of events, and probability reflects this merging. On the other hand, conditioning creates separate universes, each of them with a consistent definition of probability. The universe serves as a unified foundation for defining the relationships between attributes, and in turn, these attributes serve as means for characterizing the events. It is possible to construct or remove attributes as deemed suitable, and these attributes will transform the perception of the universe.

2.3.3 Probability: Frequency vs Belief

In many circumstances it is impossible to predict the outcomes exactly. I take a coin and toss it, but even if I try, I cannot perfectly control the outcome. If it is not possible to reliably predict the outcome, we can still reliably predict the *probability* of each outcome. For example, we could say that there is a 50% probability of the coin falling heads, and a 50% probability of the coin falling tails.

There are numerous interpretations of the meaning of probability, but a particularly important division is into the *frequentist* probability on one hand, and the interpretation of probability as a *degree of belief* on the other hand. Objective frequentist probabilities

are a part of the ontology, and they refer to reality. On the other hand, subjective beliefs arise from our limited knowledge about the world, and are an aspect of epistemology. The worldview with frequentist probability takes the reality as inherently unpredictable, but guided by a true model. The true model is identifiable, should an infinite number of observations be made. Probability is defined through the long-run frequency of an event. Learning is referred to as estimation, and seeks to minimize the fixed utility or risk.

On the other hand, the subjective view considers probabilities as resulting from the lack of knowledge. The coin toss, for example, appears random purely because the conditions of each experiment are not precisely controlled. Learning is referred to as inference. The probability is thus seen just as a way of representing the *degree of belief*, the ignorance, the inability or reluctance to state a specific model. Probability as belief refers to statements in a language, not to objects in the world. It is the model that is unable to predict the outcome, perhaps due to bad quality of the data, not due to the inherent unpredictability of the actuality. An ideal observer with all the information would be able to get a model with less uncertainty. A purely subjective interpretation of an unpredictable quantum phenomenon tolerates both options: either we do not know what is inside, or that the inside is inherently unknowable. The process of learning seeks to maximize the utility of the model, but the utility and the probability are dependent and inherently entangled (Rubin, 1987). It is possible, however, to use proper score functions and noninformative priors that favor probabilities that are calibrated and have good properties with respect to the frequentist criteria.

CHAPTER 3

An Information-Theoretic View of Interactions

3.1 Basics of Information Theory

In the previous section we have described the three crucial elements needed to discuss entropy: the universe $\mathcal{E}$, the probability P and the attributes $A, B, C, \ldots$. We can now begin to disentangle the model with information theory. We will show the connection between entropy, investment and growth. In the second subsection, we will justify other information-theoretic expressions through questions we can ask about the truth. We will use the results and expressions of (Shannon, 1948).

Let us examine an attribute, A. Shannon's entropy measured in bits is a measure of its unpredictability:

$$H(A) \triangleq -\sum_{a \in \Re_A} P(a) \log_2 P(a) \tag{3.1}$$

By definition, $0 \log_2 0 = 0$. The higher the entropy, the less reliable are our predictions about A. We can understand $H(A)$ as the amount of uncertainty about A, as estimated from its probability distribution.

Another key concept is the *Kullback-Leibler divergence* or relative entropy (Kullback and Leibler, 1951). It is a measure of divergence between two probabilistic models P and Q, both defined on the same range $\Re_X$.

$$D(P\|Q) \triangleq \sum_{x \in \Re_X} P(x) \log_2 \frac{P(x)}{Q(x)} \tag{3.2}$$

The unit of measure is a bit. KL-divergence has also been referred to as the 'expected log-factor' (logarithm of a Bayes factor), expected weight of evidence in favor of p as against q given p, and cross-entropy (Good, 1963). KL-divergence is zero only when the two functions are equal. It is not a symmetric measure: P is the *reference* model, and the KL-divergence is the expected loss incurred by the *alternative* model Q when approximating P. We can understand empirical entropy through KL-divergence.

Using conditional KL-divergence, it is possible to compare two conditional probability models, something particularly useful in supervised learning, when Y is labelled, and X is not:

$$D(P(Y|X)\|Q(Y|X)) \triangleq \sum_{x\in\Re_X, y\in\Re_Y} P(y,x)\log_2\frac{P(y|x)}{Q(y|x)} \tag{3.3}$$

This way, we compare how well the model Q approximates the true distribution of Y in the context of X, $P(Y|X)$. Observe, however, that the conditional KL-divergence cannot be computed without a joint probability model of $P(X,Y)$.

3.1.1 Interpretations of Entropy

It is extremely important to note that our universe is a model. It is not necessarily a true model of reality, but of a partial view of reality. It is the goal of statistical mechanics to provide a good model of reality through probabilistic modelling, but we can use the same tools to model anything, such as patients entering a doctor's office. And in such circumstances there is little similarity between Shannon's entropy and Boltzmann's 'quantity called H' (Tolman, 1979) which refers to molecules of gas. In retrospect, it was not a good decision to call Shannon's entropy entropy: a more appropriate term would be neginformation. For contrast, we will now present two interpretations of entropy that manifest its decision-theoretic and game-theoretic nature.

Entropy as Loss

We can express entropy and divergence in the terms of loss functions. Consider that the player whose betting portfolio is q. He suffers the loss of $-\log_2 q(e)$ in the case of the event e. This means that we have a loss function $L(e,q) = -\log_2 q(e)$: the less the player bet, the more he lost. This specific loss function is used in data compression, where we pay each symbol proportionally to the logarithm of the probability with which we predicted it, with the number of bits. Data compression programs, such as *zip*, are nothing else than successful probabilistic gamblers.

The *expected loss* is the expectation of player's loss. The player is using an imperfect model of reality with q instead of the true probability P. The Shannon entropy corresponds to the minimum expected loss, suffered by the omniscient player: $H(\mathcal{E}) = \inf_q \mathbb{E}_{e\sim P}\{L(e,q)\}$. The KL-divergence thus corresponds to the player's expected loss beyond the omniscient player's: $D(P\|q) = \mathbb{E}_{e\sim P}\{L(e,q) - L(e,P)\}$. We could also understand these expressions as definitions of entropy and divergence based on some loss function. Entropy and divergence are just specific definitions of loss and gain, and we may introduce quantities corresponding to entropy and divergence with different definitions of this loss (Grünwald and Dawid, 2004). Of course, not all properties would be retained.

Entropy and the Rate of Growth

We will now consider the definition of entropy through gambling, following a popular information theory textbook (Cover and Thomas, 1991). Assume that we are playing a game, trying to predict what event will take place. We start with K coins, and place a bet on each of the events in the universe, expressing it as a proportion of K. So for event e_i, our bet is $b(e_i)$, while $\sum_{e\in\mathcal{E}} b(e) = 1$. We now let some event e' happen, and our gain

is $MKb(e')$, where M is the maximum reward multiplier: had we bet everything on e', $b(e') = 1$, our funds would increase M-fold. Therefore, our funds multiply by $Mb(e')$.

Clearly, we would achieve the highest growth of funds by putting all the money on the single most likely event, but would also lose everything if that event did not happen. Alternatively, we minimize the chances by betting on every outcome equally, but if there are too many possible events, we would be losing in every game. It can be shown that the maximum rate of growth out of all possible betting portfolios is achieved by betting proportionally to event probabilities, so that $P(e) = b(e)$, and this is called the Kelly gambling scheme. The doubling rate of the horse race using the proportional gambling is $\log_2 M + \sum_{e\in\mathcal{E}} P(e)\log_2 P(e)$. It is easy to see that for an omniscient player the game is worth playing only if $\log_2 M > H(\mathcal{E})$, or in other words, if the logarithm of the rewards exceeds the *information entropy* of the universe. Of course, it is impossible to stop playing with reality.

Realistic observers, however, are not omniscient, and their portfolio b deviates from the true distribution P. For them, the doubling rate is $\log_2 M - H(\mathcal{E}) - D(P\|b)$, where $D(P\|b)$ is the Kullback-Leibler divergence or relative entropy between the truth P and their belief b. It is important to understand that a linear change either in entropy or in KL-divergence corresponds to a linear change in the rate of growth. Entropy is the minimum rate of growth for an omniscient predictor. Furthermore, the rate of growth or demise is essentially linked with the ability to place bets well. The same conclusion is valid also if a different M is specified for each event $m(e)$, only the $\log_2 M$ would be replaced by $\sum P(e)\log_2 m(e)$.

3.2 Entropy Decompositions

3.2.1 Entropy Calculus for Two Attributes

In addition to the attribute A, let us now introduce a new attribute, B. We have observed the joint probability distribution, $P(A, B)$. We are interested in predicting A with the knowledge of B. At each value of B, we observe the probability distribution of A, and this is expressed as a conditional probability distribution, $P(A|B)$. Conditional entropy, $H(A|B)$, quantifies the remaining uncertainty about A with the knowledge of B:

$$H(A|B) \triangleq -\sum_{a,b} P(a,b)\log_2 P(a|b) = D(P(B)\|P(A,B)) = H(A,B) - H(B) \tag{3.4}$$

We quantify the 2-way interaction between two attributes with *mutual information*:

$$\begin{aligned} I(A;B) &\triangleq \sum_{a\in\Re_A, b\in\Re_B} P(a,b)\log_2 \frac{P(a,b)}{P(a)P(b)} = D(P(A,B)\|P(A)P(B)) \\ &= H(A) + H(B) - H(A,B) = H(A) - H(A|B) = I(B;A) = H(B) - H(B|A) \end{aligned} \tag{3.5}$$

In essence, $I(A;B)$ is a measure of correlation between attributes, which is always zero or positive. It is zero if and only if the two attributes are independent, when $P(A,B) = P(A)P(B)$. Observe that the mutual information between attributes is the *average* mutual information between values in attributes' ranges.

If A is an attribute and Y is the labelled attribute, $I(A;Y)$ measures the amount of information provided by A about Y: in this context it is often called *information gain*.

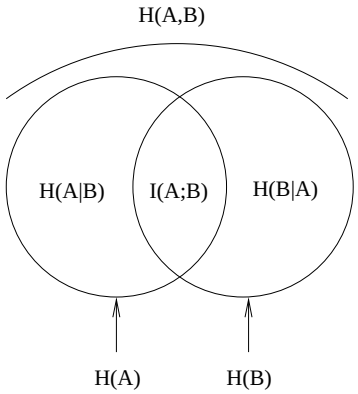


Figure 3.1: A graphical illustration of the relationships between information-theoretic measures of the joint distribution of attributes A and B. The surface area of a section corresponds to the labelled quantity. This illustration is inspired by Cover and Thomas (1991).

It is easy to see that the information gain corresponds to the prediction error made by predicting the label without the information of the attribute as assessed with the Kullback-Leibler divergence:

$$I(A;B) = D(P(A|B)\|P(A)) = D(P(B|A)\|P(B)) \tag{3.6}$$

A 2-way interaction helps reduce our uncertainty about either of the two attributes with the knowledge of the other one. We can calculate the amount of uncertainty remaining about the value of A after introducing knowledge about the value of B. This remaining uncertainty is $H(A|B)$, and we can obtain it using mutual information, $H(A|B) = H(A) - I(A;B)$. Sometimes it is worth expressing it as a percentage, something that we will refer to as *relative* mutual information. For example, after introducing attribute B, we have $100\% \cdot H(A|B)/H(A)$ percent of uncertainty about A remaining. For two attributes, the above notions are illustrated in Fig. 3.1.

3.2.2 Entropy Calculus for Three Attributes

Let us now introduce the third attribute, C. We could wonder how much uncertainty about A remains after having obtained the knowledge of B and C: $H(A|BC) = H(ABC) - H(BC)$. We might also be interested in seeing how C affects the interaction between A and B. This notion is captured with *conditional mutual information*:

$$\begin{aligned} I(A;B|C) &\triangleq \sum_{a,b,c} P(a,b,c) \log_2 \frac{P(a,b|c)}{P(a|c)P(b|c)} = H(A|C) + H(B|C) - H(AB|C) \\ &= H(A|C) - H(A|B,C) = H(AC) + H(BC) - H(C) - H(ABC). \end{aligned} \tag{3.7}$$

Conditional mutual information is always positive or zero; when it is zero, it means that A and B are unrelated given the knowledge of C, or that C completely explains the association between A and B. From this, it is sometimes inferred that A and B are both consequences of C. If A and B are conditionally independent, we can apply the naïve Bayesian classifier for predicting C on the basis of A and B with no remorse. Conditional mutual information is a frequently used heuristic for constructing Bayesian networks (Cheng et al., 2002).

Conditional mutual information $I(A;B|C)$ describes the relationship between A and B in the context of C, but we do not know the amount of influence resulting from the introduction of C. This is achieved by the measure of the intersection of all three attributes, or *interaction information* (McGill, 1954) or *McGill's multiple mutual information* (Han, 1980):

$$\begin{aligned} I(A;B;C) &\triangleq I(A;B|C) - I(A;B) = I(A,B;C) - I(A;C) - I(B;C) \\ &= H(AB) + H(BC) + H(AC) - H(A) - H(B) - H(C) - H(ABC). \end{aligned} \tag{3.8}$$

Interaction information among attributes can be understood as the amount of information that is common to all the attributes, but not present in any subset. Like mutual information, interaction information is symmetric, meaning that $I(A;B;C) = I(A;C;B) = I(C;B;A) = \ldots$. Since interaction information may be negative, we will often refer to the absolute value of interaction information as *interaction magnitude*. Again, be warned that interaction information among attributes is the average interaction information among the corresponding values.

Interaction information has proven to be a considerably better predictor of validity of the naïve Bayesian classifier assumption in classification tasks than conditional mutual information $I(A;B|C)$. This can be apparent from the identity, remembering (3.3):

$$I(A;B;C) = D\left(P(C|A,B)\middle\|P(C)\frac{P(A|C)P(B|C)}{P(A)P(B)}\right) = D\left(P(C|A,B)\middle\|\frac{P(C|A)P(C|B)}{P(C)}\right)$$

The two right-hand models closely resemble the non-normalized naïve Bayesian classifier (NBC). This non-normalization is what yields a negative interaction information, and $I(A;B;C)$ should really be seen as an approximate model comparison (but with other convenient properties). Conditional mutual information tends to overestimate the deviation, as it is derived from a joint model comparison, and not a conditional one (Fig 3.2). However, conditional mutual information can be seen as an upper bound for the actual NBC loss.

The concept of *total correlation* (Watanabe, 1960) describes the total amount of dependence among the attributes:

$$\begin{aligned} C(A,B,C) &\triangleq H(A) + H(B) + H(C) - H(ABC) \\ &= I(A;B) + I(B;C) + I(A;C) + I(A;B;C) \\ &= D(P(A,B,C)\|P(A)P(B)P(C)). \end{aligned} \tag{3.9}$$

It is always positive, or zero if and only if all the attributes are independent, $P(A,B,C) = P(A)P(B)P(C)$. However, it will not be zero even if only a pair of attributes are dependent. For example, if $P(A,B,C) = P(A,B)P(C)$, the total correlation will be non-zero, but only A and B are dependent. Hence, it is not justified to claim an interaction among all three attributes. For such a situation, interaction information will be zero, because $I(A;B|C) = I(A;B)$.

3.2.3 Quantifying n-Way Interactions

In this section, we will generalize the above concepts to interactions involving an arbitrary number of attributes. Assume a set of attributes $\mathcal{A} = \{X_1, X_2, \ldots, X_n\}$. Each attribute

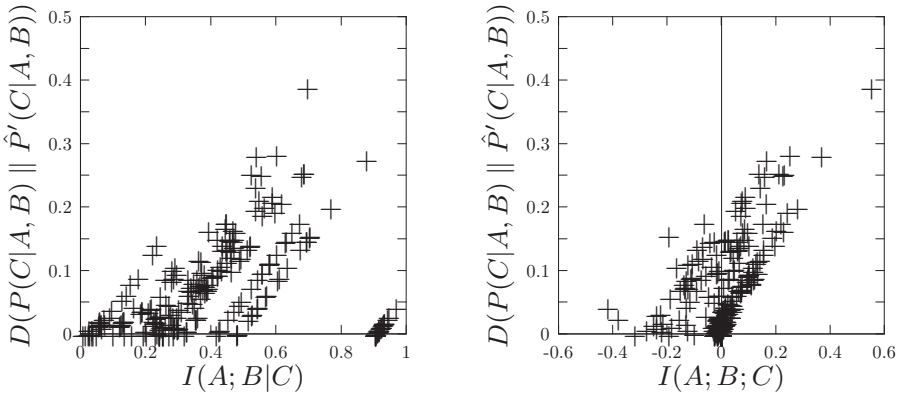


Figure 3.2: Conditional mutual information $I(A;B|C)$ is often used as the predictor of the loss caused by the conditional independence assumption in the naïve Bayesian approximation $\hat{P}'(C|A,B) \propto P(C)P(A|C)P(B|C)$. Interaction information $I(A;B;C)$ works better on this UCI Mushroom data set, as the conditional mutual information often overestimates the true loss: for some of the most conditionally dependent attributes, the actual prediction loss was quite low.

$X \in \mathcal{A}$ has a range $\Re_X = \{x_1, x_2, \ldots, x_p\}$. If we consider the whole set of attributes $\mathcal{A}$ as a multivariate or a vector of attributes, we have a joint probability distribution, $P(\mathbf{a})$. $\Re_{\mathbf{A}}$ is the Cartesian product of individual attributes' ranges, $\Re_{\mathbf{A}} = \Re_{X_1} \times \Re_{X_2} \times \cdots \times \Re_{X_n}$, and $\mathbf{a} \in \Re_{\mathbf{A}}$. We can then define a marginal probability distribution for a subset of attributes $\mathcal{S} \subseteq \mathcal{A}$, where $\mathcal{S} = \{X_{i(1)}, X_{i(2)}, \ldots, X_{i(k)}\}$:

$$P(\mathbf{s}) \triangleq \sum_{\substack{\mathbf{a} \in \Re_{\mathbf{A}}, \\ \mathbf{s}_j = \mathbf{a}_{i(j)}, \\ j=1,2,\ldots,k}} P(\mathbf{a}). \tag{3.10}$$

Next, we can define the entropy for a subset of attributes:

$$H(\mathcal{S}) \triangleq -\sum_{\mathbf{v} \in \boldsymbol{\mathcal{S}}} P(\mathbf{v}) \log_2 P(\mathbf{v}) \tag{3.11}$$

We define *k-way interaction information* by generalizing from formulae in (McGill, 1954) for $k = 3, 4$ to an arbitrary k:

$$I(\mathcal{S}) \triangleq -\sum_{\mathcal{T} \subseteq \mathcal{S}} (-1)^{|\mathcal{S} \setminus \mathcal{T}|} H(\mathcal{T}) = I(\mathcal{S} \setminus X|X) - I(\mathcal{S} \setminus X),\ X \in \mathcal{S}, \tag{3.12}$$

k-way multiple mutual information is closely related to the lattice-theoretic derivation of multiple mutual information (Han, 1980), $\Delta h(\mathcal{S}) = -I(\mathcal{S})$, and to the set-theoretic derivation of multiple mutual information (Yeung, 1991) and co-information (Bell, 2003) as $I'(\mathcal{S}) = (-1)^{|\mathcal{S}|} I(\mathcal{S})$. In this paper, we will neglect to distinguish the applications of these two formulations, except when discussing positive or negative interactions which are based on (3.12).

Finally, we define *k-way total correlation* as (Watanabe, 1960, Han, 1980):

$$C(\mathcal{S}) \triangleq \sum_{X \in \mathcal{S}} H(X) - H(\mathcal{S}) = \sum_{\mathcal{T} \subseteq \mathcal{S}, |\mathcal{T}| \geq 2} I(\mathcal{T}) = D\left(P(\mathcal{S}) \middle\| \prod_{X \in \mathcal{S}} P(X)\right). \tag{3.13}$$

We can see that it is possible to arrive at an estimate of total correlation by summing all the interaction information existing in the model. Interaction information can hence be seen as a decomposition of a k-way dependence into a sum of l, $l \leq k$ dependencies.

3.2.4 A Brief History of Entropy Decompositions

Although the idea of mutual information has been formulated (as 'rate of transmission') already by Shannon (1948), the seminal work on higher-order interaction information was done by McGill (1954), with application to the analysis of contingency table data collected in psychometric experiments, trying to identify multi-way dependencies between a number of attributes. The analogy between attributes and information theory was derived from viewing each attribute as an information source. These concepts have also appeared in biology at about the same time, as Quastler (1953) gave the same definition of interaction information as McGill, but with a different sign. Later, McGill and Quastler (1955) both agreed on using the McGill's version.

The concept of interaction information was discussed in early textbooks on information theory (e.g. Fano, 1961). A formally rigorous study of interaction information was a series of papers by Han, the best starting point to which is the final one (Han, 1980). A further discussion of mathematical properties of positive versus negative interactions appeared in (Tsujishita, 1995). Bell (2003) discussed the concept of co-information, closely related to the Yeung's notion of multiple mutual information, and suggested its usefulness in the context of dependent component analysis.

In physics, Cerf and Adami (1997) associated positive interaction information of three variables (referred to as ternary mutual information) with the non-separability of a system in quantum physics. Matsuda (2000) applied interaction information (referred to as higher-order mutual information) and the positive/negative interaction dichotomy to the study of many-body correlation effects in physics, and pointed out an analogy between interaction information and Kirkwood superposition approximation. In ecology, Orlóci et al. (2002) referred to interaction information as 'the mutual portion of total diversity' and denoted it as $I(ABC)$. Yairi et al. (1998) employed interaction information in robotics. Leydesdorff and Meyer (2003) applied interaction information to analyzing the relations between universities, industry and government, referring to it as mutual information in three dimensions. In the field of neuroscience, Brenner et al. (2000) noted the utility of interaction information for three attributes, which they referred to as synergy. They used interaction information for observing relationships between neurons. Gat (1999) referred to positive interactions as synergy, while and to negative interactions as redundance. Demšar (2002) referred to it as the relative information gain.

The concept of interactions also appeared in cooperative game theory with applications in economics and law. The issue is observation of utility of cooperation to different players, for example, a coalition is an interaction between players which might either be of negative or positive value for them. Grabisch and Roubens (1999) formulated the Banzhaf interaction index, which proves to be a generalization of interaction information, if negative entropy is understood as game-theoretic value, attributes as players, and all other players are disregarded while evaluating a coalition of a subset of them (Jakulin, 2003). Kojadinovic (2003) discusses the relationship between independence and interaction. Gediga and Düntsch (2003) applied these notions to rough set analysis.

Watanabe (1960) was one of the first to discuss total correlation in detail, even if the

same concept had been described (but not named) previously by McGill (1954). Paluš (1994) and Wienholt and Sendhoff (1996) refer to it as redundancy. Studenỳ and Vejnarovà (1998) investigated the properties of conditional mutual information as applied to conditional independence models. They discussed total correlation, generalized it, and named it multiinformation. Multiinformation was used to show that conditional independence models have no finite axiomatic characterization. More recently Wennekers and Ay (2003) have referred to total correlation as stochastic interaction, and Sporns et al. (2000) as integration. Chechik et al. (2002) investigated the similarity between total correlation and interaction information. Vedral (2002) has compared total correlation with interaction information in the context of quantum information theory. Total correlation is sometimes even referred to as multi-variate mutual information (Boes and Meyer, 1999).

The topic of interactions was a topic of extensive investigation in statistics, but our review will be an very limited one: a large proportion of work was concerned about how to include interactions into the models, but a lesser one concerned ways of quantifying or testing interactions. Darroch (1974) surveyed two definitions of interactions, the multiplicative, which was introduced by Bartlett (1935) and generalized by Roy and Kastenbaum (1956), and the additive definition due to Lancaster (1969). Darroch (1974) preferred the multiplicative definition, and described a 'partition' of interaction which is equivalent to the entropy-based approach described in the present text. Other types of partitioning were discussed by Lancaster (1969), and more recently by Amari (2001) in the context of information geometry.

In general, variance partitioning and ANOVA have much in common with entropy decompositions, something that was noticed already by Han (1977). Variance can be interpreted as a kind of loss. In the context of multiple and logistic regression, positive interaction roughly corresponds to the notion of suppression while negative interaction has much in common with confounding and multicollinearity (Lynn, 2003). These connections are not direct, however.

The relationship between set theory and entropy

Interaction information is similar to the notion of intersection of three sets. It has been long known that these computations resemble the inclusion-exclusion properties of set theory Yeung (1991). We can view mutual information $(;)$ as a set-theoretic intersection $(\cap)$, joint entropy $(,)$ as a set-theoretic union $(\cup)$ and conditioning $(|)$ as a set difference $(-)$. The notion of entropy or information corresponds to μ, a signed measure of a set, which is a set-additive function. Yeung defines μ to be an I-measure, where μ^* of a set is equal the entropy of the corresponding probability distribution, for example $\mu^*(\tilde{X}) = H(X)$. Yeung refers to diagrammatic representations of a set of attributes as *information diagrams*, similar to Venn diagrams. Some find these diagrams misleading for more than two information sources (MacKay, 2003). One reason may be the correct interpretation of negative interaction information. Another reason is the lack of a clear concept of what the elements of the sets are. Finally, it is not always possible to keep the surface areas proportional to the actual uncertainty.

Through the principle of inclusion-exclusion and understanding that multiple mutual information is equivalent to an intersection of sets, it is possible to arrive to a slightly different formulation of interaction information (e.g. Yeung, 1991). This formulation is the most frequent in recent literature, but it has a counter-intuitive semantics, as illustrated

by Bell (2003): n-parity, a special case of which is the XOR problem for $n = 2$, is an example of a purely n-way dependence. It has a positive co-information when n is even, and a negative co-information when n is odd. For that reason we adopted the original definition of (McGill, 1954).

3.3 Visualizing Entropy Decompositions

Interactions among attributes are often very interesting for a human analyst (Freitas, 2001). We will propose a novel type of a diagram in this section to present interactions in a probabilistic model. Entropy and interaction information yield easily to graphical presentation, as they are both measured in bits. In our analysis, we have used the 'census/adult', 'mushroom', 'pima', 'zoo', 'Reuters-21578' and 'German credit' data sets from the UCI repository (Hettich and Bay, 1999). In all cases, we used maximum likelihood probability estimates.

3.3.1 Positive and Negative Interactions

A useful discovery is that attributes A and B are independent, meaning that $P(A, B)$ can be approximated with $P(A)P(B)$. If so, we say that A and B do not 2-interact, or that there is no 2-way interaction between A and B. Unfortunately, attribute C may affect the relationship between A and B in a number of ways. Controlling for the value of C, A and B may prove to be dependent even if they were previously independent. Or, A and B may actually be independent when controlling for C, but dependent otherwise.

If the introduction of the third attribute C affects the dependence between A and B, we say that A, B and C 3-interact, meaning that we cannot decipher their relationship without considering all of them at once. An appearance of a dependence is an example of a *positive interaction*: positive interactions imply that the introduction of the new attribute increased the amount of dependence. A disappearance of a dependence is a kind of a *negative interaction*: negative interactions imply that the introduction of the new attribute decreased the amount of dependence. If C does not affect the dependence between A and B, we say that there is no 3-interaction.

There are plenty of real-world examples of interactions. Negative interactions imply redundance, which may be complete or partial. For example, weather attributes clouds and lightning are dependent, because they occur together. But there is a negative interaction between thunder, clouds and lighting. Should we wonder whether there is lightning, the information that there are clouds would contribute no additional information beyond what we learned by hearing the thunder. The redundancy of clouds for predicting lightning in the context of thunder is thus complete. On the other hand, there is only a partial redundancy between wind and thunder when predicting the rain: there may be just the wind with rain, just the thunder with rain, or both thunder and wind with rain.

Positive interactions imply synergy instead. For example, employment of a person and criminal behavior are not particularly dependent attributes (most unemployed people are not criminals, and many criminals are employed), but adding the knowledge of whether the person has a new sports car suddenly makes these two attributes dependent: it is a lot more frequent that an unemployed person has a new sports car if he is involved in criminal behavior; the opposite is also true: it is somewhat unlikely that an unemployed person will have a new sports car if he is not involved in criminal behavior.

The above intuitive ideas can be mathematically dealt with through interaction information of Sect. 3.2.2. Interaction information can either be positive or negative. Perhaps the best way of illustrating the difference is through the equivalence $I(A;B;C) = I(A,B;C) - I(A;C) - I(B;C)$: Assume that we are uncertain about the value of C, but we have information about A and B. Knowledge of A alone eliminates $I(A;C)$ bits of uncertainty from C. Knowledge of B alone eliminates $I(B;C)$ bits of uncertainty from C. However, the joint knowledge of A and B eliminates $I(A,B;C)$ bits of uncertainty. Hence, if interaction information is positive, we benefit from a synergy. A well-known example of such synergy is the exclusive or: $C = A + B \pmod 2$. If interaction information is negative, we suffer diminishing returns by several attributes providing overlapping, redundant information. Another interpretation, offered by McGill (1954), is as follows: Interaction information is the amount of information gained (or lost) in transmission by controlling one attribute when the other attributes are already known.

3.3.2 Information Graphs

We can illustrate the interaction quantities we discussed with *information graphs*, introduced in (Jakulin, 2003) as interaction diagrams. They are inspired by Venn diagrams, which we render as an ordinary graph, while the surface area of each node identifies the amount of uncertainty. The representation of information with the surface area is the novelty of our approach.

White circles indicate the positive 'information' of the model, the entropy eliminated by the joint model. Gray circles indicate the two types of negative 'entropy', the initial uncertainty of the attributes and the negative interactions indicating the redundancies. Redundancies can be interpreted as overlapping of information, while information is overlapping of entropy. For example, in a redundancy $I(A;B)$ can be seen as overlapping with $I(A;C)$ in the context of A. In a synergy, $I(A;B;C)$ is overlapping of $H(A)$, $H(B)$ and $H(C)$ that is not accounted for by the 2-way interactions. The joint entropy of the attributes, or any subset of them, is obtained by summing all the gray nodes and subtracting all the white nodes linked to the relevant attributes.

Mutual Information

We start with a simple example involving two attributes from the 'census/adult' data set, illustrated in Fig. 3.3. The instances of the data set are a sample of adult population from a census database. The occupation is slightly harder to predict a priori than the education

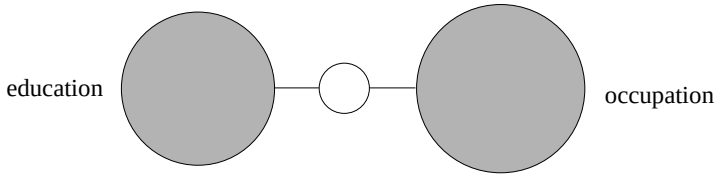


Figure 3.3: Education and occupation do have something in common: the area of the white circle indicates that the mutual information $I(\text{education};\text{occupation})$ is non-zero. This is a 2-way interaction, since two attributes are involved in it. The areas of the gray circles quantify entropy of individual attributes: $H(\text{education})$ and $H(\text{occupation})$.

because occupation entropy is larger. Because the amount of mutual information is fixed, the knowledge about the occupation will eliminate a larger proportion of uncertainty about the level of education than vice versa, but there is no reason for asserting directionality merely from the data, especially as such predictive directionality could be mistaken for causality.

A Negative Interaction

The relationship between three characteristics of animals in the 'zoo' database is rendered in Fig. 3.4. All three attributes are 2-interacting, but there is an overlap in the mutual information among each pair, indicated by a negative interaction information. It is illustrated as the gray circle, connected to the 2-way interactions, which means that they have a shared quantity of information. It would be wrong to subtract all the 2-way interactions from the sum of individual entropies to estimate the complexity of the triplet, as we would underestimate it. For that reason, the 3-way negative interaction acts as a correcting factor.

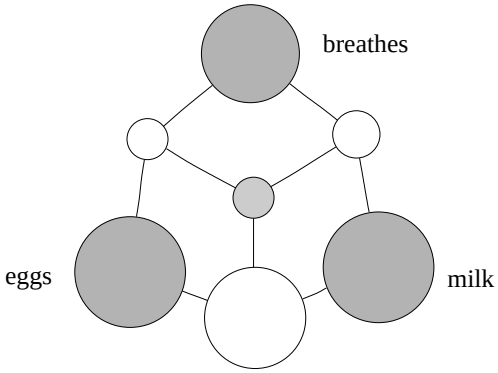


Figure 3.4: An example of a 3-way negative interaction between the properties 'lays eggs?', 'breathes?' and 'has milk?' for different animals. The negative interaction is indicated with the dark circle connected to positive 2-way interactions, as it can be understood as overlap between them.

This model is also applicable to supervised learning. If we were interested if an animal breathes, but knowing whether it gives milk and whether it lays eggs, we would obtain the residual uncertainty $H(\text{breathes}|\text{eggs}, \text{milk})$ by the following formula:

$$H(\text{breathes}) - (I(\text{breathes}; \text{eggs}) + I(\text{breathes}; \text{milk}) + I(\text{breathes}; \text{eggs}; \text{milk}))\,.$$

This domain is better predictable than the one from Fig. 3.3, since the 2-way interactions are comparable in size to the prior attribute entropies. It is quite easy to see that knowing whether whether an animal lays eggs provides us pretty much all the evidence whether it has milk: mammals do not lay eggs. Of course, such deterministic rules are not common in natural domains.

Furthermore, the 2-way interactions between breathing and eggs and between breathing and milk are very similar in magnitude to the 3-way interaction, but opposite in sign, meaning that they cancel each other out. Using the relationship between conditional

mutual information and interaction information from (3.8), we can conclude that:

$$I(\text{breathes}; \text{eggs}|\text{milk}) \approx 0$$
$$I(\text{breathes}; \text{milk}|\text{eggs}) \approx 0$$

Therefore, if the 2-way interaction between such a pair is ignored, we need no 3-way correcting factor. The relationship between these attributes can be described with two Bayesian network models, each assuming that a certain 2-way interaction does not exist in the context of the remaining attribute:

$$\text{breathes} \leftarrow \text{milk} \rightarrow \text{eggs}$$
$$\text{breathes} \leftarrow \text{eggs} \rightarrow \text{milk}$$

If we were using the naïve Bayesian classifier for predicting whether an animal breathes, we might also find out that feature selection could eliminate one of the attributes: Trying to decide whether an animal breathes, and knowing that the animal lays eggs, most of the information contributed by the fact that the animal doesn't have milk is redundant. Of course, during classification we might have to classify an animal only with the knowledge of whether it has milk, because the egg-laying attribute value is missing: this problem is rarely a concern in feature selection and feature weighting.

A Positive Interaction

Most real-life domains are difficult, meaning that it is hopeless trying to predict the outcome deterministically. One such problem domain is a potential customer's credit risk estimation. Still, we can do a good job predicting the changes in risk for different attribute values. The 'German credit' domain describes credit risk for a number of customers. Fig. 3.5 describes a relationship between the risk with a customer and two of his characteristics. The mutual information between any attribute pairs is low, indicating high uncertainty and weak predictability. The interesting aspect is the positive 3-interaction, which additionally reduces the entropy of the model. We emphasize the positivity by painting the circle corresponding to the 3-way interaction white, as this indicates information.

It is not hard to understand the significance of this synergy. On average, unemployed applicants are riskier as customers than employed ones. Also, applying for a credit to finance a business is riskier than applying for a TV set purchase. But if we heard that an unemployed person is applying for a credit to finance purchasing a new car, it would provide much more information about risk than if an employed person had given the same purpose. The corresponding reduction in credit risk uncertainty is the sum of all three interactions connected to it, on the basis of employment, on the basis of purpose, and on the basis of employment and purpose simultaneously.

It is extremely important to note that the positive interaction coexists with a mutual information between both attributes. If we removed one of the attributes because it is correlated with the other one in a feature selection procedure, we would also give up the positive interaction. In fact, positively interacting attributes are often correlated.

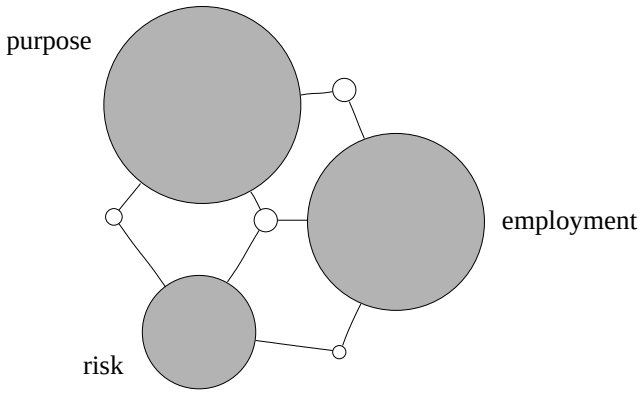


Figure 3.5: An example of a 3-way positive interaction between the customer's credit risk, his purpose for applying for a credit and his employment status.

Zero Interaction

The first explanation for a situation with zero 3-way interaction information is that an attribute C does not affect the relationship between attributes A and B, thus explaining the zero interaction information $I(A;B|C) = I(A;B) \Rightarrow I(A;B;C) = 0$. A homogeneous association among three attributes is described by all the attributes 2-interacting, but not 3-interacting. This would mean that their relationship is fully described by a loopy set of 2-way marginal associations. Although one could imagine that Fig. 3.6 describes such a homogeneous association, there is another possibility.

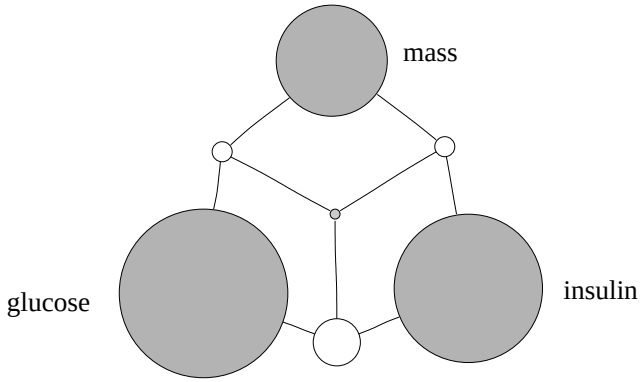


Figure 3.6: An example of an approximately homogeneous association between body mass, and insulin and glucose levels in the 'pima' data set. All the attributes are involved in 2-way interactions, yet the negative 3-way interaction is very weak, indicating that all the 2-way interactions are independent. An alternative explanation would be a mixture of a positive and a negative interaction.

Imagine a situation which is a mixture of a positive and a negative interaction. Three attributes A, B, C take values from $\{0, 1, 2\}$. The permissible events are $\{e_1 : A = B + C \pmod 2,\ e_2 : A = B = C = 2\}$. The event e_1 is the familiar XOR combination, denoting a positive interaction. The event e_2 is an example of perfectly correlated attributes, an

example of a negative interaction. In an appropriate probabilistic mixture, for example $\Pr\{e_1\} \approx 0.773$, $\Pr\{e_2\} \approx 0.227$, the interaction information $I(A;B;C)$ approaches zero. Namely, $I(A;B;C)$ is the average interaction information across all the possible combinations of values of A, B and C. The distinctly positive interaction for the event e_1 is cancelled out, on average, with the distinctly negative interaction for the event e_2. The benefit of joining the three attributes and solving the XOR problem exactly matches the loss caused by duplicating the dependence between the three attributes.

Hence, 3-way interaction information should not be seen as a full description of the 3-way interaction but as the interaction information averaged over the attribute values, even if we consider interaction information of lower and higher orders. These problems are not specific only to situations with zero interaction information, but in general. If a single attribute contains information about complex events, much information is blended together, which should rather be kept apart. Not to be misled by such mixtures, we may represent a many-valued attribute A with a set of binary attributes, each corresponding to one of the values of A. Alternatively, we may examine the value of interaction information at particular attribute values. The visualization procedure may assist in determining the interactions to be examined closely by including bounds or confidence intervals for interaction information across all combinations of attribute values; when the bounds are not tight, a mixture can be suspected.

Patterns of Interactions

If the number of attributes under investigation is increased, the combinatorial complexity of interaction information may quickly get out of control. Fortunately, interaction information is often low for most combinations of unrelated attributes. We have also observed that the average interaction information of a certain order is decreasing with the order in a set of attributes. A simple approach is to identify N interactions with maximum interaction magnitude among the n. For performance and reliability, we also limit the maximum interaction order to k, meaning that we only investigate l-way interactions, $2 \leq l \leq k \leq n$. Namely, it is difficult to reliably estimate joint probability distributions of high order. The estimate of $P(X)$ is usually more robust than the estimate of $P(X, Y, Z, W)$ given the same number of instances.

Mediation and Moderation A larger scale information graph with a selection of interactions in the 'mushroom' domain is illustrated in Fig. 3.7. Because edibility is the attribute of interest (the label), we center our attention on it, and display a few other attributes associated with it. The informativeness of the stalk shape attribute towards mushroom's edibility is very weak, but this attribute has a massive synergistic effect if accompanied with the stalk root shape attribute. We can describe the situation with the term *moderation* (Baron and Kenny, 1986): stalk shape 'moderates' the effect of stalk root shape on edibility. Stalk shape is hence a moderator variable. It is easy to see that such a situation is problematic for feature selection: if our objective was to predict edibility, a myopic feature selection algorithm would eliminate the stalk shape attribute, before we could take advantage of it in company of stalk root shape attribute. Because the magnitude of the mutual information between edibility and stalk root shape is similar in magnitude to the negative interaction among all three, we can conclude that there is a conditional independence between edibility of a mushroom and its stalk root shape given

the mushroom's odor. A useful term for such a situation is *mediation* (Baron and Kenny, 1986): odor 'mediates' the effect of stalk root shape on edibility.

The 4-way interaction information involving all four attributes was omitted from the graph, but it is distinctly negative. This can be understood by looking at the information gained about edibility from the other attributes and their interactions with the actual entropy of edibility: we cannot explain 120% of entropy, unless we are counting the evidence twice. The negativity of the 4-way interaction indicates that a certain amount of information provided by the stalk shape, stalk root shape and their interaction is also provided by the odor attribute.

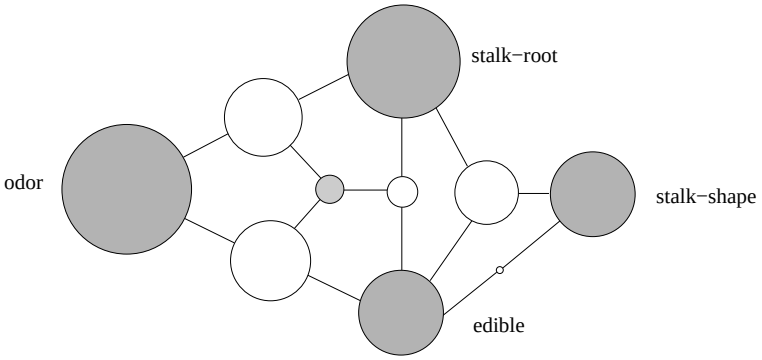


Figure 3.7: A selection of several important interactions in the 'mushroom' domain.

Conditional interaction graphs The visualization method applied to the 'mushroom' domain is unsupervised in the sense that it describes 3-way interactions outside any context, for example $I(A;B;C)$. However, it was customized for supervised learning by only examining the interactions that involve the labelled attribute, edibility. We now focus on higher-order interactions in the context of the label, such as $I(A;B|Y)$ and $I(A;B;C|Y)$ where Y is the label. These are useful for verifying the grounds for taking the conditional independence assumption in the naïve Bayesian classifier. Such assumption may be problematic if there are informative conditional interactions between attributes with respect to the label.

In Fig. 3.8 we have illustrated the informative conditional interactions with large magnitude in the 'adult/census' data set, with respect to the label – the salary attribute. Learning with the conditional independence assumption would imply that these interactions are ignored. The negative 3-way conditional interaction with large magnitude involving education, years of education and occupation (in the context of the label) offers a possibility for simplifying the domain. Other attributes from the domain were left out from the chart, as only the race and the native country attribute were conditionally interacting.

Synonymy and Polysemy In complex data sets, such as the ones for information retrieval, the number of attributes may be measured in tens of thousands. Interaction analysis must hence stem from a particular reference point. For example, let us focus on the keyword 'franc', the currency, in the 'Reuters' data set. This keyword is not a label, but merely a determiner of context. We investigate the words that co-appear

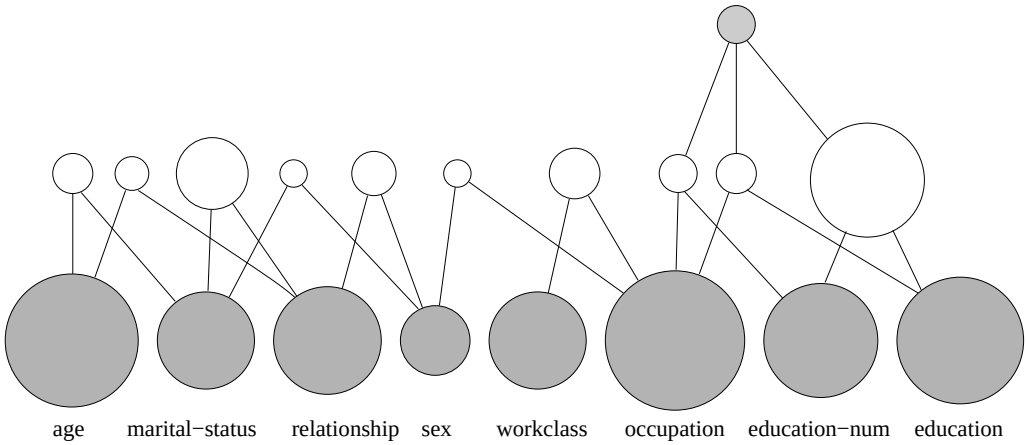


Figure 3.8: An informative conditional interaction graph illustrating the informative conditional interactions between the attributes with respect to the salary label in the 'adult/census' domain.

with it in news reports, and identify a few that are involved in 2-way interactions with it. Among these, we may identify those of the 3-way interactions with high normed interaction magnitude. The result of this analysis is rendered in Fig. 3.9. We can observe the positive interaction among 'Swiss', 'French' and 'franc' which indicates that 'franc' is polysemous. There are two contexts in which the word 'franc' appears, but these two contexts do not mix, and this causes the interaction to be positive. The strong 2-way interaction between 'franc' and 'francs' indicates a likelihood of synonymy: the two words are frequently both present or both absent, and the same is true of pairs 'French'-'Paris' and 'Swiss'-'Zurich'. Looking at the mutual information (which is not illustrated), the two negative interactions are in fact near conditional independencies, where 'Zurich' and 'franc' are conditionally independent given 'Swiss', while 'French' and 'franc' are conditionally independent given 'Paris'. Hence, the two keywords that are best suited to distinguish the contexts of the kinds of 'franc' are 'Swiss' and 'Paris'. These three are positively interacting, too.

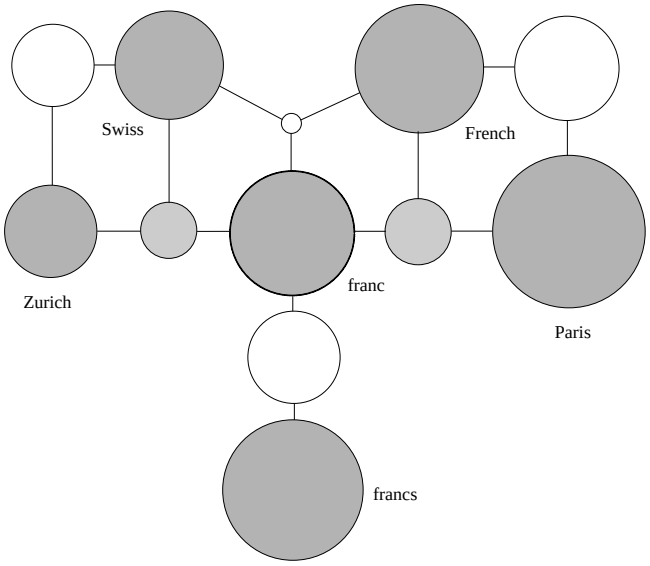


Figure 3.9: A selection of interactions involving the keyword 'franc' in news reports shows that interaction analysis can help identify useful contexts, synonyms and polysemy in information retrieval. Because not all 2-way interactions are listed, the negative 3-way interaction is attached to the attributes and not to the underlying 2-way interactions.

CHAPTER 4

The Statistics of Interactions

In Chapter 3 the probability model was given and interactions were examined in the context of it. Although we did examine the interactions on data, we did not focus on how the probability model was estimated. The purpose of this chapter is to tackle the issues of Sects. 2.2.4 and 2.2.5: we may believe that several models are valid in explaining the data, or that there are several data sets consistent with a given model.

The probability model is derived from the data using either Bayesian or frequentist methods. There may be several probability models, and we may be uncertain about them. For each probability model, it is possible to induce the probability or belief distributions over the information-theoretic quantities. Mutual information is never less than zero, yet sometimes it is preferable to assume independence. We approach this problem by contrasting it with the notion of a significance test. We integrate significance tests with the notion of loss functions. We compare cross-validation, goodness-of-fit testing, permutation and bootstrap tests, and a novel formulation of fully Bayesian significance testing.

In Sect. 4.4 we re-examine the relationship between interaction information and particular models that underlie it. We define the underlying models as part-to-whole approximations, and list three ways of coming up with them: by approximation, by maximum entropy and by maximum likelihood.

4.1 Two Kinds of 'Interaction'

Before we start, we need a clarification. There are two different concepts, entropy decompositions (Chapter 3) and model comparisons. We will now summarize both, pointing out the similarities and stressing the differences.

4.1.1 Entropy Decompositions

The starting point in information theory is often a single unambiguous joint probability model $P(X_1, X_2, \ldots, X_n)$. Interaction information, mutual information and total correlation take the joint model and decompose the joint entropy $H(X_1, X_2, \ldots, X_n)$ into a sum of terms through an analogy with set theory (Yeung, 1991).

We can approximate the joint entropy by adding and subtracting the terms, facilitated by the useful properties of Shannon entropy. Examples of such approximations to joint entropy are the Bethe and Kikuchi approximations (Yedidia et al., 2004). We can evaluate the quality of certain types of approximations using the notions of mutual information (Shannon, 1948), and total correlation (Watanabe, 1960). Furthermore, we can create lattice structures in the entropy space that can be used to reconstruct the entropy of subsets of attributes (Han, 1975). The terms in these lattice-based decompositions, such as interaction information (McGill, 1954), are informative and provide insight into the joint probability model. Slightly different approaches to decomposing entropy were proposed by Han (1978) and Amari (2001).

Assume that $\hat{P}_{BN}$ is a *Bayesian network* (Pearl, 1988) based on P. It can then be represented as a directed acyclic graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The approximate global joint probability model $\hat{P}_{BN}(\mathcal{X})$ based on $\mathcal{G}$ is defined as:

$$\hat{P}_{BN}(\mathcal{V}) = \prod_{X \in \mathcal{V}} P(X|\mathbf{Parents}_X) \tag{4.1}$$

Thus, $\hat{P}$ is based on conditioning and marginalization of P, and the chain rule is used to merge these portions together. If the Kullback-Leibler divergence $D(P\|\hat{P})$ is used as the loss function, it can be expressed as:

$$D(P\|\hat{P}_{BN}) = \sum_{X \in \mathcal{V}} H(X|\mathbf{Parents}_X) - H(\mathcal{V}) \tag{4.2}$$

Here, the entropy H is based on the underlying P. The chain rule guarantees that the joint probability model $\hat{P}_{BN}$ requires no normalization, and that the parameters of $\hat{P}_{BN}$ can be computed in closed form.

It is characteristic of entropy decompositions that the assumption of no interaction cannot provide a positive utility: the global joint model $P(\cdot)$ is assumed to be 'true'.

4.1.2 Model Comparisons

In statistics, model comparisons start with *two* inferred or estimated probability models, the *reference* P, and the *approximation* Q. The quantification of an interaction corresponds to the loss incurred or the gain obtained by the approximation that does not assume that interaction. The definition of interaction thus corresponds to the restriction imposed on the approximation. The quantification of the interaction depends on the loss function used.

P and Q need not be joint models, and can also be conditional models of the type $P(\mathcal{X}|\mathcal{Y})$ and $Q(\mathcal{A}|\mathcal{B})$. The sufficient usual requirement for meaningful comparison here is that $\mathcal{X} = \mathcal{A}$ and $\mathcal{Y} \subseteq \mathcal{B}$. More generally, however, there must be a function f that relates $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{A} \times \mathcal{B}$:

$$f : \mathcal{X} \times \mathcal{Y} \to \mathcal{A} \times \mathcal{B} \tag{4.3}$$

To obtain the expected loss, the joint reference $P(\mathcal{X}, \mathcal{Y})$ must be defined. It quantifies the weight of loss that corresponds to an individual case $L(P(x|y), Q(a|b))$. With the joint reference, the expected loss is then:

$$L(P, Q) = \mathbb{E}_{x,y \sim P(\mathcal{X},\mathcal{Y}),\, (a,b)=f(x,y)}\{L(P(x|y), Q(a|b))\} \tag{4.4}$$

Given a finite data set, the statistical models of either P or Q are generally not unique and cannot be determined with precision. For that reason, the result of a model comparison, the loss $c = L(P, Q)$, is itself a random quantity, and can itself be modelled with a probability model $P(c)$. $P(c)$ can be summarized with confidence intervals, point-estimated using the expected loss.

It is characteristic for a model comparison that the removal of an interaction may reduce the loss. Because the joint model $P(\cdot)$ is inferred from the data, a simpler model may involve less risk. However, the amount of reduction or gain is generally dependent on other interactions in a global model, and depends on whether the model is conditional or joint.

Although entropy decompositions are not directly concerned with model comparisons, they can still be used as heuristics. However, when the entropy decomposition does not correspond to a Bayesian network, we must be careful: the implicit alternative model may not be normalized, as happens in the case of negative interaction information.

4.2 Probability Estimation

Probabilities are not given a priori: all we have is the data. The fundamental model normally used to connect the probabilities with the data is the *multinomial distribution* $\text{Multinomial}_k(n, p_1, p_2, \ldots, p_k)$, where $p_i \geq 0$ and $\sum_{i=1}^{k} p_i = 1$. There are k types of events, and each event has a specified probability. If $k = 2$, we speak of the *binomial distribution*, and when $k = 2$ and $n = 1$ we have the *Bernoulli distribution*.

The attribute X with the range $\Re_X = \{1, 2, \ldots, k\}$ is given, along with the data $\mathcal{D}$. We can examine the distribution of the attribute using the number of occurrences of each of the k values, $n_i = \#_{\mathcal{D}}\{X = i\}$, $\sum_i n_i = n = |\mathcal{D}|$. The resulting count vector $\mathbf{n} = [n_1, \ldots, n_k]^T$ is a *sufficient statistic*: all that we seek to recover from the data is the count vector, and other properties of the data are ignored, including the ordering of the instances, or other attributes. Each instance is an event, and each attribute value is the type of the event, counted separately.

The vector of counts $\mathbf{n}$ can be assumed to have the multinomial distribution, which we express as

$$\mathbf{n} \sim \text{Multinomial}_k(n, p_1, p_2, \ldots, p_k).$$

We can summarize the vector of probabilities as $\mathbf{p} = [p_1, p_2, \ldots, p_k]^T$. The probability of an occurrence of the data with such a sufficient statistic from a particular model $\mathbf{p}$ under the assumption of $\mathbf{n}$ having a multinomial distribution can be expressed as a probability *mass* function:

$$P(\mathbf{n}|\mathbf{p}) \triangleq \binom{n}{n_1, n_2, \ldots, n_k} \prod_{i=1}^{k} p_i^{n_i} \tag{4.5}$$

For each $\mathbf{p}$ we can imagine the multinomial model in the generative sense as a *dice-tossing machine*, which explains the working of the model. We construct a dice with k sides, biased in such a way that the i-th side falls with the probability of p_i. We now toss the dice n times, and obtain a data set of size n. Each toss is independent of all others. We now perform this experiment infinitely many times, where each experiment is also independent of all others. Then, $P(\mathbf{n}|\mathbf{p})$ is the proportion of the experiments that

resulted in the same value $\mathbf{n}$ of the sufficient statistic, based on an infinite sequence of experiments made with the dice-tossing machine corresponding to $\mathbf{p}$.

In an empirical situation, the the probability vector $\mathbf{p}$ is unknown, and we have just the data. The data is summarized through the sufficient statistic, obtaining the counts $\mathbf{n}$, where $n = \sum_{i=1}^{k} n_i$. *Relative frequencies* are often used as approximate estimates of probabilities, and are defined as $\hat{\pi}_i \triangleq \frac{n_i}{n}$. It can be shown that relative frequencies are the maximum likelihood estimates. This means that among all multinomial dice-tossing machines, the ones parameterized by the relative frequencies will yield the data sets having such counts most frequently.

4.2.1 Frequentist Vagueness

We now assume that the relative frequencies $\hat{\pi}$ are indeed the true probabilities. We can generate new random data sets of equal size with the dice-tossing machine. If we estimate the relative frequencies in the data sets that come out from the machine, they will generally differ from the original ones $\hat{\pi}$, unless the sample is infinite. Therefore, the true probability based on a finite sample is a vague concept, something we have addressed in Sect. 2.2.5.

Parametric Bootstrap. The approach of generating random data sets through an assumed model is referred to as the *parametric bootstrap* (Efron and Gong, 1983). The above dice-tossing machine is nothing but a a metaphor for parametric bootstrap. Using the relative frequencies, we now generate B *resamples* or resampled data sets $\{\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}\}$ of the same size as the original one $|\mathcal{D}^{*\cdot}| = |\mathcal{D}| = n$. Each instance in each data set is independent of all others, but that instances are distributed in accordance with the estimated model Multinomial$(n, \hat{\boldsymbol{\pi}})$. Ideally, $B \to \infty$, but choices of $B = 100$ or $B = 1000$ are used in practice. From each resample $\mathcal{D}^{*i}$, we can re-estimate the relative frequencies, and end up with a vector of *bootstrap replications*, $\{\hat{\boldsymbol{\pi}}^{*1}, \hat{\boldsymbol{\pi}}^{*2}, \ldots, \hat{\boldsymbol{\pi}}^{*B}\}$.

Nonparametric Bootstrap. The *nonparametric bootstrap* does not operate on sufficient statistics. Instead, each instance $\mathbf{d}_i \in \mathcal{D}$ is considered to be a value of a n-valued attribute D, where $\Re_D = \mathcal{D}$, and $n = |\mathcal{D}|$. Each value d has the equal probability of $\frac{1}{n}$, so the probability vector is $\mathbf{p}_D = [\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}]^T$. Nonparametric bootstrap thus draws samples of size n from the multinomial distribution parameterized by $\mathbf{p}_D$. For each value of D with a non-zero count, the corresponding number of instances are drawn from $\mathcal{D}$. Thus, $D \sim \text{Multinomial}_n(n, \mathbf{p}_D)$ results in the resample $\mathcal{D}^* = \{\mathbf{d}^{(i)};\ i \in \{1, 2, \ldots, n\}\}$. In the case of the multinomial distribution, the parametric and nonparametric bootstrap are equivalent, but nonparametric bootstrap is useful when no parametric distribution is sufficiently persuasive.

Bayesian Bootstrap. Another interpretation of the nonparametric bootstrap is that it assigns weights to instances, so that the weights have a multinomial distribution. Weights are always integers, and Rubin (1981) suggested smoother real-valued weights instead. This is especially useful when we have small data sets. We achieve this by introducing a weight vector that has the Dirichlet distribution $\mathbf{w} \sim \text{Dirichlet}_n(1, 1, \ldots, 1)$ (Lee and Clyde, 2004). The i-th instance has weight $\mathbf{w}^{(i)}$.

Bootstrap and Inference. The frequentist approach may be misleading for inference. For example, we take a coin that falls heads with $p_H = 0.9$. We toss it three times and obtain $\{H, H, H\}$. The relative frequency would be $\hat{\pi}_H = 1$, and all generated data sets of size 3 will be also $\{H, H, H\}$. All estimates from the generated data sets will also have $\hat{\pi}^*_H = 1$, and not one of them will match the truth. If we obtained $\{H, T, H\}$, and generated data sets of the same size, the probability estimates would only take the following values $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Again, $p_H = 0.9$ does not appear in the sample. In neither case would the employment of the nonparametric bootstrap help arrive at a non-zero estimate of the true probability, although the Bayesian bootstrap would help in the latter case. Nevertheless, bootstrap 'works well in practice.'

Asymptotic Vagueness. We can approximate the multinomial distribution with a multivariate Gaussian one (Agresti, 2002), whose mean $\boldsymbol{\mu} = \mathbf{p}$, and the covariance matrix $\boldsymbol{\Sigma} = \text{diag}(\mathbf{p}) - \mathbf{p}\mathbf{p}^T$. It is easy to see that the vague probability can be modelled with a multivariate Gaussian centered at the relative frequency:

$$\mathbf{p} \sim \text{Gaussian}(\hat{\boldsymbol{\pi}}, \boldsymbol{\Sigma}/n) \tag{4.6}$$

This realization also entails the use of χ^2 distribution in testing.

4.2.2 Bayesian Vagueness

The original disposition of the Bayesian approach is that the relative frequencies are not an appropriate way of obtaining the model, not even as the center of vagueness. Instead, a prior is postulated, which models how much belief we have in different probabilities before seeing the data. Often, a conjugate prior is chosen for a number of convenient properties. For the multinomial parameters, the commonly used conjugate prior is the *Dirichlet distribution*, and obtain the probability *density* function of the probability vector given the data:

$$\mathbf{p} \sim \text{Dirichlet}_k(n_1, \ldots, n_k) \tag{4.7}$$

$$p(\mathbf{p}|\mathbf{n}) \triangleq \frac{\Gamma\left(\sum_{i=1}^k n_i\right)}{\prod_{i=1}^k \Gamma(n_i)} \prod_{i=1}^k p_i^{n_i - 1} \qquad \text{for } p_i \geq 0 \text{ and } \sum_{i=0}^k p_i = 1 \tag{4.8}$$

The use of a Dirichlet prior parameterized by $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_k]^T$ results in a Dirichlet posterior $[\alpha_1 + n_1, \ldots, \alpha_k + n_k]^T$. When all $\alpha_i \to 0$, we speak of an improper prior (which does, however, result in a proper posterior if all $n_i > 0$). When all $\alpha_i = 1$, we have the usual noninformative uniform prior, which is the basis for the Laplace probability estimate. A special case of the Dirichlet distribution for $k = 2$ is the *beta distribution*.

We can sample from a Dirichlet distribution using the following approach (Gelman et al., 2004a): draw $x_1, \ldots, x_k$ from independent gamma distributions with shape parameters $\alpha_1, \ldots, \alpha_k$ and common scale, then form each of the probabilities by normalizing $p_i = x_i / \sum_i x_i$.

4.2.3 Vagueness of Loss

Entropy and mutual information can be interpreted as loss, as we have seen in Sect. 3.1.1. Specifically, the model P suffers the loss of $L(\mathbf{x}, P) = -\log_2 P(\mathbf{x})$ for an instance $\mathbf{x} \in \mathcal{D}$.

The empirical entropy of $\mathbf{X}$ given P and the Kullback-Leibler divergence between P and Q can then be defined as:

$$H(\mathbf{X}|P) = \mathbb{E}_{\mathbf{x}\sim P}\{L(\mathbf{x}, P)\} \tag{4.9}$$

$$D(P(\mathbf{X})\|\hat{P}(\mathbf{X})) = \mathbb{E}_{\mathbf{x}\sim P}\{L(\mathbf{x}, \hat{P}) - L(\mathbf{x}, P)\} \tag{4.10}$$

It is possible to keep P and Q fixed, and only vary the distribution over which the expectation is computed. This way, a large variance in loss over the whole range will be reflected in an increased variance in entropy or divergence. Specifically, we can compute the expectation of loss on various resamples of the data set, but maintaining the two models. For example:

$$\hat{I}(A;B)^{*,i} = \frac{1}{|\mathcal{D}^{*,i}|}\sum_{\mathbf{x}\in\mathcal{D}^{*,i}}\Big(L(\mathbf{x}, \hat{P}) - L(\mathbf{x}, P)\Big) \tag{4.11}$$

4.3 Case Study: Distribution of Mutual Information

All the methods of the previous section result in a probability distribution on probability vectors. Individual probability vectors are parameters to the multinomial distribution. The bootstrap methods yield a sequence of the probability vector estimates, the asymptotic approach results in a multivariate normal approximation to the estimate of the probability vector, and the Bayesian approach result in a posterior distribution over the probability vectors. In every case, we have a number or a continuum of probability models consistent with the data. We can interpret them as inverse probabilities, second-order probability distributions, or conditional probability distributions. We will refer to them as *belief distributions* or *ensemble models*, because we have non-zero belief in a number of probability models, and because our uncertain knowledge can be represented with an ensemble of probabilities.

While it may be interesting to examine the ensembles by themselves, as we have done in Fig. 2.7, the interpretation soon becomes overwhelming without making additional assumptions. Entropy and mutual information are always expressed relative to a particular probability model $\mathbf{p}$ that assigns probabilities to an arbitrary combination of the values of the attributes. We make this dependence explicit by writing $I(A;B|\mathbf{p})$ and $H(A|\mathbf{p})$.

Miller (1955) noticed that entropy and mutual information based on the relative frequency estimates are biased: based on count data we often underestimate entropy, and overestimate mutual information. He proposed the following unbiased estimates of entropy and mutual information:

$$\mathbb{E}'\{H(X)\} = H(X) + \log_2 e\left(\frac{|\Re_X| - 1}{2n} - \frac{1 - \sum_{x\in\Re_X}\hat{p}(x)^{-1}}{12n^2}\right) + O(n^{-3}) \tag{4.12}$$

$$\mathbb{E}'\{I(X;Y)\} = I(X;Y) - \log_2 e\frac{(|\Re_X| - 1)(|\Re_Y| - 1)}{2n} + O(n^{-2}) \tag{4.13}$$

Wolpert and Wolf (1995) theoretically analyzed the problem of Bayesian estimation of mutual information, and have suggested using the uniform Dirichlet prior. When $\mathbf{p}$ is the probability model based on the data $\mathcal{D}$, we can speak of the *posterior probability model of*

entropy:

$$\Pr\{H(\mathbf{X}|\mathcal{D}) \leq w\} = \int \mathbb{I}\{H(\mathbf{X}|\mathcal{D},\mathbf{p})) \leq w\} p(\mathbf{p}|\mathcal{D}) d\mathbf{p} \tag{4.14}$$

Here, $\mathbb{I}\{C\}$ is the indicator function, taking the value of 1 when the condition C is fulfilled and 0 otherwise. We can summarize the posterior probability model of entropy with the *expected posterior entropy*:

$$\mathbb{E}_{\mathbf{p}\sim p(\mathbf{p}|\mathcal{D})}\{H(\mathbf{X}|\mathbf{p})\} \tag{4.15}$$

In most cases, however, *entropy given the posterior predictive model of* $\mathbf{p}$ is used,

$$H(\mathbf{X}|\mathbb{E}_{\mathbf{p}\sim p(\mathbf{p}|\mathcal{D})}\{\mathbf{p}\}) = H\left(\mathbf{X}\middle|\mathcal{D}, \int P(\mathbf{X},\mathbf{p}|\mathcal{D}) d\mathbf{p}\right) \tag{4.16}$$

We will generally use the posterior distribution of entropy (4.14), but will sometimes summarize it with the expected posterior entropy (4.15). We will not employ the posterior predictive models of $\mathbf{p}$ for evaluating entropy, because it makes the misleading impression of entropy being a scalar when it is really a random quantity.

Hutter and Zaffalon (2005) combined the Bayesian approach with asymptotic approximations, and provide the approximate moments of the mutual information in accessible closed form. The equations end up being very similar to the ones of Miller (1955), and are as follows:

$$\mathbb{E}''\{I(X;Y)\} = I(X;Y) - \log_2 e \frac{(|\Re_X|-1)(|\Re_Y|-1)}{2n+1} \tag{4.17}$$

$$\begin{aligned}\mathbb{V}\mathrm{ar}''\{I(X;Y)\} &= \mathbb{E}\left\{(I(X;Y) - \mathbb{E}\{I(X;Y)\})^2\right\} \\ &= \frac{1}{n+1}\left(\sum_{x\in\Re_X, y\in\Re_Y} \left(P(x,y)I(x;y)^2\right) - I(X;Y)^2\right)\end{aligned} \tag{4.18}$$

As an example, we examined a realistic pair of attributes from a concrete data set: 'Harris hip score' and 'diabetes' from the HHS data set discussed in Sect. 7.3.3. The *contingency table* lists the number of patients for every combination of the attribute values:

	HHS = excellent	HHS = good	HHS = bad
diabetes = no	37	30	28
diabetes = yes	6	4	7

We have employed both asymptotic corrections along with other approaches. We performed the nonparametric bootstrap and the Bayesian bootstrap with $B = 20000$. We also took 20000 samples from $\mathbf{p}$'s posterior based on the improper Dirichlet prior $\boldsymbol{\alpha} \to \mathbf{0}$ (Bernardo and Smith, 2000) (also referred to as the Haldane prior) and the uniform noninformative prior $\forall i : \alpha_i = 1$ (Wolpert and Wolf, 1995) (also referred to as the Laplacean prior). For each $\mathbf{p}$ we computed the mutual information. The results obtained with these methods differ:

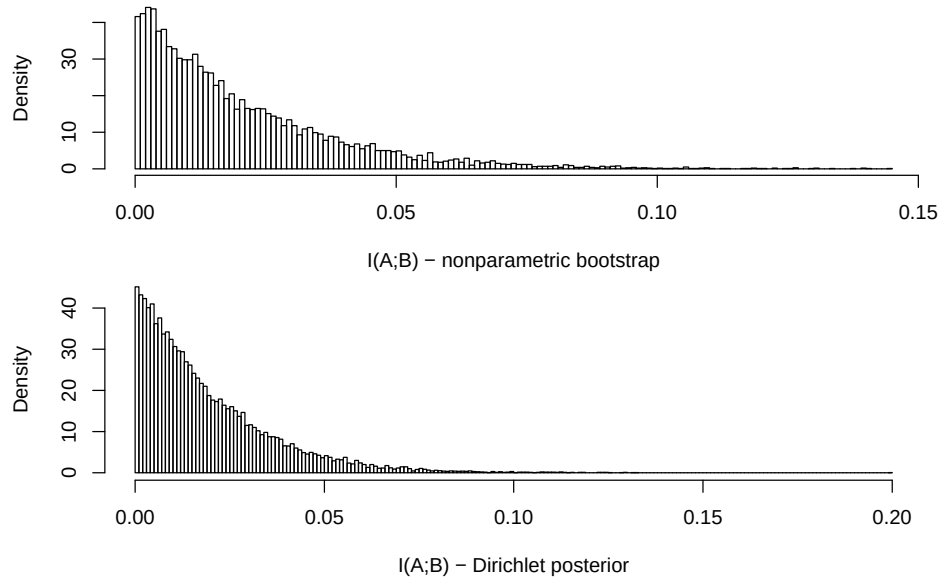


Figure 4.1: We can obtain the belief distribution of mutual information estimates by resampling from the count data using the nonparametric bootstrap (top), and by sampling from the Bayesian posterior using the improper Dirichlet prior $\boldsymbol{\alpha} \to \mathbf{0}$ (bottom).

$I(A;B\|\hat{\boldsymbol{\pi}})$	relative frequency estimate	0.006
$\mathbb{E}\{I(X;Y\|\hat{\boldsymbol{\pi}}^*)\}$	bootstrap	0.020
$\mathbb{E}\{I(X;Y\|\hat{\boldsymbol{\pi}}^*)\}$	Bayesian bootstrap	0.019
$\mathbb{E}_{\mathbf{p}\sim\text{Dirichlet}(n,n_1,\ldots,n_k)}\{I(X;Y\|\mathbf{p})\}$	Dirichlet posterior ($\boldsymbol{\alpha} \to \mathbf{0}$)	0.019
$\mathbb{E}_{\mathbf{p}\sim\text{Dirichlet}(n+k,n_1+1,\ldots,n_k+1)}\{I(X;Y\|\mathbf{p})\}$	(Wolpert and Wolf, 1995)	0.017
$\mathbb{E}'\{I(X;Y)\}$	(Miller, 1955)	-0.007
$\mathbb{E}''\{I(X;Y)\}$	(Hutter and Zaffalon, 2005)	-0.007
$\mathbb{V}\text{ar}''\{I(X;Y)\}^{1/2}$	(Hutter and Zaffalon, 2005)	0.013

Interestingly, the asymptotic corrections result in negative values. More illuminating than the mean values may be the histogram of bootstrapped mutual information estimates shown in Fig. 4.1.

There is a clear similarity between the two asymptotic corrections, and between the Bayesian bootstrap and the improper Dirichlet prior. However, other results seem to differ, and there is little empirical ground for making a choice: one has to pick the method one trusts the most. Nevertheless, all methods seem to agree that the 'diabetes' attribute does not have a consistently high information gain.

The mean of the mutual information is not a good representation for the distributions like those in Fig. 4.1, not even if accompanied by variance. Because the distribution is asymmetric, saying that the mean is 0.019 ± 0.017 (bootstrap) or -0.007 ± 0.013 (asymptotic) might be misleading. A better summary is a *confidence interval*, expressed in terms of the distribution percentiles. The 99% confidence interval reaches between the mutual information at the 0.5%-th percentile and the 99.5%-th percentile of the posterior belief. We summarize some of the distributions listed earlier:

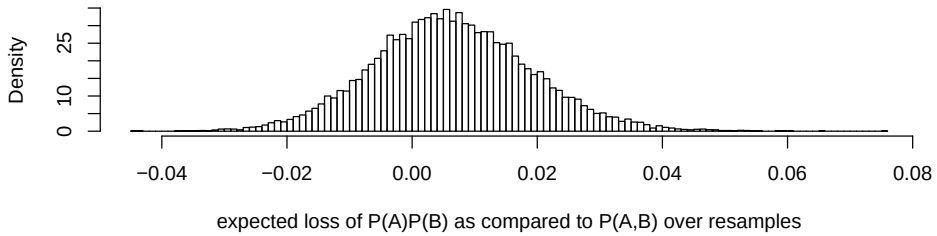


Figure 4.2: The distribution of mutual information obtained by through vague loss is quite different from other approaches.

bootstrap	[0.0001, 0.0969]
Bayesian bootstrap	[0.0001, 0.0876]
Dirichlet posterior: $\boldsymbol{\alpha} \to \mathbf{0}$	[0.0001, 0.0887]
Dirichlet posterior: $\forall i : \alpha_i = 1$ (Wolpert and Wolf, 1995)	[0.0001, 0.0807]

All these confidence intervals agree with the 0.005 belief that the mutual information will be less than 0.0001.

The approach with vagueness of loss $\hat{I}(A;B)$ (4.11), estimated using the Bayesian bootstrap, resulted in a distinctly different distribution, as shown in Fig. 4.2. The 99% confidence interval of $\hat{I}(A;B)$ was [-0.0255,0.0418].

4.4 Part-to-Whole Approximations

We are often not interested in the estimate of mutual and interaction information itself, but have other goals. For example, mutual and interaction information are used for attribute selection or split selection, as we will do in Sect. 7.3. In such circumstances, we do not seek the quantity of interaction, but whether to assume it or not. Why not assume an interaction? We are often better off assuming that a dependence between two attributes does not exist if there is not enough information to support this increase in model's complexity.

A model comparison contrasts two models: one that assumes that there is an interaction, and another one that does not. As a principled definition of what is the defining property of an interaction, we propose the notion of the 'part-to-whole' approximation.

An interaction can be understood as an irreducible whole. This is where an interaction differs from a mere dependency. A dependency may be based on several interactions, but the interaction itself is such a dependency that cannot be broken down. To dispel the haze, we need a practical definition of the difference between the whole and its reduction. One view is that the whole is reducible if we can predict it *without observing all the involved attributes at the same time*. We do not observe it directly if every measurement of the system is limited to a part of the system. In the language of probability, a view of a part of the system results from marginalization: the removal of one or more attributes, achieved by summing it out or integrating it out from the joint probability model.

Not to favor any attribute in particular, we will observe the system from all sides, but always with one or more attributes missing. Formally, to verify whether $P(A,B,C)$ can be factorized, we should attempt to approximate it using the set of all the attainable marginals: $\mathcal{M} = \{P(A,B), P(A,C), P(B,C), P(A), P(B), P(C)\}$, but not $P(A,B,C)$ it-

self. Such approximations will be referred to as *part-to-whole approximations*. If the approximation of $P(A, B, C)$ so obtained from these marginal densities fits $P(A, B, C)$ well, there is no interaction. Otherwise, we have to accept an interaction, and potentially seek a tractable model to model it.

4.4.1 Examples of Part-to-Whole Approximations

Marginal and Conditional Independence

Using the chain rule, we can decompose any joint probability model into a product of its marginalizations and conditionings, based on the assumptions of mutual and conditional independence. There are several models of that type, as we have seen in Ch. 7. In the simple case of three attributes, there are three models that involve two marginals of two attributes each: $P(B|A)P(C|A)P(A)$, $P(A|B)P(C|B)P(B)$, $P(A|C)P(B|C)P(C)$. For example, the first model assumes that B and C are independent in the context of A. There are three models that involve one marginal of two attributes: $P(A, B)P(C)$, $P(B, C)P(A)$, $P(A, C)P(B)$. Finally, there is a single model with only one-attribute marginals: $P(A)P(B)P(C)$. Neither of these approximations alone is a proper part-to-whole approximation, however, all the available parts are not employed at the same time; for example, the first model disregards the dependence between B and C.

It is possible to combine several such models in a mixture (Meilă and Jordan, 2000). However, trees only make use of two-attribute marginals, and more sophisticated models based on the chain rule and conditional independence may be used for k-way interactions $k > 3$ (Matúš, 1999).

Kirkwood Superposition Approximation

Kirkwood superposition approximation (Kirkwood and Boggs, 1942) combines all the available pairwise dependencies in closed form in order to construct a complete joint model. Matsuda (2000) phrased KSA in terms of an approximation $\hat{P}_K(A, B, C)$ to the joint probability density function $P(A, B, C)$ as follows:

$$\hat{P}_K(a, b, c) \triangleq \frac{P(a, b)P(a, c)P(b, c)}{P(a)P(b)P(c)} = P(a|b)P(b|c)P(c|a). \tag{4.19}$$

Kirkwood superposition approximation has many analogies with the *cluster-variation method* (CVM) (Yedidia et al., 2004), and the *Kikuchi approximation* of free energy (Kikuchi, 1951).

Kirkwood superposition approximation does not always result in a normalized PDF: $Z = \sum_{a,b,c} \hat{P}_K(a, b, c)$ may be more or less than 1, thus violating the normalization condition. We define the normalized KSA as $\frac{1}{Z}\hat{P}_K(a, b, c)$. It is not necessary to find a global normalizing factor: if A and B are known to have the values of a and b, we need to compute the normalizing factor only for different values of C.

Instead of normalization, one can employ an alternative to KL-divergence proposed by Cziszár (1998):

$$D'(P\|Q) \triangleq \sum_{\mathbf{x}\in\Re_{\mathbf{X}}} \left(P(\mathbf{x}) \log_2 \frac{P(\mathbf{x})}{Q(\mathbf{x})} - P(\mathbf{x}) + Q(\mathbf{x}) \right) \tag{4.20}$$

However, the properties of this divergence function are unclear.

In a general case for a k-way interaction, the Kirkwood superposition approximation for a set of attributes $\mathcal{V}$ can be phrased simply as:

$$\hat{P}_K(\mathcal{V}) \triangleq \prod_{\mathcal{T} \subset \mathcal{V}} P(\mathcal{T})^{(-1)^{1+|\mathcal{V} \setminus \mathcal{T}|}} \tag{4.21}$$

It should be quite easy to see the connection between interaction information and the non-normalized Kirkwood superposition approximation:

$$I(\mathcal{V}) = D(P(\mathcal{V}) \| \hat{P}_K(\mathcal{V})) \tag{4.22}$$

We can interpret the interaction information as the approximate weight of evidence in favor of not approximating the joint probability model with the generalized Kirkwood superposition approximation. Because the approximation is inconsistent with the normalization condition, the interaction information may be negative, and may underestimate the true loss of the approximation. Therefore, the Kirkwood superposition approximation must be normalized before computing the divergence.

Parts-as-Constraints, Loglinear Models and Maximum Entropy

The loglinear part-to-whole model employs $\mathcal{M}$ as the set of association terms (Agresti, 2002). If we view each lower-order interaction as a constraint, we may seek a joint probability model that complies with all the constraints. There may be many models that agree with the constraints, and there are two basic strategies for choosing a single one.

The observed data has the highest likelihood given the *maximum likelihood estimate*, which has to be compliant with the constraints. In this case, however, we would be observing the joint PDF and retro-fitting an approximation, which we should not, according to the above definition of a part-to-whole model, unless we assure that no information enters the model beyond what is given by the constraints. On the other hand, the *maximum entropy estimate* is the most uncertain probability density function that is still compliant with the constraints. There have been many justifications for the maximum entropy estimate, but the most important one is that it carries no additional information beyond what is provided by the constraints (Good, 1963).

The principled definition of the part-to-whole approximation is that it should be the *worst* joint model that still satisfies the given constraints. If entropy is our loss function, then the maximum entropy estimate fulfills the principle. For a different loss function, a different estimate should be used.

In the context of determining whether there exists a k-way interaction among attributes $\mathcal{V} = \{A, B, C\}$, the approximation to the joint PDF should be based on its marginals $\mathcal{M}$, but now each of them is viewed as a constraint. Because the $(k-1)$-order constraints, such as $P(A, B)$ subsume all the underlying restraints, such as $P(A)$ and $P(B)$, the resulting set of constraints is simply $\mathcal{C} = \{(P(\mathcal{V} \setminus X|\boldsymbol{\theta}));\ X \in \mathcal{V}\}$. Note that all constraints are in the context of the same parameter, $\boldsymbol{\theta}$. This common context assures that the constraints are consistent.

Although there are far more efficient constraint satisfaction procedures, we will make use of the *generalized iterative scaling* method to obtain the loglinear part-to-whole model $\hat{P}_{IS}$ (Darroch and Ratcliff, 1972):

1. Estimate $\hat{\boldsymbol{\theta}}$ from the data by whatever means preferred (maximum likelihood, Laplace estimate, picking a sample from the posterior, ...). Use the resulting joint probability model $P(\mathcal{V}|\hat{\boldsymbol{\theta}})$ to construct a consistent set of constraints $\mathcal{C}$.

2. Initialize $\hat{P}^0_{IS}(\mathbf{V})$ as the uniform distribution, which is also the unconstrained maximum entropy distribution. This assures that no information is passed to the optimization procedure: $\hat{P}^0_{IS}(\mathbf{v}) = |\Re_{\mathcal{V}}|^{-1}$ for all $\mathbf{v} \in \Re_{\mathbf{V}}$.

3. Employ iterative scaling to obtain $\hat{P}_{IS}$ from $\mathcal{C}$, iteratively cycling through $X \in \mathcal{V}$:

$$\hat{P}^{(n+1)}_{IS}(\mathbf{V}) = \hat{P}^{(n)}_{IS}(\mathbf{V}) \prod_{P(\mathcal{V}\setminus X|\hat{\boldsymbol{\theta}})\in\mathcal{C}} \frac{P(\mathbf{V}\setminus X|\hat{\boldsymbol{\theta}})}{\hat{P}^{(n)}_{IS}(\mathbf{V}\setminus X)}, \tag{4.23}$$

   where

$$\hat{P}_{IS}(\mathbf{v}\setminus X) \triangleq \sum_{\mathbf{w}_X = x,\, x\in\Re_X} \hat{P}_{IS}(\mathbf{w}),\ \mathbf{w}_Y = \mathbf{v}_Y,\, Y \in \mathcal{V}\setminus X$$

   is just another notation for marginalization.

It is not necessary to employ an iterative algorithm for part-to-whole approximations when using binary attributes. Good (1965) notes a closed form expression resembling the Fourier transform. The expression for k-way interaction derived through information geometry by Amari (2001) for the case of 3-attribute part-to-whole model appears identical. The existence of maximum likelihood estimators has recently been discussed by Eriksson et al. (2004).

Recently, the MaxEnt and Bayesian methods have been unified (Cheeseman and Stutz, 2004). One can obtain the posterior distribution of MaxEnt models simply from the joint prior distribution for the constraints. Essentially, for a given set of data, there are many imaginable configurations of constraints, and we are uncertain about which configuration to choose. However, for each configuration of constraints, there is a unique posterior MaxEnt model. A simple implementation might sample posterior constraints, and from each choice of constraints create the MaxEnt posterior that satisfies the constraints. MaxEnt methods may also be interpreted as 'maximum independence' models.

More specifically, in a Bayesian context we can define a prior $P(\boldsymbol{\Theta})$ rather than estimate $\hat{\boldsymbol{\theta}}$ from the data. We can then sample from the posterior $P(\boldsymbol{\Theta}|\mathcal{D})$. For each sample $\boldsymbol{\theta}$ we can obtain both the joint and the set of constraints that correspond to $\boldsymbol{\theta}$; we can speak of the posterior distribution of constraints $P(\mathcal{C}|\boldsymbol{\Theta})$. For the set of constraints that correspond to $\boldsymbol{\theta}$ we employ generalized iterative scaling to obtain the maximum entropy model $\hat{P}_{IS}(\mathcal{V}|\boldsymbol{\theta})$. Across the samples of $\boldsymbol{\Theta}$, this yields the posterior distribution of maximum entropy models.

There is also a Bayesian version of GIS, referred to as Bayesian IPF (Gelman et al., 2004a), that replaces the step in (4.23) by a stochastic version:

$$\hat{P}^{(n+1)}_{BIPF}(\mathbf{V}) = \frac{A}{2P(\mathbf{V}\setminus X)}\hat{P}^{(n)}_{BIPF}(\mathbf{V}) \prod_{(P,X)\in\mathcal{C}} \frac{P(\mathbf{V}\setminus X)}{\hat{P}^{(n)}_{BIPF}(\mathbf{V}\setminus X)} \tag{4.24}$$

Here, A is a random draw from a $\chi^2_{2|\Re_{\mathbf{V}\setminus X}|}$ distribution. The result is a sampling from the posterior given a conjugate Dirichlet-like prior. $P(\mathbf{V}\setminus X)$ must be a posterior based on this prior.

The MaxEnt model is the worst of the models by considering the joint entropy. In some cases, however, we employ conditional models for predicting a particular labelled attribute Y. The jointly MaxEnt model is the worst-case joint model, but may not be the worst-case conditional model. For that purpose we may maximize the conditional entropy of Y given the unlabelled attributes (Lafferty et al., 2001), or minimize the mutual information between Y and the unlabelled attributes (Globerson and Tishby, 2004), while maintaining the agreement with the parts.

Theorems of (Good, 1963) In the context of maximum entropy, it should be helpful to reproduce a somewhat more restricted form of the Theorems 2 and 3 by Good (1963): *The following propositions are equivalent:*

1. *Let $\hat{P}(\mathcal{V})$, $\mathcal{V} = \{X_1, X_2, \ldots, X_k\}$ be the maximum entropy model consistent with the constraints $\mathcal{C} = \{P(\mathcal{V} \setminus X;\ X \in \mathcal{V}\}$, and also that $\hat{P} = P$. This means that there is no k-way interaction among $\mathcal{V}$ given the model P.*

2. *$P(\mathcal{V})$ is a product of k positive functions $F_1(\mathcal{V} \setminus X_1), F_2(\mathcal{V} \setminus X_2), \ldots, F_k(\mathcal{V} \setminus X_k)$. This corresponds to the ability to express P using a Boltzmann model (Sect. 8.2.3), where there is a bijection between potentials $\phi(\mathcal{S})$ and constraints $P(\mathcal{S}) \in \mathcal{C}$.*

3. *For every configuration of values $\mathbf{v} \in \Re_{\mathcal{V}}$, there is no k-way interaction among the following k binary attributes $\dot{X}_1, \dot{X}_2, \dot{X}_k$, where:*

$$\dot{x}_i \triangleq \begin{cases} 1 & ;\ X_i = \mathbf{v}_i \\ 0 & ;\ \textit{otherwise.} \end{cases}$$

4. *There are no k-way interactions in any superset of attributes $\mathcal{V} \subseteq \mathcal{W}$.*

The first item (corresponding to Good's Theorem 2) defines an interaction with a particular choice of the maximum entropy method for obtaining the part-to-whole model. Furthermore, it provides the means of expressing the model purely through the set of its marginals. The second item in the theorem associates the absence of k-way interactions a Boltzmann model with a particular set of potentials. The third item explains that each interaction can be localized for particular attribute value combinations. The fourth item assures that no additional attribute can eliminate the existence of an interaction.

4.4.2 A Comparison of the Part-to-Whole Approximations

The advantage of loglinear models fitted by iterative scaling is that the addition of additional consistent constraints can only improve the fit of the model. Kirkwood superposition approximation has the advantage of making use of all the available data in theory, and of being phrased in closed form. For that reason, it is highly efficient. However, it does not obey any objective criteria, and its performance is not guaranteed. Finally, we can make use of the conditional independence approximations, but giving up the information about certain marginals that cannot be integrated with the chain rule.

We have taken 16 data sets from the UCI repository, and for each pair of attributes in each domain, we have investigated the 3-way interaction between the pair and the label. We compared the Kullback-Leibler divergence between the maximum likelihood joint

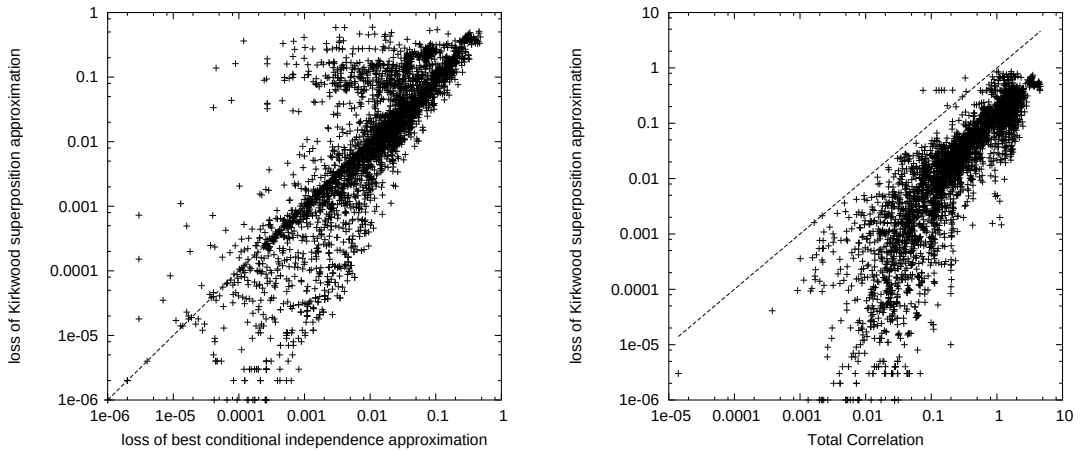


Figure 4.3: The Kirkwood superposition approximation does not always outperform the best of the conditional independence models (left), but practically always outperforms the fully factorized model $P(A)P(B)P(C)$ (right). The error of the fully factorized model is measured by total correlation.

probability model $P(A, B, C)$ and its part-to-whole approximation $\hat{P}(A, B, C)$. We have used all the three approximations: the normalized Kirkwood superposition approximation, the loglinear model, and the single best one of the three conditional independence approximations. The number of iterative scaling iterations was bounded at 10, and there were only few situations where the convergence did not occur in this time.

It turns out that the conditional independence model was somewhat more frequently worse than the Kirkwood superposition approximation (1868 vs 1411), but the average error was almost 8.9 times lower than that of the Kirkwood superposition approximation (Fig. 4.3). On the other hand, it also shows that models that include KSA may achieve better results than those that are limited to models with conditional independence.

The Kirkwood superposition approximation is high both with distinctly negative interaction information and with distinctly positive interaction information, as shown in Fig. 4.5. However, the error may be high even if the interaction information is near zero. This possibility has been discussed in Sect. 3.3.2, and arises when there is a mixture of value combinations, some involved in synergies and some in redundancies. The particular pattern of bad behavior of the Kirkwood superposition approximation can be predicted on the basis of its deviation from normalization, as shown in Fig. 4.6. This provides both an effective heuristic for deciding when to apply iterative scaling. Furthermore, positive interaction information is quite informative about the quality of the iterative scaling approximation (Fig. 4.7).

Summary of the Results The results plotted in Fig. 4.4 indicate that iterative scaling always gives better performance than either KSA or conditional independence, and is hence better suited to play the role of the part-to-whole approximation.

Apart from Good (1963), several authors have spoken in favor of a constraint-based view of part-to-whole modelling. Amari (2001) has discussed a decomposition of entropy that is based on loglinear models without interactions. Meo (2002) has proposed using such a 'maximum independence' model as the foundation for expressing a generalization

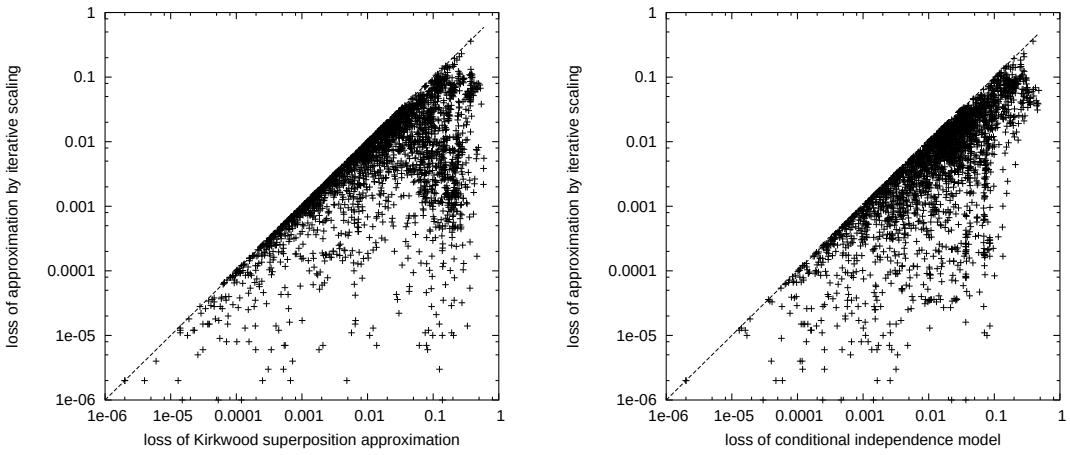


Figure 4.4: The approximation divergence achieved by iterative scaling consistently outperforms both the Kirkwood superposition approximation (left) and the best of the conditional independence models (right).

of mutual information to multiple attributes, also providing an algorithm for obtaining the part-to-whole model. Nemenman (2004) and Schneidman et al. (2004) have applied the same notion for investigating correlations in networks.

When we are building models, however, there is a trade-off involved. On one hand, we can gain in computational performance by using the Kirkwood superposition approximation or assume a pattern of conditional and mutual independencies. With these procedures, we have to include more parts, but we can work efficiently with them. On the other hand, we can use fewer parts and work with iterative methods, such as iterative scaling.

4.5 Model Comparisons

In Sect. 4.4 we have compared several methods for coming up with two models, where one that assumes an interaction and one that does not. The approximation loss $D(P\|\hat{P})$ is a measure of how important the interaction is. The approximation loss depends on what part-to-whole approximation method we use. It also depends on what probability model we employ: although we have only discussed the multinomial model, there are very many possible models.

It is exceedingly unlikely that the loss would be exactly zero, and it would seem that there are always interactions. Of course, some interactions are stronger than others. We could sort them by mutual information, and then pick them one by one. Or, we could pick them by the reduction in loss that they yield, as it has been done by Della Pietra et al. (1997). But almost everything would eventually become an interaction if we proceeded this way. And, whether something is or isn't an interaction is dependent upon the interactions already in the model. The alternative is to state that a certain amount of loss caused by ignoring the interaction is insignificant. Or, we can evaluate the loss in such a way that penalizes the complexity of the model.

We will now provide ways of testing whether an interaction is significant or not. We will employ the standard statistical hypothesis testing infrastructure, both frequentist and

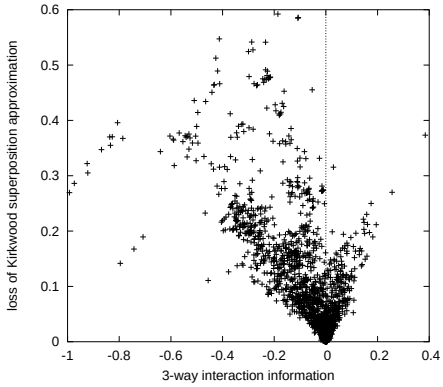


Figure 4.5: It is possible to predict the performance of the Kirkwood superposition approximation using interaction information. Both distinctly high and distinctly low interaction information is associated with high KSA divergence. However, even low interaction information may result in high KSA divergence.

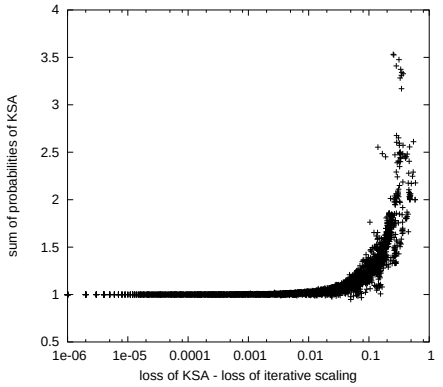


Figure 4.6: The Kirkwood superposition approximation deviates from the iterative scaling approximation when the normalization coefficient is large.

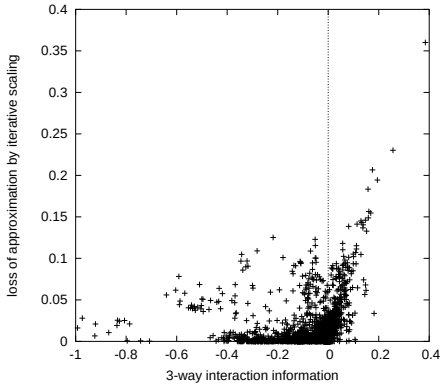


Figure 4.7: Interaction information is not particularly suggestive about the divergence of the iterative scaling approximation.

Bayesian. The fundamental idea will be that of the distribution of loss. We will address this distribution in two ways, approximately using the null model, or for two models at the same time via paired comparisons. The latter approach can be generalized to a comparison of K models at the same time.

4.5.1 The Null Model

We often have a default, preferred model. We refer to it as the *null model*, and we stick to it unless it can be refuted by the data. Usually, the default model is simple and skeptical: smoking is not associated with cancer, the drug does not cure a disease, the attribute does not predict the label. It is not necessary, however, for the null model to be simple: we can start from the assumption that working with computers hurts the eyesight. Through statistical inference we attempt to refute this model and replace it with a better one, with the *alternative model*. We achieve this is by examining the distribution of loss as expected under the null model, and examining the expected loss of the alternative model.

In the context of interactions, the null model can either be the model with the interaction or the model without the interaction. We confirm the interaction when we can reject the no-interaction model by providing a distinctly better alternative, or when we cannot reject the interaction model by providing a good-enough alternative.

The fundamental concept used in testing is perhaps best described as *self-loss*: even if the model is correct, it will generally achieve non-zero loss on a random sample from the model itself. Alternatively, if we estimate a model from the sample, it will generally deviate from the true model. The statistical testing itself is based upon examining how the approximation loss compares to the self-loss distribution. Naturally, the testing depends on the choice of the loss function. As usual, we will employ the Kullback-Leibler divergence.

The Asymptotic Approach

Assume an underlying null probability model P of categorical attributes. We then generate a sample of n instances from P, and estimate $\hat{P}$ with from relative frequencies in the sample. The KL-divergence between P and its estimate $\hat{P}$ multiplied by $2n/\log_2 e$ is equal to the Wilks' likelihood ratio statistic G^2. For large n, the G^2 in such a context follows a χ^2_{df} distribution with df degrees of freedom:

$$\frac{2n}{\log_2 e} D(\hat{P}\|P) \underset{n\to\infty}{\sim} \chi^2_{|\Re_{\mathcal{V}}|-1} \tag{4.25}$$

By the guideline (Agresti, 2002), the asymptotic approximation is poor when $n/df < 5$. For example, to evaluate a 3-way interaction of three 3-valued attributes, where $df = 26$, there should be least 135 instances. Without that much data, the comparison itself is meaningless. The choice of degrees of freedom df depends on the properties of the null model. There are two alternatives, using the null assumption of interaction, and using the null assumption of no interaction. We now describe each possibility.

Null: Interaction The Pearson's approach to selecting the number of degrees of freedom disregards the complexity of the approximating model. $df = |\Re_{\mathcal{V}}| - 1$ is based on the cardinality of the set of possible combinations of attribute values $|\Re_{\mathcal{V}}|$. $\Re_{\mathcal{V}}$ is a subset

of the Cartesian product of ranges of individual attributes. Namely, certain value combinations are impossible, where the joint domain of two binary attributes A and B, where $b = \neg a$, $\mathcal{V}$ should be reduced to only two possible combinations, $\mathcal{V} = \{(a, \neg b), (\neg a, b)\}$. The impossibility of a particular value conjunction is often inferred from the zero count in the set of instances, and we followed this approach in this paper. A more detailed discussion of these *structural zeros* appears in (Krippendorff, 1986).

The P-value ϕ (or the weight of evidence for rejecting the null model) is defined to be

$$\phi \triangleq \Pr\left\{\chi^2_{df}(x) \geq \frac{2n}{\log_2 e} D(P\|\hat{P})\right\} \tag{4.26}$$

The P-value can also be interpreted as the probability that the average loss incurred by P on an *independent* sample from the null Gaussian model approximating the multinomial distribution parameterized by P itself, is greater or equal to the average loss incurred by the approximation $\hat{P}$ in the original sample.

This P-value can be interpreted as the lower bound of P-values of all the approximations. Using *df* assures us that no simplification would be able to reduce the P-value, regardless of its complexity.

Null: No Interaction In Fisher's scheme (Fisher, 1922), we are not examining the self-loss of the interaction-assuming null. Instead, we are examining the self-loss of the no-interaction-assuming null. Asymptotically, it also results in a χ^2 distribution, but with a different setting of the degrees of freedom. The residual degrees of freedom should instead be $df' = \prod_{X \in \mathbf{V}} (|\Re_X| - 1)$.

For discrete data there are two ways of interpreting the no-interaction null. We can maintain the marginal counts and re-shuffle the values among instances. This is the foundation for the hypergeometric or multiple hypergeometric model of independence. Alternatively, we can fix marginal probabilities and then form the joint probability model under the assumption of independence: the model is a product of multinomials.

The Resampling Approach

Bootstrapped P-Values We can randomly generate independent *bootstrap samples* of size n' from the original training set. Each bootstrap sample is created by randomly and independently picking instances from the original training set with replacement. This non-parametric bootstrap corresponds to an assumption that the training instances themselves are samples from a multinomial distribution, parameterized by P. For each bootstrap sample we estimate the model P^*, and compute the loss incurred by our prediction for the actual sample $D(P^*\|P)$. We then observe where $D(P\|\hat{P})$ lies in this distribution of self-losses. The P-value is $P(D(P^*\|P) \geq D(P\|\hat{P}))$ in the set of bootstrap estimates P^*. If we perform B replications $P^{*1}, \ldots, P^{*B}$, we obtain the bootstrapped P-value as a probability from the proportion of resamples where the alternative model outperformed the null model. We employ the uniform prior to estimate the probability from the frequency:

$$\hat{P}(D(P^*\|P) \geq D(P\|\hat{P})) \triangleq \frac{\#_{i=1}^{B}\{D(P^{*i}\|P) \geq D(P\|\hat{P})\} + 1}{B + 2} \tag{4.27}$$

This way we prevent the extreme P-values of 0 and 1 from occurring with a finite number of bootstrap resamples.

The bootstrap sample size n' is a nuisance parameter which affects the result: the larger value of n', the lower the deviation between P' and P. The P-value is conditional on n'. Usually, the size of the bootstrap sample is assumed to be equal to the original sample $n' = n$. The larger the n', the more likely is the rejection of an approximation with the same KL-divergence.

The default nonparametric bootstrap corresponds to sampling from the null multinomial interaction model, e.g., $P(A, B)$. We can perform the parametric bootstrap from $P(A)P(B)$ in order to treat the no-interaction model as the null.

Permutation Tests While bootstrap generates independent samples with the dice-tossing machine, the permutation tests start from the original data, breaking down the structure in data in particular ways (Frank and Witten, 1998). For example, we can permute the values for each attribute independently, maintaining the frequency distribution of the values, but eliminating any kind of dependence between the values of two attributes.

Permutation testing only works for the no-interaction model as the null, and corresponds to the multiple hypergeometric distribution. The exhaustive computation using the odds ratio as a 'loss' function on a 2×2 contingency table in this manner corresponds to Fisher's exact test of independence.

The Bayesian Approach

Suppose that we have two models, $\mathbf{\Theta}_1$ and $\mathbf{\Theta}_2$ for the same data. This should be interpreted as having two hypothesis spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, where $\mathbf{\Theta}_1$ is in the context of $\mathcal{H}_1$ and $\mathbf{\Theta}_2$ in the context of $\mathcal{H}_2$. We can examine the divergence between posteriors as follows:

$$\Pr\{D_{\mathbf{\Theta}_1|\mathbf{\Theta}_2}(P\|P) \leq w\} \triangleq \Pr\{D(P(\mathbf{X}|\mathbf{\Theta}_1)\|P(\mathbf{X}|\mathbf{\Theta}_2)) \leq w\} \propto \iint \mathbb{I}\{D(P(\mathbf{X}|\boldsymbol{\theta}_1)\|P(\mathbf{X}|\boldsymbol{\theta}_2)) \leq w\} P(\boldsymbol{\theta}_1|\mathcal{D})P(\boldsymbol{\theta}_2|\mathcal{D})d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_2 \quad (4.28)$$

We may build these models with the same prior and with the same data, but independently one of the another. This means that they share the hypothesis space, and it implies $P(\mathbf{\Theta}_1 = \boldsymbol{\theta}) = P(\mathbf{\Theta}_2 = \boldsymbol{\theta})$ and $P(\mathbf{\Theta}_1 = \boldsymbol{\theta}|\mathcal{D}) = P(\mathbf{\Theta}_2 = \boldsymbol{\theta}|\mathcal{D})$. In such a case, the self-loss distribution is described by $\Pr\{D(P_{\mathbf{\Theta}|\mathbf{\Theta}}(P\|P) \leq w)$.

This approach can be contrasted with the definition of a Bayesian P-value (Gelman et al., 2004a) based on a statistic T and on sampling new data sets from the posterior model:

$$p_B = \iint \mathbb{I}\{T(\mathcal{D}^*, \boldsymbol{\theta}) \geq T(\mathcal{D}, \boldsymbol{\theta})\} P(\mathcal{D}^*|\boldsymbol{\theta})P(\boldsymbol{\theta}|\mathcal{D})d\mathcal{D}^* d\boldsymbol{\theta} \quad (4.29)$$

The B-values do not involve any sampling of data sets, just of parameters.

Making Decisions with P-Values

On the basis of the thus obtained P-value, we can decide whether an interaction exists or not. P-value identifies the probability that the loss of $D(P\|\hat{P})$ or greater is obtained by the *true* model on a finite sample. For example, the P-value of 0.05 means that the loss incurred by the null model will be greater or equal to the loss obtained by the approximation $\hat{P}$ on the training sample in 5 independent samples out of 100 on the

average. P-value thus only provides a measure of robustness of the interaction, but not of its importance.

Now assume that the P-value is ϕ. With the interaction null, we may classify the situation into two types: the interaction is discovered when $\phi \leq \alpha$, and the interaction is rejected when $\phi > \alpha$. The lower the value of α, the more unwilling we are to choose the alternative model. It must be noted that α does *not* directly correspond to the the Type I error probability.

Confidence Intervals

A P-value measures how unusual is the error of the alternative model as compared to the distribution of the null model. In many situations, however, it is interesting to examine the distribution of the alternative model's error. Although we have examined the distribution of mutual information in Sect. 4.3, the problem has to be interpreted in a different sense: in terms of the difference between the loss achieved by the alternative model as compared to the loss of the null model across the circumstances.

Although the approach that we will present can be applied in a variety of circumstances, let us adopt the context of using bootstrap under the assumption of the interaction null. Under the interaction null we generate a number of resample-derived models $P^{*1}, \ldots, P^{*B}$. A symmetric 99% confidence interval of the alternative model loss is based on two numbers $w_<$ and $w_>$, so that

$$\Pr\{D(P^*\|\hat{P}) - D(P^*\|P) \leq w_<\} = (100\% - 99\%)/2 \quad \wedge \tag{4.30}$$

$$\Pr\{D(P^*\|\hat{P}) - D(P^*\|P) \geq w_>\} = (100\% - 99\%)/2 \tag{4.31}$$

The probability that the difference results in a negative number corresponds to the P-value. A practical implementation will obtain a number of bootstrap resamples, record a list of the performance differences $D(P^*\|\hat{P}) - D(P^*\|P)$, sort them, and retrieve $w_<$ and $w_>$ through appropriate quantiles.

4.5.2 Paired Comparison

A disadvantage of the null model significance testing approach is that the alternative and the null models are estimated only once. We can address this problem either by performing cross-validation and learning both the alternative and the null models and comparing them. Alternatively, we can follow a Bayesian approach by assuming independence between the null and the alternative models.

Cross-Validation

For each replication and fold, the set of instances $\mathcal{D}$ is partitioned into subsets for testing ($\dot{\mathcal{D}}$) and for training ($\mathcal{D}'$). For F folds, we have F configurations $\{\dot{\mathcal{D}}^1, \mathcal{D}'^1, \ldots, \dot{\mathcal{D}}^F, \mathcal{D}'^F\}$. As in bootstrap, the number of folds is a nuisance parameter, but with the choice of 2 folds, the training and the test set have equal size, so the probability estimates are equally reliable in both. To avoid the dependence on the particular choice of the partitioning, we should perform B replications of cross validation $\{\dot{\mathcal{D}}^{1,\cdot}, \mathcal{D}'^{1,\cdot}, \dot{\mathcal{D}}^{B,\cdot}, \mathcal{D}'^{B,\cdot}\}$.

For each pair of subsets, we estimate two null probability models: the training null model P' and the testing null model $\dot{P}$. We then also construct the part-to-whole approximation $\hat{P}'$. The CV-value ν is defined as:

$$\nu \triangleq \Pr\{D(\dot{P}\|\hat{P}') \geq D(\dot{P}\|P')\} \tag{4.32}$$

in a large set of cross-validated estimates of $(p', \hat{p}', \dot{p})$. In the context of cross-validation, we estimate the number of times the null model achieved a greater loss than the approximation model in predicting the test set. This way we penalize the variance of the probability model.

The Bayesian Approach

Bayes Factor The usual approach to Bayesian hypothesis testing are Bayes factors. Assuming two models of data $\mathcal{D}$ parameterized by $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$, the *Bayes factor* is defined as (Bernardo and Smith, 2000):

$$BF_{\mathcal{D}}(\boldsymbol{\Theta}_1\|\boldsymbol{\Theta}_2) \triangleq \frac{P(\boldsymbol{\Theta}_1)\int P(\boldsymbol{\theta}_1)P(\mathcal{D}|\boldsymbol{\theta}_1)d\boldsymbol{\theta}_1}{P(\boldsymbol{\Theta}_2)\int P(\boldsymbol{\theta}_2)P(\mathcal{D}|\boldsymbol{\theta}_2)d\boldsymbol{\theta}_2} \tag{4.33}$$

The prior beliefs in each model are $P(\boldsymbol{\Theta}_1)$ and $P(\boldsymbol{\Theta}_2)$. If the testing is objective, these beliefs are equal, and sometimes the Bayes factor is defined already with this assumption (Gelman et al., 2004a, Berger, 2003). The Bayes factor may be calculated for any data set, or even for an individual instance.

B-values The Bayes factor is a scalar value, and the variation of the posterior model in the context of the data is not included into consideration. There are means of addressing this issue, such as DIC (Speigelhalter et al., 2003). However, we can extend the Bayesian self-loss distribution (4.28) also in the context of comparing the null, an independent null and the alternative model.

The expected self-divergence $\mathbb{E}\{D_{\boldsymbol{\Theta}|\boldsymbol{\Theta}}(P\|P)\}$ of a reference posterior model $P(\boldsymbol{\Theta}|\mathcal{D})$ should be an order of magnitude lower than the expected loss of a model $\hat{P}$ based with respect to the reference model P, $\mathbb{E}\{D_{\boldsymbol{\Theta}|\hat{\boldsymbol{\Theta}}}(P\|\hat{P})\}$. If this is not the case, the reference model would be too complex given the data, and the variance of the estimate is not low enough to reliably estimate the bias.

Both comparisons can be joined into a unique probability corresponding to a B-value β for the comparison between the null posterior $P(\boldsymbol{\Theta}|\mathcal{D})$ and the alternative posterior $P(\hat{\boldsymbol{\Theta}}|\mathcal{D})$, where KL-divergence is used instead of a test statistic. There is no need to draw samples from the model as with Bayesian P-values because KL-divergence compares probabilistic models directly. The probability that $P(\hat{\boldsymbol{\Theta}}|\mathcal{D})$ is worse than an independent $P(\boldsymbol{\Theta}'|\mathcal{D})$, where $\boldsymbol{\Theta} = \boldsymbol{\Theta}'$, is:

$$\beta \triangleq \iiint \mathbb{I}\{D_{\boldsymbol{\theta}|\boldsymbol{\theta}'}(P\|P') \geq D_{\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}}(P\|\hat{P})\}P(\boldsymbol{\theta}|\mathcal{D})P(\boldsymbol{\theta}'|\mathcal{D})P(\hat{\boldsymbol{\theta}}|\mathcal{D})d\boldsymbol{\theta}d\boldsymbol{\theta}'d\hat{\boldsymbol{\theta}} \tag{4.34}$$

The B-values do not have several of the properties of P-values.

4.5.3 Multiple Comparisons

P-values are said to be meaningful if performing a single hypothesis test, but an analysis of the whole domain involves a large number of tests. We have to account for the consequently increased risk of making an error in any of them. The best-case approach is to assume that all P-values are perfectly correlated, and we can use them without adjustment. The worst-case approach is to assume that all P-values are perfectly independent, and adjust them with Bonferroni correction.

A number of researchers have proposed ways of avoiding this problem, such as the false detection rate paradigm of Benjamini and Hochberg (1995) and the q-values by Storey (2003). The fundamental idea is to define a measure of error across all the decisions, and then assume the decisions to be independent. It can still be seen as an oversimplification to assume independence between models, especially as these models relate to the same data and potentially to the same attributes. This oversimplification can be demonstrated by either duplicating a significant attribute (where a single attribute will now have two P-values), or by adding a perfectly independent and random attribute (which will affect the P-values of non-random attributes in the model). Methods have been developed that account for assumptions about the pattern of dependence (Yekutieli and Benjamini, 1999).

Without such assumptions, multiple testing can be examined in our earlier framework: all that needs to be changed is the indicator function. Correct multiple testing becomes increasingly difficult with a large number of attributes without making strong prior assumptions, because a complex global model needs to be built. For example, the B-value corresponding to the test of the conjunction of statements "$\mathbf{X}$ is independent of $\mathbf{Y}$" and "$\mathbf{X}$ is independent of $\mathbf{Z}$" would have to be based on the model $P(\mathbf{X}, \mathbf{Y}, \mathbf{Z}|\mathcal{D})$ conditioned on the prior $\hat{\boldsymbol{\Theta}}$ and an independent pair of identical priors $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$:

$$\iiint \mathbb{I}\{D_{\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}}(P(\mathbf{X},\mathbf{Y})\|P(\mathbf{X})P(\mathbf{Y})) \leq D_{\boldsymbol{\theta}|\boldsymbol{\theta}'}(P(\mathbf{X},\mathbf{Y})\|P(\mathbf{X},\mathbf{Y})) \wedge D_{\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}}(P(\mathbf{X},\mathbf{Z})\|P(\mathbf{X})P(\mathbf{Z})) \leq D_{\boldsymbol{\theta}|\boldsymbol{\theta}'}(P(\mathbf{X},\mathbf{Z})\|P(\mathbf{X},\mathbf{Z}))\} P(\boldsymbol{\theta}|\mathcal{D})P(\boldsymbol{\theta}'|\mathcal{D})P(\hat{\boldsymbol{\theta}}|\mathcal{D})d\boldsymbol{\theta}d\boldsymbol{\theta}'d\hat{\boldsymbol{\theta}} \quad (4.35)$$

The same approach can be used to estimate the error rate. Furthermore, in appropriate circumstances, the bootstrap may be used in the place of the Bayesian integral over the parameter space.

4.5.4 Value at Risk

The usefulness of P-values is somewhat restricted. A P-value is only an indication of how likely it is that using the interaction will not be beneficial. At the same time, we do not know how large will the benefit be: it might be risk-free but trivially low. For that reason, the concept of *value at risk* (VaR) is interesting and important. The value of risk at the confidence level ϕ is the level under which the utility will drop only with the probability of ϕ (Cherubini et al., 2004). Synonymously, it is the level of losses that will only be exceeded with the probability of ϕ. If we employ the KL-divergence as our loss function, the VaR can be defined as:

$$\Pr\{D(P\|\hat{P}) \geq \mathrm{VaR}_{\phi,\hat{P}}\} = \phi \quad (4.36)$$

In effect, VaR is a one-tailed confidence interval, so that the tail reaches towards pessimism. The justification is that in the worst case, an act will always result in a net loss. However, we can attempt to bound our risks so that the net loss will occur only rarely. VaR is directly associated with the utility, yet it has been discounted to include risk. Expected utility does not distinguish between high and low risk, while P-values do not distinguish between high and low utilities. VaR combines the benefits of both measures.

A certain model of vagueness needs to be presupposed in order to come up with the distribution of KL-divergence. We may apply the vagueness of loss approach if we desire efficiency. The self-loss distribution can be used if we trust the model P. If we are forced to approximate with $\hat{P}$, we may examine the distribution of loss that captures the imperfection of $\hat{P}$ approximating the reference P.

4.5.5 Anomalies and Testing Procedures

There are several problems with certain testing procedures applied to extreme cases. Partly the problems may be remedied using paired testing which includes the variation to the alternative model, but not always.

Perfect Independence

Consider the following discrete coin toss data set with coins A and B, resembling the experimental setup of (Frank and Witten, 1998):

A	B
H	H
H	T
T	H
T	T

1. Assume the null model to be $P(A, B)$ and the alternative to be $P(A)P(B)$. This is the context for the goodness-of-fit with Pearson's protocol, and for the bootstrap with infinitely many replications. The resulting P-value will tend towards 1.0: an impossibly small value given the circumstances.

2. Assume the null model to be $P(A)P(B)$ and the alternative to be $P(A, B)$. This is the context for permutation tests with infinitely many replications. The resulting P-value also tends towards 1.0: impossibly large, as it is foreseeable that the two coins are indeed dependent.

Fortunately, perfect independencies break down through paired testing.

Perfect Dependence

Consider this discrete coin toss data set with coins A and B (Frank, 2004):

A	B
H	H
H	H
T	T
H	H
T	T
T	T

It is a rather small sample, and the two coins always matched perfectly: it can be imagined that such a situation may occur purely by chance. The coins are matched perfectly in every resample also if we use cross validation or bootstrap. So cross validation and bootstrap will be overconfident about the dependence. Therefore, cross-validation cannot prevent certain types of overfitting, something that we have already noticed in Sect. 7.3.1.

4.5.6 An Empirical Comparison

We compared the 2-way interactions between each attribute and the label in several standard benchmark domains with the number of instances in the order of magnitude of 100: 'soybean-small', 'lung', 'horse-colic', 'post-op', 'lymphography' and 'breast-cancer'. In domains with more instances, the 2-way interactions are practically always significant, which means that there is enough data to make it worth to model them. But on such small domains, it is sometimes better to disregard weakly relevant attributes, as they may cause overfitting.

Comparing χ^2 and resampled P-values We examined the similarity between the P-values obtained with the assumption of χ^2 distribution of KL-divergence, and the P-values obtained through the bootstrap procedure. The match shown in Fig. 4.8 is good enough to recommend using χ^2-based P-values as a reasonable heuristic which perhaps tends to slightly underestimate. The number of bootstrap samples was 10000. The agreement between the permutation test and the χ^2-based P-values in Fig. 4.9 is also close, but less reliable.

On the difference between P-values and cross-validated CV-values We compared the P-values obtained with bootstrap with similarly obtained CV-values, using 500 replications of 2-fold cross-validation. The result is illustrated in Fig. 4.10 and shows that the two estimates of significance are correlated, but behave somewhat differently. P-values are more conservative, while very low and very high P-values do not guarantee an improvement or deterioration in CV performance. Although CV-values might seem intuitively more appealing (even if the number of folds is another nuisance parameter), we are not aware of suitable asymptotic approximations that would allow quick estimation. The permutation test seems not to be as predictive of CV-values as bootstrap, as is shown in Fig. 4.11.

P-values and cross-validated performance We have employed cross-validation to verify whether a classifier benefits from using an attribute, as compared to a classifier based just on the prior label probability distribution. In other words, we are comparing classifiers $P(Y)$ and $P(Y|X = x) = P(Y, X = x)/P(X = x)$, where Y is the label and X

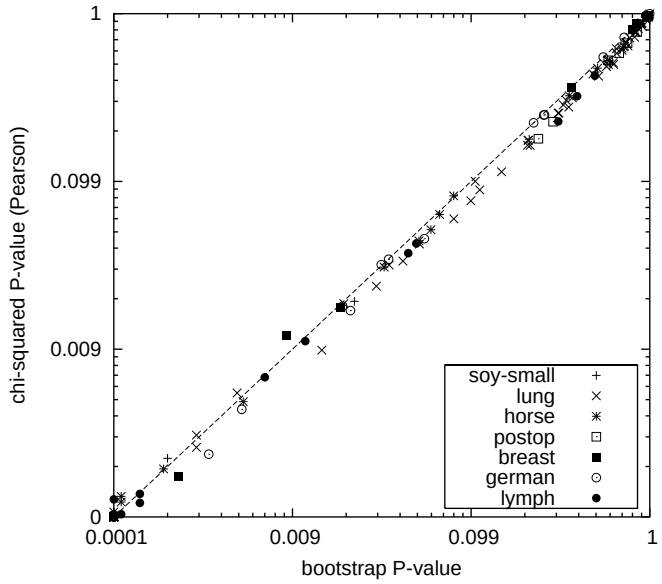


Figure 4.8: A comparison of P-values estimated by using the bootstrap and by assuming the χ^2 distribution with $df = |\Re_{\mathcal{V}}| - 1$.

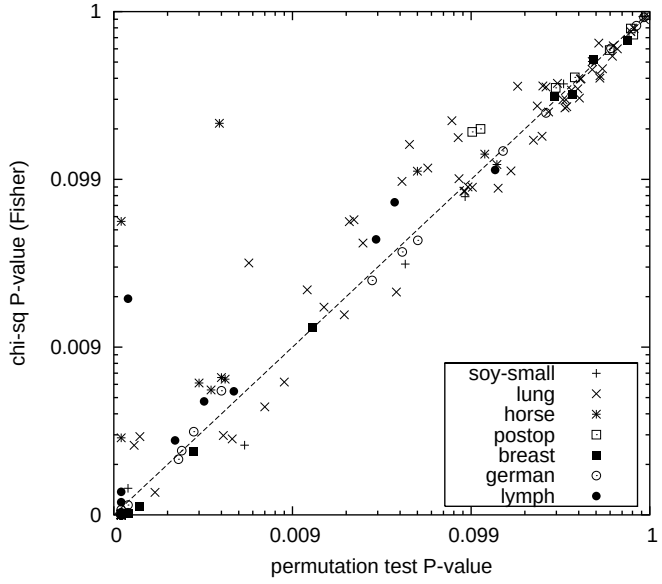


Figure 4.9: A comparison of P-values estimated by using the permutation test and by assuming the χ^2 distribution with $df = (|\Re_A| - 1)(|\Re_B| - 1)$.

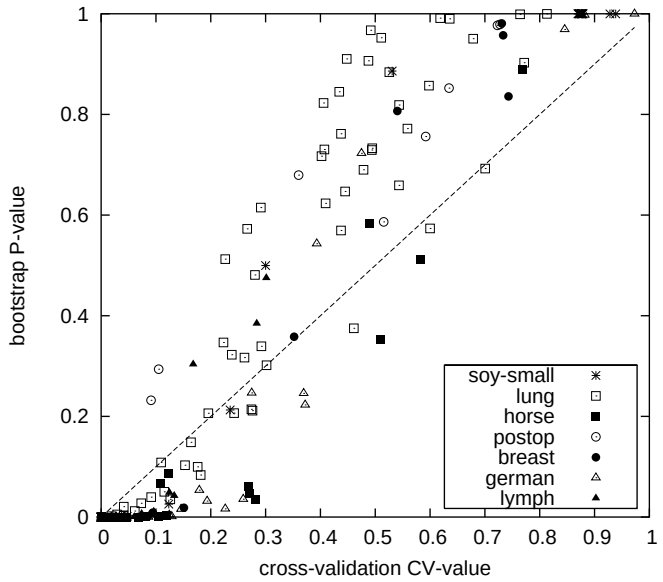


Figure 4.10: A comparison of P-values estimated with the bootstrap with the probability that the test set loss of the interaction-assuming model was not lower than that of the independence-assuming one in 2-fold cross-validation (CV-value).

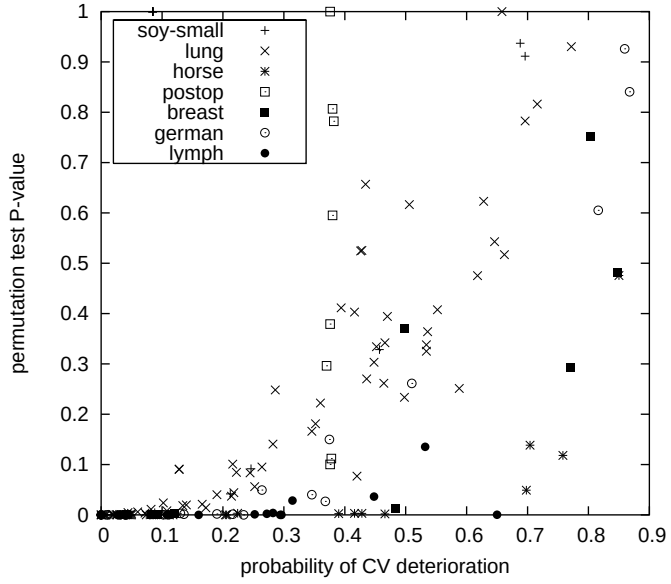


Figure 4.11: A comparison of P-values estimated with the permutation test with the probability that the test set loss of the interaction-assuming model was not lower than that of the independence-assuming one in 2-fold cross-validation (CV-value).

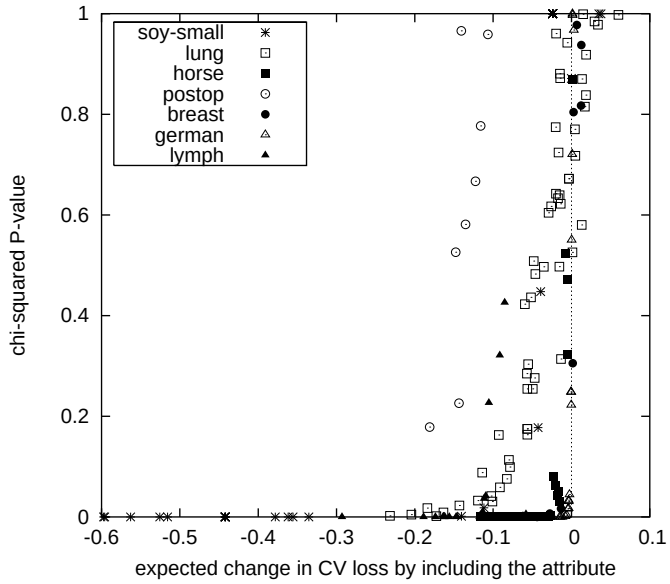


Figure 4.12: A comparison of P-values assuming χ^2 distribution with the average change in log-likelihood of the data given the information about the attribute value.

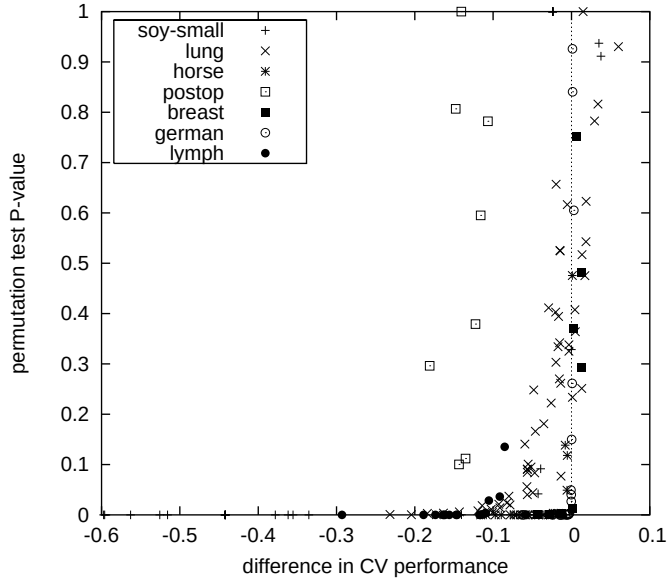


Figure 4.13: A comparison of P-values from the permutation test with the average change in log-likelihood of the data given the information about the attribute value.

is an attribute. The loss function was the expected change in negative log-likelihood of the label value of a test set instance when given the instance's attribute value. This way, we use no probabilistic model of the testing set $\dot{p}$, but instead merely consider instances as samples from it. The probabilities were estimated with the uniform prior to avoid zero probabilities and infinitely negative log-likelihoods. We employed 2-fold cross-validation with 500 replications. The final loss was the average loss per instance across all the instances, folds and replications. The results in Fig. 4.12 show that the goodness-of-fit P-value was a very good predictor of the increase in loss. The permutation test seems not to perform as well again (Fig. 4.13).

Focusing on the results obtained with the goodness-of-fit P-values, the useful attributes appear on the left hand side of the graph. If we pick the first 100 of the 173 total attributes with $\phi < 0.3$, there will not be a single one of them that would increase the loss. On the other hand, if we picked the first 100 attributes on the basis of mutual information or information gain, we would end up with a deterioration in 7 cases, which is still a two-fold improvement upon the base rate, where 14.4% of all the attributes yield a deterioration in this experiment.

On the other hand, it must be noted that 60% of the 30 most insignificant attributes with $\phi > 0.9$ also result in a decrease of prediction loss! The cut-off used for detecting overfitting through an increase in loss by cross-validation is obviously somewhat ad hoc, especially as both CV-values and P-values turned to be largely equivalent in this experiment. For that reason we should sometimes be skeptical of the performance-based results of cross-validation. Significance can be seen as a necessary condition for a model, carrying the aversion to chance and complexity, but not a sufficient one, neglecting the expected performance difference.

B-values and cross-validated performance We have assumed the uniform prior both for the joint multinomial model of $P(A, B)$ and for each of the marginals $P(A), P(B)$. There are now three independent models, sampled from the posterior: the reference joint model, the independence-assuming model and the dependence-assuming model. The B-value indicates the proportion of times the independence-assuming model matched or outperformed the dependence-assuming model.

The results in Figs. 4.14 and 4.15 show that it is a reasonably associated with cross-validation. More interesting, however, is the clustering of domains. The fact that some domains are skewed to the left or to the right indicates that data set size influences the difference between B-values and cross-validation.

Summary From experiments we made, there seems to be a difference between the bootstrap formulation and the cross-validated formulation of hypothesis testing, but the two are not considerably different when it comes to judging the risk of average deterioration. This conclusion has been disputed, but a tenable explanation for our results could be that all our evaluations were based on the Kullback-Leibler divergence, while earlier experiments tried to employ statistical testing based on probabilistic statistics for improving classification performance assessed with a conceptually different notions of classification accuracy (error rate) or instance ranking (area under the ROC).

The permutation test seems to differ in its nature from both the bootstrap and the cross-validation, and it proved to be a less reliable heuristic. Furthermore, it was not

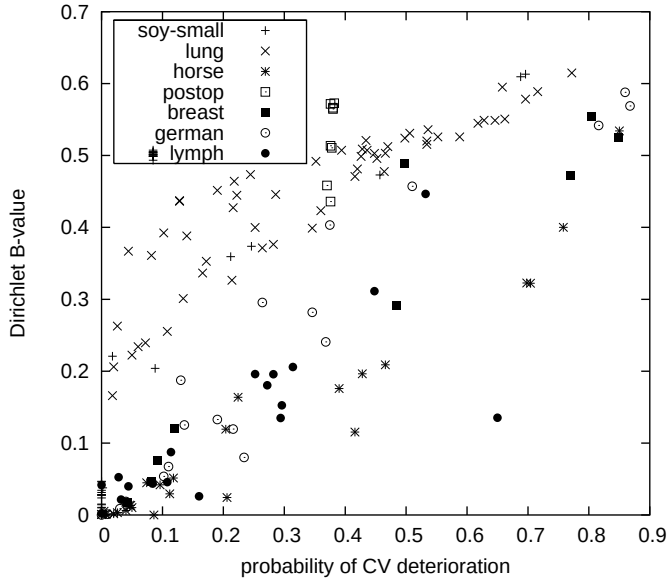


Figure 4.14: A comparison of B-values based on the uniform prior and the multinomial model with the CV-values obtained through 2-fold cross-validation.

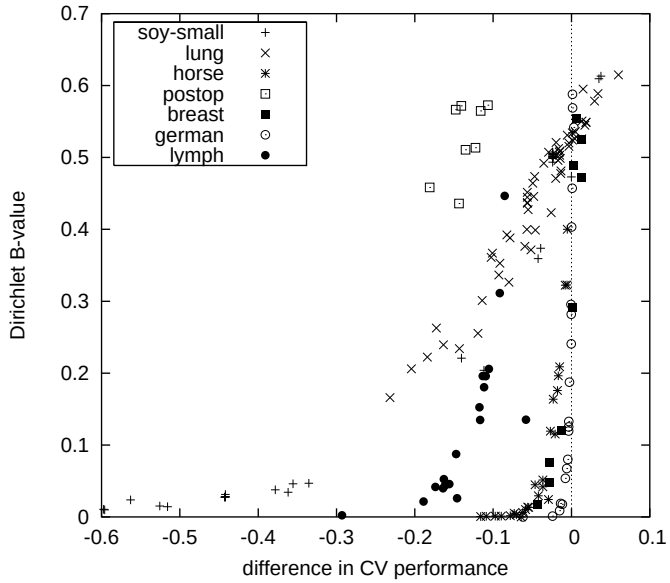


Figure 4.15: A comparison of B-values based on the uniform prior and the multinomial model with with the average change in log-likelihood of the data given the information about the attribute value.

reliably approximated by the χ^2 distribution. The best-performing were B-values, and it would be interesting to find more efficient approximations.

We should choose the evaluation method based on the reasonableness of the assumptions it makes, rather than try to pick a favorite (such as cross-validation or the Bayesian approach) and then attempt to approximate it by other methods. The comparisons we have made are intended to demonstrate similarities and differences between the methods. Our formulation of the significance testing methods includes the assumption of the loss function, which is used as a statistic.

CHAPTER 5

Interactions among Continuous Attributes

5.1 Differential Entropy

Although entropy is often computed for an attribute or a set of them, Shannon did not define entropy for attributes, but for a joint probability model of the attributes, a particular joint probability mass function P. Entropy should be seen as a characteristic of a model and not of an attribute or a data set. That is why expressing entropy as $H(A)$ is somewhat misleading; a more appropriate expression is $H(A|P, \mathcal{D})$, where $\mathcal{D}$ is the data and P is the probability model $P(A|\boldsymbol{\theta}, \mathcal{D})$, the one actually used for computing the entropy. Although we will not always express entropy conditionally, the assumptions are always implicit: entropy is usually computed by assuming a maximum likelihood multinomial model.

Entropy H is defined for probability mass functions, not for probability density functions. For a multivariate joint density function p modelling an attribute vector $\mathbf{X}$, we can define a somewhat different concept of *differential entropy* h, also measured in bits (Cover and Thomas, 1991):

$$h(\mathbf{X}|p) \triangleq -\int_{\Re_{\mathbf{X}}} p(\mathbf{x}) \log_2 p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_p\{-\log_2 p(\mathbf{x})\} \tag{5.1}$$

The properties of differential entropy do not fully match those of ordinary entropy (Shannon, 1948). For example, differential entropy may be negative or even zero. For example:

$$\sigma \leq 1/\sqrt{2\pi e} : h(X|X \sim \text{Normal}(\mu, \sigma)) \leq 0 \tag{5.2}$$

Differential entropy is also sensitive to the choice of the coordinate system. Nonetheless, the magnitude of entropy and the sign of changes in entropy remain meaningful: the higher the entropy, the harder the predictions. Entropy should be understood as the expected loss of the model, given the model itself. Shannon entropy results from the choice of the logarithmic loss function. Other loss or utility functions may be employed and a corresponding generalized notion of entropy thus derived (Grünwald and Dawid, 2004), but its properties might not match those of Shannon entropy.

An analytical derivation of differential entropy has been made only for a few model families. Therefore, *empirical entropy* (Yeung, 2002), sometimes also referred to as sample

entropy (Roberts et al., 1999), often proves to be a useful approximation. If the data is $\mathcal{D}$ sampled from $\Re_{\mathbf{X}}$, a probability model $p(\mathbf{X}|\mathcal{D})$ can be learned from it. If the modelling is reliable, $\mathcal{D}$ can be understood as a representative random sample drawn from $\Re_{\mathbf{X}}$ using p. The approximation to h is the expected negative log-likelihood of the model p:

$$\hat{h}(\mathbf{X}|p,\mathcal{D}) \triangleq -\frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \log_2 p(\mathbf{x}). \tag{5.3}$$

The resulting differential empirical entropy is the average negative log-likelihood of the model p. Observe that $1/|\mathcal{D}|$ is the probability of choosing a certain instance in $\mathcal{D}$. The resulting sum can then be understood as the expectation of entropy given a uniform probability distribution over the data: all instances in $\mathcal{D}$ have equal probability, and those outside are impossible.

KL-divergence or relative entropy $D(p\|q)$ (Kullback and Leibler, 1951) assesses the difference between two probability density functions p and q:

$$D(p\|q) \triangleq \int_{\Re_{\mathbf{X}}} p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{5.4}$$

KL-divergence is zero only when the two functions are equal. It is not a symmetric measure: P is the *reference* model, and the KL-divergence is the expected loss incurred by the *alternative* model Q when approximating P. We can understand empirical entropy through KL-divergence. If $U_{\mathcal{D}}$ is the uniform probability mass function on the data $\mathcal{D}$:

$$\hat{H}(\mathbf{X}|P,\mathcal{D}) = D(U_{\mathcal{D}}\|P) - H(U_{\mathcal{D}}), \tag{5.5}$$

$$U_{\mathcal{D}}(\mathbf{x}) \triangleq 1 - \frac{|\mathcal{D}\setminus\{\mathbf{x}\}|}{|\mathcal{D}|} \tag{5.6}$$

The same formula can be used to compute the differential empirical entropy of a PDF p, mixing probability mass and probability density functions, but ultimately yielding the same result as (5.3). If we interpret entropy as defined using KL-divergence, some problems of differential entropy, such as negativity, would be remedied with a better choice of the reference model U.

An important connection between entropy and KL-divergence appears when q is a marginalization of p: $\int p(x,y)dx = q(y)$. In such a case, $D(p\|q) = h(p) - h(q)$. If q is a factorization of p, the KL-divergence can be expressed as a sum of entropies. Generally, the KL-divergence between p and any product of probability mass or density functions obtained by conditioning or marginalization p is expressible by adding or subtracting entropies of p's marginals. For example, the divergence between $p(X,Y,Z)$ and $q(X,Y,Z) = p(X|Z)p(Y|Z)p(Z)$ is

$$D(p(X,Y,Z)\|p(X|Z)p(Y|Z)p(Z)) = \\ h(X,Z|p) + h(Y,Z|p) - h(Z|p) - h(X,Y,Z|p) \tag{5.7}$$

Mutual and conditional mutual information provide a short-hand notation, in this case: $D(p\|q) = I(X;Y|Z)$. Conditional and marginal entropies can be calculated through KL-divergence. Assuming $\mathbf{x} = [\mathbf{a},\mathbf{b},\mathbf{c}]$, marginalization over $\mathbf{C}$, the entropy of $\mathbf{a}$ conditioned on $\mathbf{B}$ is $h(\mathbf{A}) - h(\mathbf{A},\mathbf{B})$ or:

$$h(\mathbf{A}|\mathbf{B}) = \int_{\Re_{\mathbf{x}}} p(\mathbf{x}) \log_2 p(\mathbf{a}_{\mathbf{x}}|\mathbf{b}_{\mathbf{x}}) d\mathbf{x} = D(p(\mathbf{A},\mathbf{B})\|p(\mathbf{B})) \tag{5.8}$$

Using conditional KL-divergence, it is possible to compare two conditional probability density functions, something particularly useful in supervised learning:

$$D(p(\mathbf{X}|\mathbf{Y})\|q(\mathbf{X}|\mathbf{Y})) = \iint p(\mathbf{x},\mathbf{y}) \log_2 \frac{p(\mathbf{x}|\mathbf{y})}{q(\mathbf{x}|\mathbf{y})} d\mathbf{x}d\mathbf{y} \tag{5.9}$$

Observe, however, that conditional KL-divergence cannot be computed without a joint probability model of both $\mathbf{X}$ and $\mathbf{Y}$, $p(\mathbf{X},\mathbf{Y})$.

Fisher Information and Invariantized Entropy

The fact that differential entropy is sensitive to parametrization is a major hindrance to the applicability of differential entropy. For that reason, Good (1983) proposed *invariantized entropy* with the help of *Fisher's information matrix*. For an appropriate probability model $p(\mathbf{X}|\boldsymbol{\Theta})$, where $\boldsymbol{\Theta}$ is a k-dimensional vector: $\boldsymbol{\Theta} = [\theta_1, \theta_2, \ldots, \theta_k]^T$. We define Fisher's information matrix $\mathsf{I}(\boldsymbol{\theta})$ as a $k \times k$ matrix (Bernardo and Smith, 2000):

$$(\mathsf{I}(\boldsymbol{\theta}))_{i,j} \triangleq -\int p(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} \tag{5.10}$$

$\mathsf{I}(\boldsymbol{\theta})$ is symmetric and positive semidefinite, and can also be written as (Amari and Nagaoka, 2000):

$$(\mathsf{I}(\boldsymbol{\theta}))_{i,j} = \int p(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_i} \log p(\mathbf{x}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_j} \log p(\mathbf{x}|\boldsymbol{\theta}) d\mathbf{x} = 4\int \frac{\partial}{\partial\theta_i}\sqrt{p(\mathbf{x}|\boldsymbol{\theta})} \frac{\partial}{\partial\theta_j}\sqrt{p(\mathbf{x}|\boldsymbol{\theta})} d\mathbf{x} \tag{5.11}$$

We can now define invariantized entropy $h_\infty(\mathbf{X}|\boldsymbol{\theta})$ using the determinant of the Fisher information matrix $|\mathsf{I}(\boldsymbol{\theta})|$ (Good, 1983):

$$h_\infty(\mathbf{X}|\boldsymbol{\theta}) \triangleq -\int p(\mathbf{x}|\boldsymbol{\theta}) \log_2 \frac{p(\mathbf{x}|\boldsymbol{\theta})}{\sqrt{|\mathsf{I}(\boldsymbol{\theta})|}} d\mathbf{x} \tag{5.12}$$

Invariantized entropy can too be negative or zero. Besides, Kullback-Leibler divergence $D(p\|q)$ is already invariant to such transformations.

Some definitions of entropy express it based on a particular 'prior' π (Caticha, 2000):

$$h_\pi(\mathbf{X}|p) \triangleq -\int p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{\pi(\mathbf{x})} d\mathbf{x} = -D(p\|\pi) \tag{5.13}$$

One should note that Caticha's 'prior' does not correspond to the Bayesian notion of a prior. Nonetheless, Good's invariantized entropy corresponds to Caticha's definition of entropy with the choice of Jeffreys' prior for 'prior' π. Jeffreys' prior is constant across $\mathbf{x}$ given a fixed value of the parameters.

5.2 Multivariate Normal Distribution

We can now study individual models of continuous attributes. The most familiar one is the multivariate normal distribution. The vector of attributes $\mathbf{X}$ is now treated as a single multi-dimensional attribute. If a d-dimensional attribute $\mathbf{X} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})\right\} \tag{5.14}$$

$I(A;B)$	0.001	0.01	0.05	0.1	0.2	0.3	0.4	0.5	1.0
ρ	0.037	0.117	0.259	0.360	0.492	0.583	0.652	0.707	0.866
R^2	0.1%	1.4%	6.7%	12.9%	24.2%	34.0%	42.6%	50.0%	75.0%
$I(A;B)$	1.5	2.0	2.5	3.0	3.5	4.0	4.5	5.0	5.5
ρ	0.935	0.968	0.984	0.992	0.996	0.998	0.999	1.000	1.000
R^2	85.5%	93.7%	96.9%	98.4%	99.2%	99.6%	99.8%	99.9%	100.0%

Table 5.1: A comparison of mutual information $I(A;B)$ for a bivariate normal model, measured with bits of information, correlation coefficients ρ, and percentages of explained variance $R^2 = \rho^2$.

If $d = 1$, the differential entropy can be expressed in closed form as (Cover and Thomas, 1991):

$$h(X|\mu,\sigma) = \frac{1}{2}\log_2(2\pi e\sigma^2) \tag{5.15}$$

And in the general case the differential entropy is (Billinger, 2004):

$$h(\mathbf{X}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{2}\log_2(|2\pi e\boldsymbol{\Sigma}|) \tag{5.16}$$

Here, $|\cdot|$ denotes the determinant.

A simple reference model p that allows correlation is a multivariate normal distribution. The d-dimensional attribute vector $\mathbf{X} = [X_1, \ldots, X_d]$ is treated as a single multi-dimensional attribute. On the other hand, the alternative model q models each attribute independently.

$$p: \qquad \mathbf{x} \sim \text{Normal}(\boldsymbol{\mu},\boldsymbol{\Sigma}) \tag{5.17}$$

$$q: \qquad \mathbf{x} \sim \prod_i^d \text{Normal}(\mu_i,\sigma_i) \tag{5.18}$$

This scheme is not limited to two dimensions, so correlations involving an arbitrary number of attributes can be investigated easily. For the case of $d = 2$, Fig. 5.1 demonstrates the relationship between the KL-divergence and the correlation coefficient ρ, which can also be expressed in closed form (Billinger, 2004):

$$D(p\|q) = -\frac{1}{2}\log_2(1-\rho^2) \tag{5.19}$$

It may be helpful to interpret the units of bits in terms of correlation coefficient values. For that reason, we provide Table 5.1. However, the table may be quite misleading for the interpretation of mutual information for discrete attributes.

In a more general situation, we may partition the attributes of $\mathbf{X}$ into two groups $\mathbf{X} = [\mathbf{A},\mathbf{B}]^T$. Here we assume that $\mathbf{A}$ is r-dimensional, and $\mathbf{B}$ s-dimensional. Similarly, we can partition the covariance matrix:

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{AA}} & \boldsymbol{\Sigma}_{\mathbf{AB}} \\ \boldsymbol{\Sigma}_{\mathbf{BA}} & \boldsymbol{\Sigma}_{\mathbf{BB}} \end{pmatrix} \tag{5.20}$$

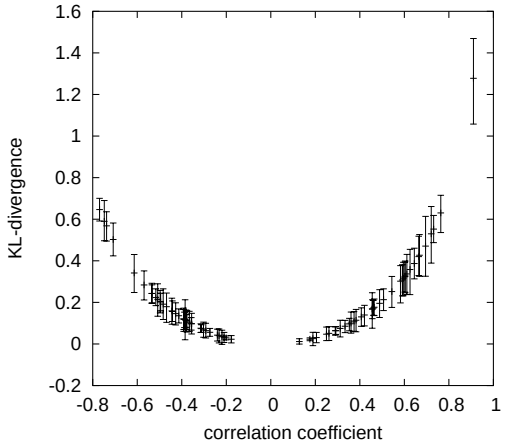


Figure 5.1: We may express correlation using KL-divergence instead of the correlation coefficient, retaining a monotonic relationship. In this figure we plot the importance of correlation for all pairs of attributes in the 'Boston housing' data. The confidence intervals are estimated with vagueness of loss. The wide confidence interval on the extreme right should raise suspicion: the high correlation for that particular pair of attributes (*property tax* and *highways*) is merely due to a few high property tax outliers. The next two correlations, (*nitric oxides* with *employment distance* and *non-retail acres*) are more meaningful and more stable.

We can now express the mutual information between $\mathbf{A}$ and $\mathbf{B}$ in closed form (Billinger, 2004):

$$D(p(\mathbf{A},\mathbf{B})\|p(\mathbf{A})p(\mathbf{B})) = I(\mathbf{A};\mathbf{B}) = -\frac{1}{2}\log_2\left(\frac{|\mathbf{\Sigma}|}{|\mathbf{\Sigma_{AA}}||\mathbf{\Sigma_{BB}}|}\right) \tag{5.21}$$

The resulting mutual information has a χ^2 asymptotic distribution (although better approximations have been proposed) (Billinger, 2004):

$$\frac{I(\mathbf{A};\mathbf{B})}{\log_2 e} \sim \chi^2_{rs} \tag{5.22}$$

Of course, we can employ this approach to obtain partial correlation coefficients, and many other interesting structures. Non-zero correlation corresponds to the existence of a 2-way interaction between the attributes. However, linear dependence is only a particularly simple type of interaction.

Interaction information using the multivariate normal model can also be positive. A possible interpretation is suppression (Lynn, 2003): an attribute B suppresses the variance in A unrelated to C, while A predicts the value of the label C.

5.3 Mixture Models

The multivariate Gaussian model is ubiquitous, but has certain limitations: it is strictly linear and strictly unimodal. This does not mean that it is inappropriate for non-linear and multimodal data. It is just that the model does not capture as much information as would be captured by a different model.

Mixture models have recently risen to prominence with their flexibility. Mixture models are based on a set of *components*, each component is a probability density function in the

attribute space. Each component has a corresponding probability of occurrence, and a point in the attribute space may have non-zero density for several components. If the set of components is finite, the model is a finite mixture (McLachlan and Peel, 2000).

We will focus on a restricted form of mixture models by making the assumption of *local independence*, so that the latent attribute Z having a range $\Re_Z = \{z_1, \ldots, z_K\}$ will account for all the dependence between attributes $\mathbf{X} = [X_1, X_2, \ldots, X_d]$:

$$p(\mathbf{X}|Z) = \sum_{k=1}^{K} \pi_k \prod_{i=1}^{d} p(X_i|\phi_{k,i}) \tag{5.23}$$

Each individual value of Z can be interpreted both as a *component*, a probabilistic prototype, a *cluster* a set of instances that correspond to the prototype (to some extent), or as an *axis* or dimension of a vector space where the instances can be represented as points. Since the value of Z is unknown, we infer X using a multinomial model for Z: $p(z_k) = \pi_k$, $\sum_k \pi_k = 1$. The naïve Bayesian classifier (NBC) of the label Y given the attributes $\mathbf{X}$ is identical to the above formulation of (5.23), but with the non-hidden label Y playing the role of the latent attribute. An added benefit of using local independence is that for computing the marginalizations of $\mathbf{X}$, all that needs to be done is to compute the product for a subset of attributes.

The choice of the functions in the mixture depends on the type of the attribute. Most implementations are based on normal or Gaussian mixtures, which work for continuous attributes, e.g. (Roberts et al., 1999). Recently, multinomial mixtures for discrete or count-based attributes have been successfully utilized in information retrieval, e.g. (Buntine, 2002). Most implementations are based on normal or Gaussian mixtures, which work for continuous attributes, e.g. (Roberts et al., 1999). Recently, multinomial mixtures for discrete or count-based attributes have been successfully utilized in information retrieval, e.g. (Buntine, 2002). The MULTIMIX program (Hunt and Jorgensen, 1999) handles both continuous and discrete attributes simultaneously with the local independence assumption, adopting the multinomial distribution for any discrete attribute X_d (5.24) and the normal distribution for any continuous attribute X_c (5.25):

$$X_d \sim \text{Multinomial}(\boldsymbol{\lambda}, 1) \qquad p(X_d = x_j|\boldsymbol{\lambda}) = \lambda_j, \ \sum_j \lambda_j = 1 \tag{5.24}$$

$$X_c \sim \text{Normal}(\mu, \sigma) \qquad p(X_c = x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{5.25}$$

Finally, mixture models are not the only choice of probability models involving continuous attributes. An active area of research in statistics, finance and econometrics involves the concept of *copulas* (Joe, 1997, Cherubini et al., 2004).

5.3.1 The EM Algorithm

We have employed the expectation-maximization algorithm to determine the parameters π and ϕ in (5.23). The EM algorithm is an iterative procedure for improving the fit of the model by interleaving two separate optimization steps. In the *expectation* step we compute the latent attribute value probabilities for each instance of the data, while keeping π and ϕ constant. In the *maximization* step, we compute the maximum likelihood

(and therefore also minimum sample entropy) parameter values for each component, given the set of instances having the latent attribute value which corresponds to the component: each instance is weighted with the probability that it belongs to the component. Because the distributions we use in the mixture are simple, the maximum likelihood equations can be solved analytically (Hunt and Jorgensen, 1999).

Instead of the common practice of using random values as initial parameter settings, each instance was assigned crisply to one of the clusters as found by *pam*, a robust greedy medoid-based clustering algorithm (Kaufman and Rousseeuw, 1990), instead of the first E step. To prevent correlated attributes from skewing the metric, the instances were presented to *pam* projected to their eigenspace using principal component analysis (PCA). Since PCA is sensitive to the scale of attributes, each attribute was standardized to have the mean of 0 and the variance of 1 beforehand.

The role of the expectation step in iteration $[t+1]$ is the computation of the probability distribution of the latent attribute values in each instance $\mathbf{x}^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_d^{(i)}]$ of the data, and for k-th of the K components, while keeping π and ϕ constant:

$$p(z_k|\mathbf{x}^{(i)}) = \tau^{(i),[t+1]} = \frac{\pi_k^{[t]} \prod_{m=1}^{d} p(x_m^{(i)}|\phi_{k,m}^{[t]})}{\sum_{k=1}^{K} \pi_k^{[t]} \prod_{m=1}^{d} p(x_m^{(i)}|\phi_{k,m}^{[t]})} \tag{5.26}$$

Since we allow the full range of probabilities, this allows each instance to be a possible member of several components. In the maximization step, we compute the maximum likelihood (and therefore also minimum sample entropy) parameter values for each component, weighing each instance according to its membership in the component. Because the model is simple, the maximum likelihood equations can be solved analytically, averaging over all the instances of the data $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\}$:

$$\pi^{[t+1]} = \frac{1}{n} \sum_{i=1}^{n} \tau^{(i),[t+1]} \tag{5.27}$$

$$\lambda_{k,m,j}^{(t+1)} = \frac{1}{n\pi^{[t+1]}} \sum_{i, x_m^{(i)}=j} \tau^{(i),[t+1]} \tag{5.28}$$

$$\mu_{k,m}^{[t+1]} = \frac{1}{n\pi^{[t+1]}} \sum_{i} \tau^{(i),[t+1]} x_m^{(i)} \tag{5.29}$$

$$\sigma^{2}{}_{k,m}^{[t+1]} = \frac{1}{n\pi^{[t+1]}} \sum_{i} \tau^{(i),[t+1]} (x_m^{(i)} - \mu_{k,m}^{[t+1]})^2 \tag{5.30}$$

If a d-dimensional attribute $\mathbf{X} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, the two corresponding M steps for the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\boldsymbol{\Sigma}$ are:

$$\boldsymbol{\mu}_{k,m}^{[t+1]} = \frac{1}{n\pi^{[t+1]}} \sum_{i} \tau^{(i),[t+1]} \mathbf{x}_m^{(i)} \tag{5.31}$$

$$\boldsymbol{\Sigma}_{k,m}^{[t+1]} = \frac{1}{n\pi^{[t+1]}} \sum_{i} \tau^{(i),[t+1]} (\mathbf{x}_m^{(i)} - \boldsymbol{\mu}_{k,m}^{[t+1]}) \cdot (\mathbf{x}_m^{(i)} - \boldsymbol{\mu}_{k,m}^{[t+1]})^T \tag{5.32}$$

Using the multivariate normal distribution for $\mathbf{X}$ is no longer consistent with the local independence model (5.23): $\mathbf{X}$ is now treated as a single multi-dimensional attribute.

5.3.2 Supervised, unsupervised and informative learning

The primary target in classification is predicting the label. The *unsupervised* learning approach of maximizing the joint likelihood or entropy, is not directly concerned with this aim, using the following criterion:

$$\arg\min_p \hat{h}(\mathbf{X}, Y|\mathcal{D}, p) \tag{5.33}$$

Instead, a separate mixture model can be built for each label value. This is referred to as *informative* learning (Rubinstein and Hastie, 1997), and the objective is to minimize the entropy of the attributes given the label:

$$\arg\min_p \hat{h}(\mathbf{X}|Y, \mathcal{D}, p) \tag{5.34}$$

Class-conditional modelling is a type of informative learning, where we fit k components for each label. This technique has been used for improving the naïve Bayesian classifier by several researchers (Vilalta and Rish, 2003, Monti and Cooper, 1999).

Class-conditional modelling of attributes is not, however, discriminative modelling of class boundaries, the true goal of pure supervised learning. In our context we can formalize the objective of non-Bayesian supervised learning as minimization of the entropy of the predictive distribution of the label given the attributes:

$$\arg\min_p \hat{h}(Y|\mathbf{X}, \mathcal{T}, p) \tag{5.35}$$

This way we are fulfilling the fundamental task of minimizing the loss in predicting the label from the attributes. It is important to distinguish these three learning objectives, as they all differ one from another, in spite of the apparent similarity between supervised and informative learning. Fig. 5.2 illustrates that neither unsupervised nor informative learning match the supervised learning objectives, and that informative learning is not necessarily better than unsupervised learning. Fig. 5.3 compares certain supervised and unsupervised discretization methods with mixture models.

In the context of MaxEnt approaches, we seek the worst model given the constraints. The basic MaxEnt ($\arg\max_p H(\mathbf{X}, Y|p)$, (Jaynes, 2003)) can be used for unsupervised learning, and either maximum conditional entropy for discriminative ($\arg\max_p H(Y|\mathbf{X}, p)$) and class-conditional ($\arg\max_p H(\mathbf{X}|Y, p)$) learning. An alternative for the non-joint case is the minimum mutual information (MinMI) principle ($\arg\min_p I(\mathbf{X}; Y|p)$, (Globerson and Tishby, 2004)).

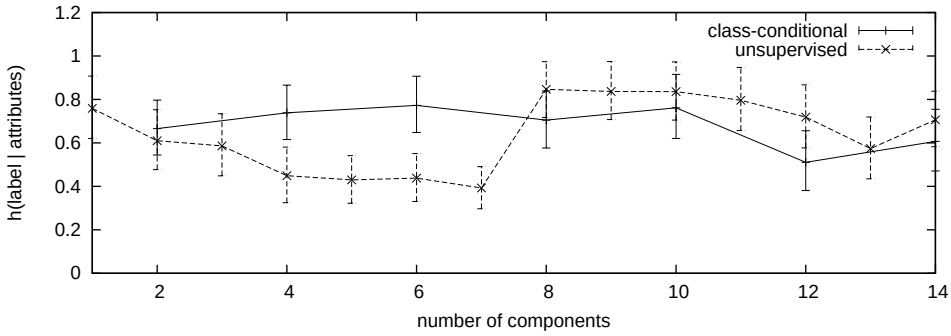


Figure 5.2: This analysis of the 'voting' data set demonstrates that increasing the number of components does not always result in better classification performance on the training set, regardless of whether an unsupervised or a class-conditional scheme is used. Furthermore, an unsupervised scheme may well yield better results than a class-conditional one, with smaller confidence intervals. The reason for this is that the maximum likelihood EM algorithm is not seeking to maximize the conditional likelihood of the label given the attributes, the goal of pure supervised learning.

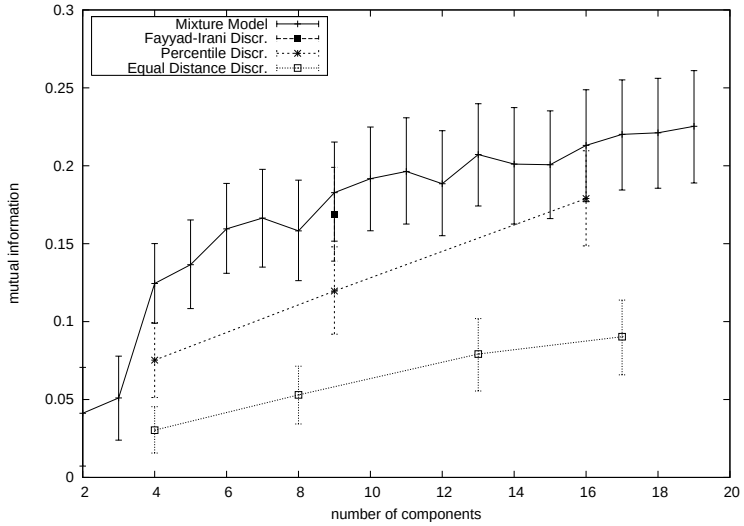


Figure 5.3: In the 'CMC' data set, we assessed the mutual information between the trichotomous label and the pair (*wife age*, *number of children*), varying the number of components and the method used to construct them. The alternative to mixture modelling is space partitioning, as it appears in classification trees and in discretization methods, but which yields mutually-exclusive partitions as components. It is possible to see that the Fayyad and Irani (1993) discretization method, which tries to capture the dependence between an attribute and the label, is competitive with mixture modelling at the same level of complexity, but not the discretization methods based on equally large or equally represented bins. Again, mutual information is not monotonically increasing with the number of components trained with an unsupervised criterion.

CHAPTER 6

Visualization with Interactions

6.1 The Methodology of Interaction Analysis

Before beginning with interaction analysis, we first need the underlying probabilistic model. The ways of coming up with the model have been discussed in Ch. 4. In practical analysis, however, it is not necessary to have a single global model: through local analysis we build a separate model for each subset of attributes under investigation, as the global model would match the local one if the attributes outside the focus were marginalized away. Namely, marginalization can be performed both on the data or on the model. The disadvantage of the local analysis is that the local models, especially if they are estimated and not modelled, might be mutually inconsistent.

The general procedure for interaction analysis in an unsupervised learning problem described with attributes $\mathcal{X} = \{A_1, A_2, \ldots, A_m\}$ thus takes the following form:

1. Form the set of one-attribute set of projections $\mathcal{S}_1 = \{\{A_1\}, \{A_2\}, \ldots, \{A_m\}\}$, two-attribute set of projections $\mathcal{S}_2 = \{\{A_1, A_2\}, \{A_1, A_3\}, \ldots, \{A_{m-1}, A_m\}\}$, and so on.

2. If there is a label Y, add it to each projection $S \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots$.

3. Build a local probability model for each S.

4. Evaluate interaction information for each S.

5. Process the results:

   - Summarize the pattern of interactions in a comprehensible form (interaction matrix, interaction dendrogram, metric scaling, Sect. 6.2).
   - Identify the most distinct interactions and visualize them (interaction graphs, Sect. 6.3).
   - Examine a distinct interaction in detail, potentially explaining it with rules (Sect. 6.4).

For tractability, we only perform interaction analysis up to a particular k. Usually $k = 2$. The complexity in such a case is quadratic, as there are $\binom{m}{k}$ ways of choosing k attributes out of m. We can label several attributes to set the context for interaction analysis, maintaining the combinatorial tractability. We never attempt to build the global probability model of $P(A_1, A_2, \ldots, A_m)$.

6.2 Interaction as Proximity

In initial phases of exploratory data analysis we might not be interested in detailed relationships between attributes, but merely wish to discover groups of mutually interacting attributes. In supervised learning, we are not investigating the relationships between attributes themselves (where mutual information would have been the metric of interest), but rather the relationships between the mutual information of either attribute with the label. In other words, we would like to know whether two attributes provide similar information about the label, or whether there is synergy between attributes' information about the label.

6.2.1 Attribute Proximity Measures

To perform any kind of similarity-based analysis, we should define a similarity or a dissimilarity measure between attributes. With respect to the amount of interaction, interacting attributes should appear close to one another, and non-interacting attributes far from one another. One of the most frequently used similarity measures for clustering is *Jaccard's coefficient*. For two sets, $\mathcal{A}$ and $\mathcal{B}$, the Jaccard's coefficient (along with several other similarity measures) can be expressed through set cardinality (Manning and Schütze, 1999):

$$J(\mathcal{A}, \mathcal{B}) \triangleq \frac{|\mathcal{A} \cap \mathcal{B}|}{|\mathcal{A} \cup \mathcal{B}|} \tag{6.1}$$

If we understand interaction information as a way of measuring the cardinality of the intersection in Fig. 3.1 and in Section 3.2.4, where mutual information corresponds to the intersection, and joint entropy as the union, we can define the *normed mutual information* between attributes A and B:

$$\bar{I}(A; B) \triangleq \frac{I(A; B)}{H(A, B)} \tag{6.2}$$

Rajski's distance is closely related to normed mutual information, and can be defined as

$$\langle A, B \rangle_R = 1 - \bar{I}(A; B) = \frac{2H(A, B) - H(A) - H(B)}{H(A, B)} = \frac{H(A|B) + H(B|A)}{H(A, B)} \tag{6.3}$$

It takes the value of 1 when the attributes are completely dependent, and 0 when the attributes are completely independent. Furthermore, it obeys the triangle inequality (Rajski, 1961). It is highly applicable for various visualizations.

Màntaras' distance (López de Màntaras, 1991) is identical to Rajski's distance, has been shown to be a useful heuristic for feature selection, less sensitive to the attribute alphabet size. Dividing by the joint entropy helps us reduce the effect of the number of attribute values, hence facilitating comparisons of mutual information between different attributes. Normed mutual information is identical to interdependence redundancy (Wong

and Liu, 1975). Rajski's distance was recently generalized in the terms of Kolmogorov complexity by Li et al. (2004). Meilă (2003) has employed the variation of information, the non-normalized Rajski's distance, to assess clusterings. She has also proved a number of interesting properties of variation of information that also hold for Rajski's distance, such as collinearity of attribute refinements and convex additivity.

For visualizing higher-order interactions in an analogous way, we can introduce a further attribute C either as context, using normed conditional mutual information:

$$\bar{I}(A, B|C) \triangleq \frac{I(A; B|C)}{H(A, B|C)} \tag{6.4}$$

Alternatively, we may employ interaction information in a normalized form. We refer to it as normed interaction magnitude:

$$|\bar{I}(A; B; C)| \triangleq \frac{|I(A; B; C)|}{H(A, B, C)} \tag{6.5}$$

Here, the interaction magnitude $|I(A; B; C)|$ is the absolute value of interaction information. C has to be fixed and usually corresponds to the label, while A and B are attributes that iterate across all combinations of remaining attributes. We define a distance as $\langle A, B, C\rangle_R = 1 - |\bar{I}(A; B; C)|$ in order to summarize interaction information across a set of attributes.

While the distance $\langle A, B, C\rangle_R$ functions as an approach to summarizing interaction information, it does not have particularly convenient metric properties. For that purpose we will define two multimetrics: interaction distance (6.6) and total correlation distance (6.7):

$$\langle \mathcal{X}\rangle_{ID} \triangleq \frac{\sum_{X\in\mathcal{X}} H(X|\mathcal{X} \setminus \{X\})}{H(\mathcal{X})} \tag{6.6}$$

$$\langle \mathcal{X}\rangle_{TD} \triangleq 1 - \frac{H(\mathcal{X})}{\sum_{X\in\mathcal{X}} H(X)} \tag{6.7}$$

Both of these distances are in the range of $[0, 1]$. The interaction distance corresponds to the proportion of joint entropy that remains unaccounted for after examining all interactions. The total correlation distance is related to the proportion between the actual joint entropy and the sum of individual attributes' entropies.

6.2.2 The Interaction Matrix

The Library of Congress in Washington maintains the THOMAS database of legislative information. One type of data are the senate roll calls. For each roll call, the database provides a list of votes cast by each of the 100 senators. There were 459 roll calls in the first session of the 108th congress, comprising the year 2003. For each of those, the vote of every senator is recorded in three ways: 'Yea', 'Nay' and 'Not Voting'. The outcome of the roll call is treated in precisely the same way as the vote of a senator, with positive outcomes (Bill Passed, Amendment Germane, Motion Agreed to, Nomination Confirmed, Guilty, etc.) corresponding to 'Yea', and negative outcomes (Resolution Rejected, Motion to Table Failed, Veto Sustained, Joint Resolution Defeated, etc.). Hence, the outcome can

be interpreted as the 101st senator. Each senator and the outcome can be interpreted as binary attributes. Each roll call can be interpreted as an instance.

Special-purpose models are normally used in political science, and they often postulate a model of rational decision making. Each senator is modelled as a position or an *ideal point* in a spatial model of preferences (Davis et al., 1970, Stewart III, 2001), where the first dimension often delineates the liberal-conservative preference, and the second region or social issues preference (McCarty et al., 2001). In the corresponding voting model senators try to maximize their utility, and the voting process is interpreted as the attempt of each senator to decide about the roll call based on his or her ideal point. In this model, it is the similarity in ideal points that accounts for the similarities between senators' votes. The algorithms for fitting the spatial models of parliamentary voting can be understood as constructive induction algorithms that try to explain the actions of each senator over all the roll calls simply with the senator's ideal point. These models can be evaluated by comparing the true votes with the votes predicted by the model. The ideal points can be obtained either by optimization, e.g., with the optimal classification algorithm (Poole, 2000), or by Bayesian modelling (Clinton et al., 2004). Of course, not all analysis methods postulate a model of decision making, e.g. (Lawson et al., 2003, de Leeuw, 2003).

Distances as plain numbers provide little insight. However, we can provide the distances between all pairs of senators in the form of a graphical matrix (Fig. 6.1). The color can be used to indicate the proximity: the darker, the higher the dependence between the two senators. Dissimilarity matrices are clearer if similar senators are adjacent to one another, so we have sorted them. Furthermore, we can exploit the symmetry of Rajski's distance by only visualizing the bottom-left half of the dissimilarity matrix. It is important to realize that high Rajski's distance may imply that the votes are systematically opposed: the low Rajski's distance between Kerry (D-MA) and the Republican senators is an example of this.

6.2.3 Interaction Matrix with the Label

All interactions in supervised learning domains include the label. The interaction matrix thereby describes the relationship between each pair of attributes with regards to the way they predict the label together. Along the diagonal, we may visualize the information gain of the attribute alone. We have employed a slightly different diverging color scheme from cartography (Brewer et al., 2003): as before, red color encodes positive, and blue color negative interaction information, but no-interaction is simply white. An example of such an interaction matrix is shown in Fig. 6.2.

6.2.4 Metric Scaling

If each senator is denoted with a point in some k-dimensional space, we can try to place these points so that the Euclidean distances between the points would match Rajski's distances. Most algorithms for metric scaling are based on iterative procedures. We have employed Torgerson-Gower scaling (Borg and Groenen, 1997), and SMACOF (de Leeuw, 1977). The Torgerson-Gower algorithm employs the scalar product algorithm and a single step of singular value decomposition. On the other hand, SMACOF is an iterative majorization algorithm which optimizes a simpler auxiliary function that bounds the true criterion of matching metric distance. In Fig. 6.3 we can see that both methods separate

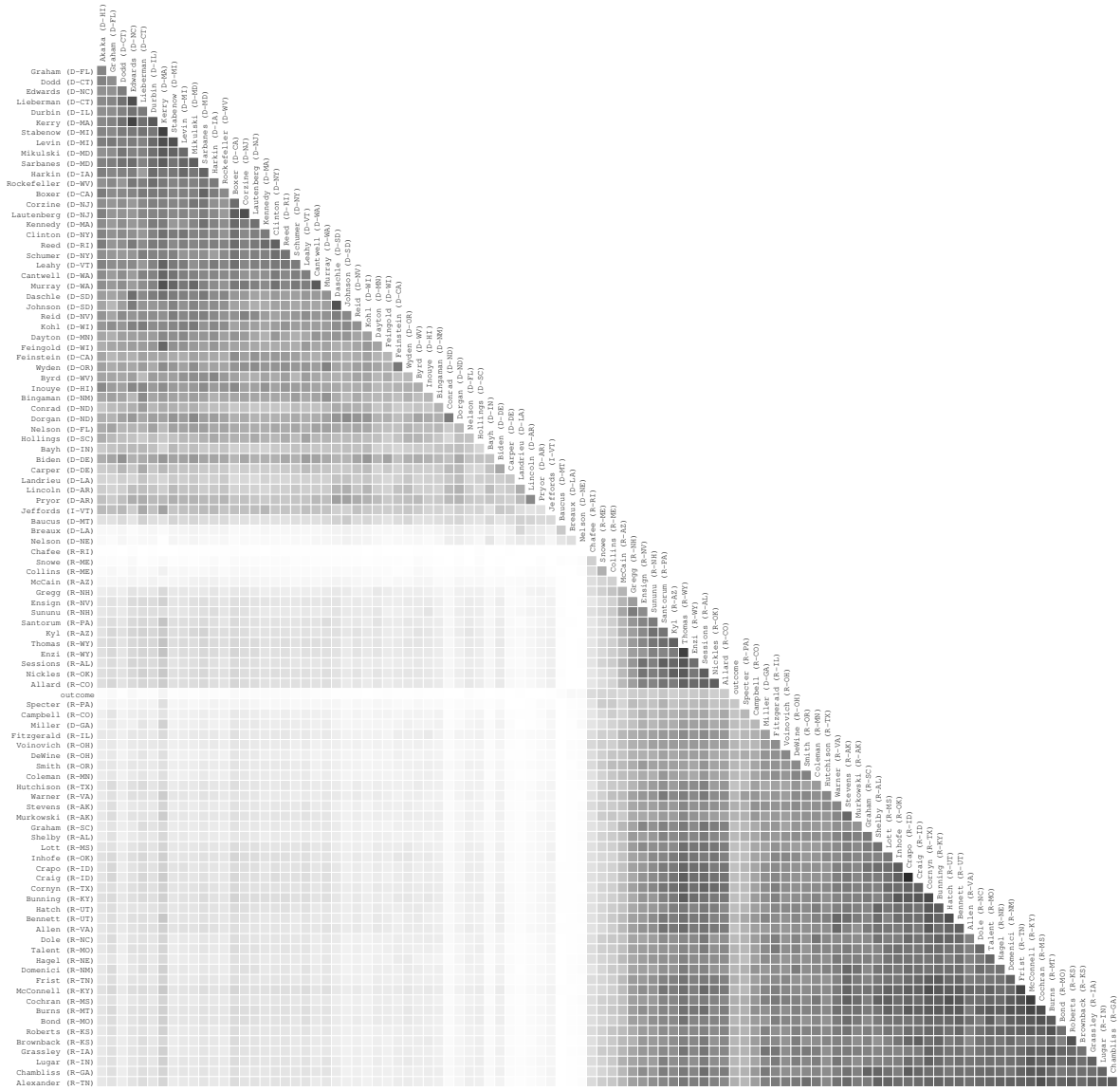


Figure 6.1: The symmetric dissimilarity matrix graphically illustrates Rajski's distance between all pairs of senators, based on their votes in 2003. Three large clusters can be identified visually from this graph, and one group of moderate senators in each party. The major clusters correspond to the political parties even if the party information was not used in the computation of distance.

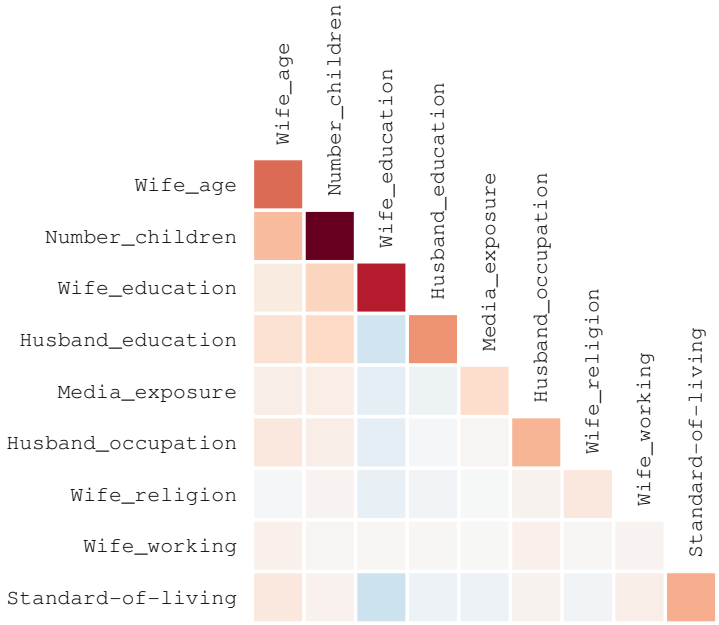


Figure 6.2: From this depiction of the 'CMC' data set, we can interpret the importance of individual attributes to predicting the label (*contraception method*), and all attribute pairs. From the diagonal, we can identify that the two most important attributes are age and education of the wife. There are two distinct negative interactions involving the education of the wife: with the standard of living and the husband's education. The most distinctly positive is the interaction between the age and the number of children.

the Republicans from the Democrats, with a single outlier (Miller, D-GA).

6.2.5 Interaction Dendrograms

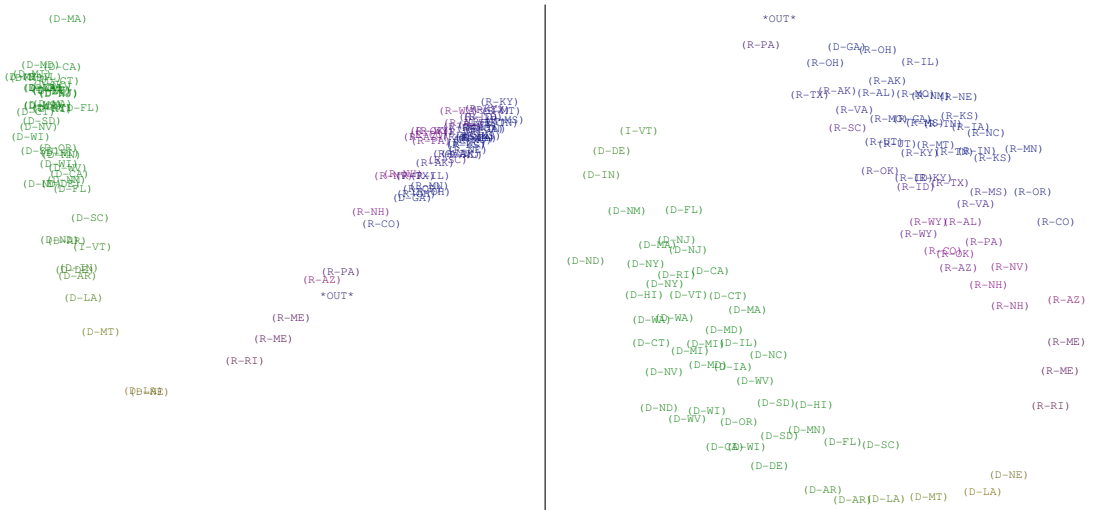
Hierarchical Clustering Algorithms

The agglomerative nesting algorithm *agnes* (Kaufman and Rousseeuw, 1990, Struyf et al., 1997) constructs a hierarchy of clusterings. At first, each observation is a small cluster by itself. Clusters are merged until only a single large cluster remains which contains all the observations. At each stage the two nearest clusters are combined to form one larger cluster. In hierarchical clustering there are $n-1$ fusion steps for n observations. The decision, which two clusters are the closest, is made with a *linkage method*:

- The average linkage method attempts to minimize the average distance between all pairs of members of two clusters. If $\mathcal{R}$ and $\mathcal{Q}$ are clusters, the distance between them is defined as

$$d(\mathcal{R},\mathcal{Q}) = \frac{1}{|\mathcal{R}||\mathcal{Q}|}\sum_{i\in\mathcal{R},j\in\mathcal{Q}} d(i,j) \tag{6.8}$$

- The single linkage method is based on minimizing the distance between the closest neighbors in the two clusters. In this case, the hierarchical clustering corresponds



Torgerson's scaling procedure SMACOF scaling procedure

Figure 6.3: Multi-dimensional scaling attempts to capture the given dissimilarity matrix with Euclidean distances between points. The outcome depends highly on the algorithm used, e.g., Torgerson's (left) and SMACOF (right). Regardless of that, the results are comparable.

to the minimum spanning tree:

$$d(\mathcal{R}, \mathcal{Q}) = \min_{i \in \mathcal{R}, j \in \mathcal{Q}} d(i, j) \tag{6.9}$$

- The complete linkage method is based on minimizing the distance between the furthest neighbors:

$$d(\mathcal{R}, \mathcal{Q}) = \max_{i \in \mathcal{R}, j \in \mathcal{Q}} d(i, j) \tag{6.10}$$

- Ward's minimum variance linkage method attempts to minimize the increase in the total sum of squared deviations from the mean of a cluster.
- The weighted linkage method is a derivative of average linkage method, but both clusters are weighted equally in order to remove the influence of different cluster size.

The linkage distance between $\mathcal{R}$ and $\mathcal{Q}$ indicates the quality of the merger. The 'height' in the graphical presentation of the cluster corresponds to $d(\mathcal{R}, \mathcal{Q})$.

Unsupervised Attribute Clustering

We can cluster the attributes by how much they are associated with one another, simply using the Rajski's distance between them. The label is just one of the attributes, treated equally. This is useful for obtaining a broad overview of the data set, as shown in Fig. 6.4.

Supervised Attribute Clustering

In the example of the 'Spam' data set, the label got assigned to a specific cluster. This may give a misleading indication that other attributes are irrelevant. This is not the

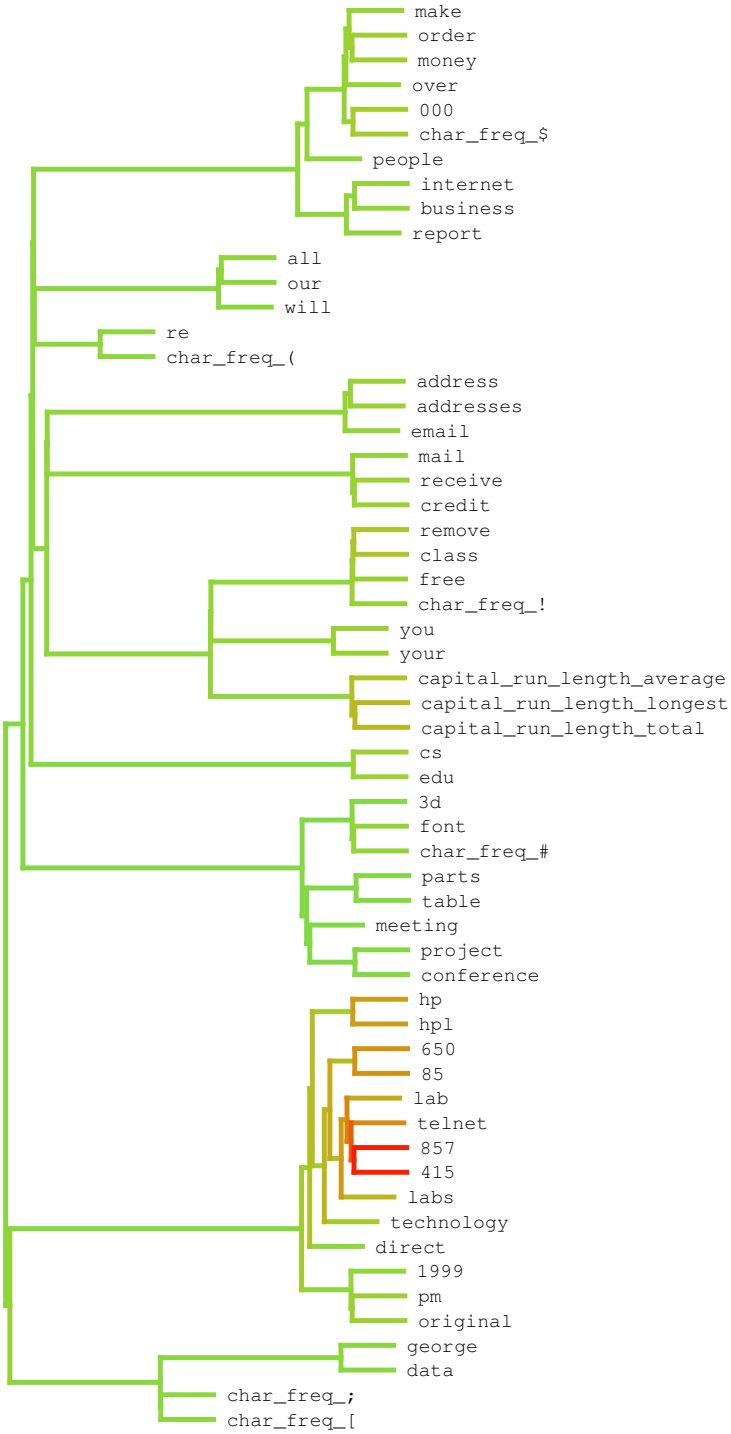


Figure 6.4: This clustering of the 'Spam' data set illustrates the dependencies between co-appearances of individual words in email messages. The color indicates the strength of the dependence: green is weak, and red is strong. Several clusters of keywords identify typical topics. The label is *class*, and it appears in a cluster along with *remove*, *free* and the exclamation mark.

case. In most classification and regression tasks we are not particularly interested in the relationships between attributes themselves. Instead, we are interested in how the attributes provide the information about the label.

Interaction information can be seen as the change in dependence between pairs of attributes after introducing the context. The direction of change is not important: we will distinguish this later. If we bind proximity in our presentation to the amount of change in dependence, regardless of whether it is positive or negative, we can employ the aforementioned generalization of Rajski's distance $\langle A, B, C\rangle_R = 1 - |\bar{I}(A; B; C)|$. The attributes that co-interact with the label will hence appear close to one another; those that do not co-interact with the label will appear far from one another.

In the example in Fig. 6.5, we have used Ward's method for agglomerative hierarchical clustering. We include each attribute's individual information gain in the form of a horizontal bar: the longer the bar, the more informative the attribute. Clustering helps identify the groups of attributes that should be investigated more closely. We can use color to convey the type of the interaction. We color the branches of the dendrogram based on the average interaction information inside the cluster, again mapping to the blue-green-red scale. For example, we color zero interactions green, positive interactions red and negative interactions blue, mixing all three color components depending on the normed interaction information. Blue clusters indicate on average negatively interacting groups, and red clusters indicate positively interacting groups of attributes.

The resulting interaction dendrogram is one approach to variable clustering, where the proximity is based on the redundancy or synergy of the attributes' information about the label. We can observe that there are two distinct clusters of attributes. One cluster contains attributes related to the lifestyle of the person: *age*, *family*, *working hours*, *sex*. The second cluster contains attributes related to the occupation and education of the person. The third cluster is not compact, and contains the information about the native country, race and work class, all relatively uninformative about the label.

6.2.6 Attribute Selection and Interaction Dendrograms

Jakulin and Leban (2003) have proposed a graphical representation of the interaction information, making use of the Venn diagram analogy, but expressing the sets as overlapping bars and not as overlapping circles. Namely, we can decompose the joint information gain $I(A, B; Y)$ into a sum of $I(A; B; Y) + I(A; Y) + I(B; Y)$. In a dendrogram the horizontal bars already indicate the mutual information of each attribute $I(\cdot; Y)$. Furthermore, interacting attributes are already adjacent in the diagram. For every pair of adjacent attributes in the dendrogram, we can express the $I(A; B; Y)$ as an arrow, directed towards left and colored blue if the interaction is negative, and towards right and colored red if the interaction is positive. The result of such visualization for the 'mushroom' data set is shown in Fig. 6.6.

6.2.7 Taxonomies and Interaction Dendrograms

In a previous study (Zupan et al., 2001) the participating physician defined an attribute taxonomy for this domain in order to construct a required concept hierarchy for the decision support model: this provided grounds for comparison with the taxonomy discovered by observing attribute interactions from the data. In Fig. 6.7, we compare the attribute

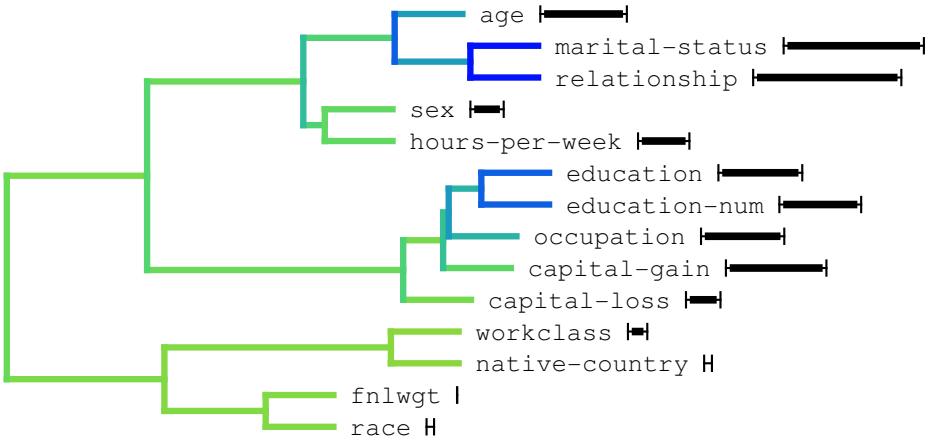


Figure 6.5: An interaction dendrogram illustrates which attributes interact, positively or negatively, with the label in the 'census/adult' data set. The label indicates the individual's income. The width of the horizontal bar indicates the amount of mutual information between an attribute and the label.

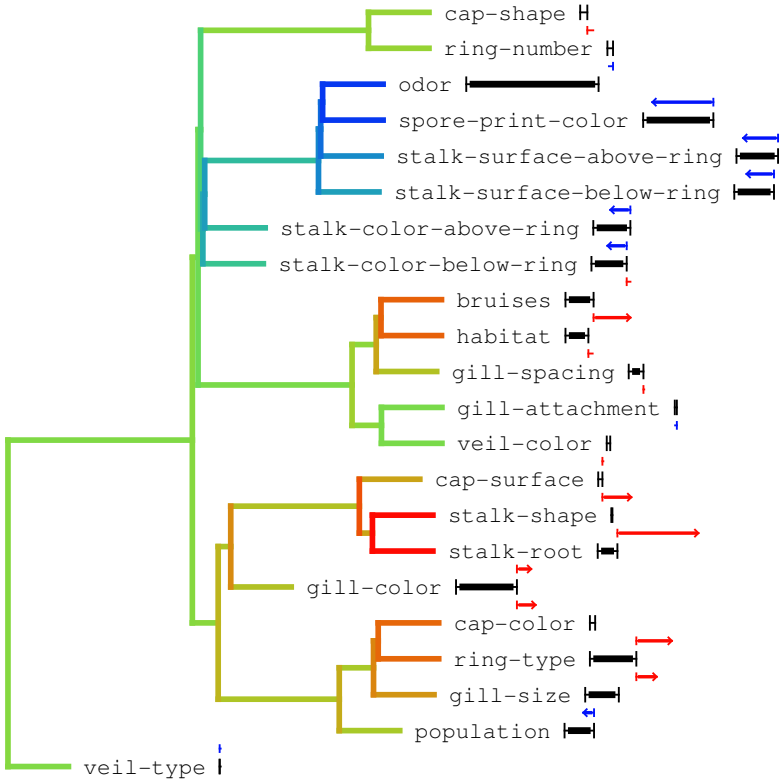


Figure 6.6: An interaction dendrogram for the 'mushroom' data set helps us perform rudimentary attribute selection and combination. The label is a mushroom's edibility. We observe that *spore-print-color* is next to being useless once *odor* has been taken into consideration. On the other hand, a holistic treatment of *bruises* and *habitat* would result in a synergy that is itself worth as much as *bruises* on its own.

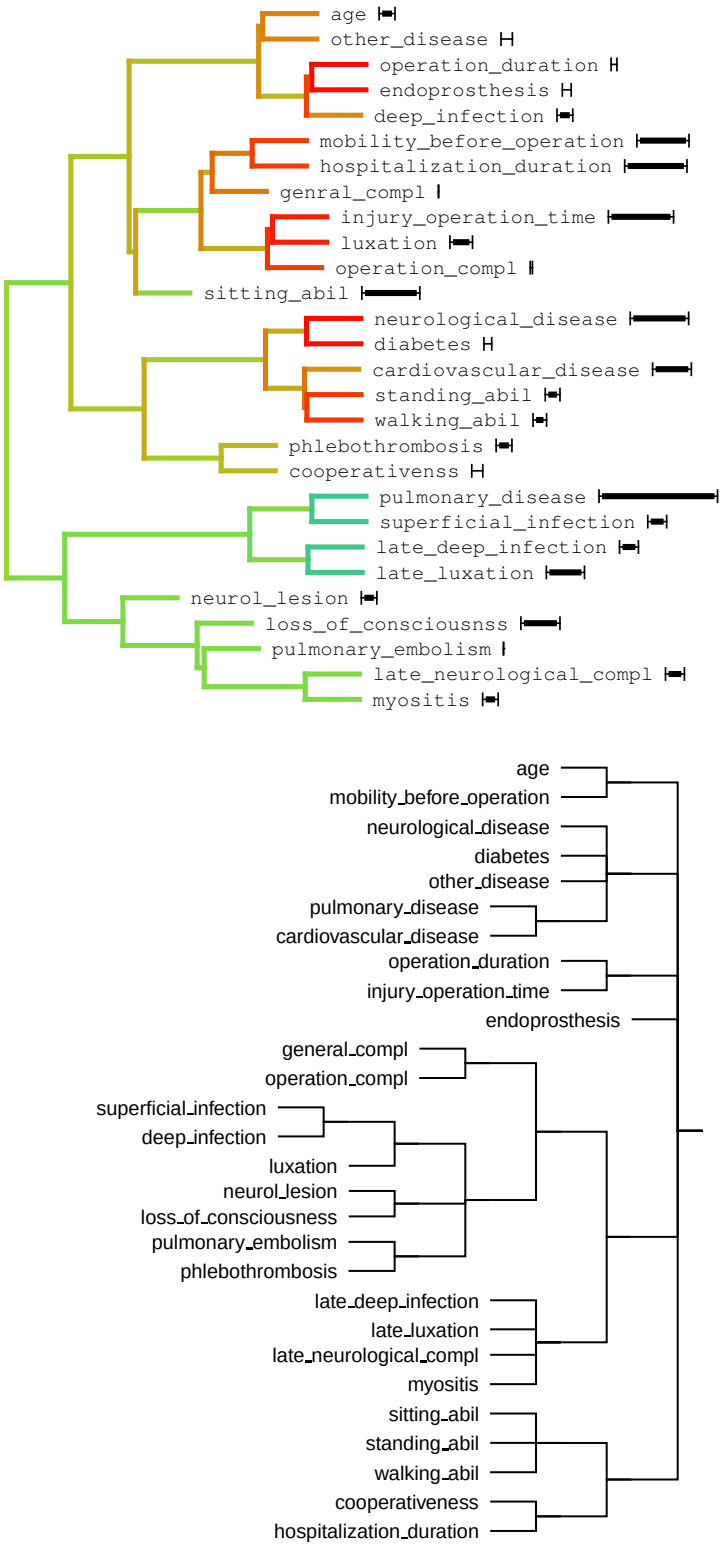


Figure 6.7: An attribute interaction dendrogram (top) illustrates which attributes interact, positively or negatively, while the expert-defined concept structure (bottom) was reproduced from (Zupan et al., 2001).

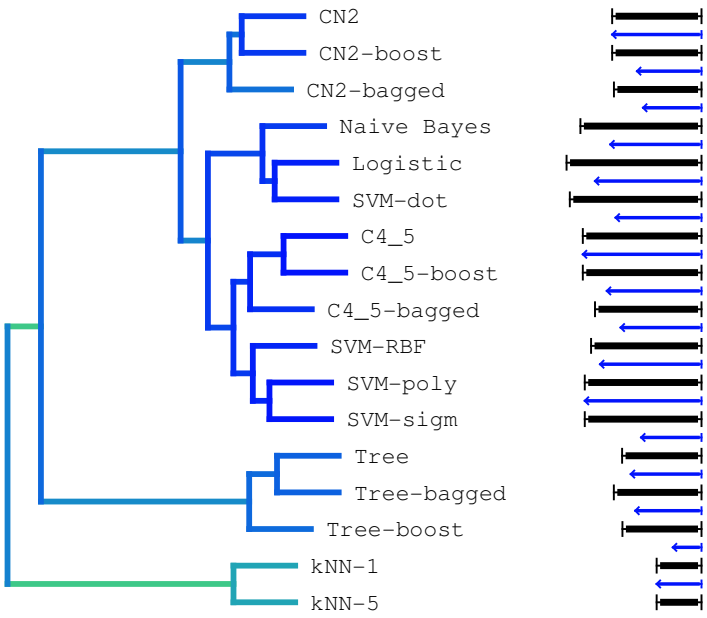


Figure 6.8: The output of a classifier can be understood as an attribute. This way, we can examine the similarities between machine learning algorithms. The ones shown are implemented in Orange (Demšar and Zupan, 2004). They were evaluated on the 'CMC' data set under the 10-fold cross-validation protocol. Logistic regression performed best, but the differences were small.

interaction dendrogram with an expert-defined concept structure (attribute taxonomy). While there are some similarities (like the close relation between the abilities to stand and to walk), the two hierarchies differ. The domain expert appears to have defined her structure on the basis of medical (anatomical, physiological) taxonomy; they do not seem to correspond to attribute interactions.

6.2.8 A Taxonomy of Machine Learning Algorithms

Assume k classification algorithms, $\mathcal{L} = \{L_1, L_2, \ldots, L_k\}$. Each of the classification algorithms is trained on the data set, and then evaluated on the test set. The predictions of the classifier are labelled attribute values. The output of each classifier can thus be understood as an attribute, and we can apply the interaction dendrogram to organize the results, as shown in Fig. 6.8. We can identify several clusters of methods; the first cluster consists of linear methods (naïve Bayesian classifier, logistic regression, SVM with a dot product kernel), the second cluster contains C4.5-based methods, the third cluster is composed of SVM with various non-linear kernels, and the final cluster comprises nearest neighbor algorithms. The CN2 classifier and an alternative implementation of classification trees are different. These taxonomies generally depend on the data set.

6.2.9 Missing Values

Attribute values are sometimes not given. The missing value is usually represented with a special attribute value if the attribute is discrete, or we only infer the model from those

instances that have no missing values. We have employed both techniques in this work.

The concept of *ignorability* (Gelman et al., 2004a) relates to whether the missing values are truly missing at random or whether there is some pattern to whether the value is missing or not. We can study the pattern of missing values with interaction analysis, however. To do this, we define the binary *inclusion attribute* I_A for an attribute A (Gelman et al., 2004a), which is defined as follows:

$$I_A \triangleq \begin{cases} 0 & ; \text{ The value of } A \text{ is missing;} \\ 1 & ; \text{ The value of } A \text{ is observed.} \end{cases} \tag{6.11}$$

Of course, we could define the inclusion attribute for several underlying attributes, e.g. I_{ABC} that would take the value of 1 when all of them would be known and 0 otherwise.

These inclusion attributes can be added to the data set, and the interaction analysis can be performed as usual. This way, we can infer whether a particular attribute is not missing at random. For experiments, we have focused on the 'pima' data set from the UCI repository (Hettich and Bay, 1999). The 'pima' data set is described as having no missing values. Yet, if we examine the histograms for attributes, as shown in Fig. 6.9, we see that there is a considerable number of anomalous values. They are often coded as 0 or as 99, because experimenters frequently impute illogical values when the true value is missing: body mass and blood pressure of 0 are obviously impossible.

For the analysis, we have included the inclusion attributes for the following attributes: *body mass*, *blood pressure*, *insulin level* and *skin fold*. From Fig. 6.10 we can see that the inclusion of *insulin level* is not importantly associated with the labelled attribute, but there are stronger associations with other inclusion attributes. From the dendrogram we can understand how the observations were recorded.

6.2.10 Concept Drift

The idea of concept drift (Kukar, 2003, Widmer and Kubat, 1996) is that the characteristics of certain attributes change with time. A practical example of drift is in medicine, as different doctors may judge different characteristics of disease differently. Moreover, the experience derived from the data gathered over a period will reflect in the actions done in the subsequent periods.

We have taken the familiar HHS data set, but extended with more recent observations (Jenul, 2003). The attribute *period* indicates whether an instance came from the first batch of observations or from the second one. It is then interesting to examine what has changed. For this, we can employ unsupervised attribute clustering. The results in Fig. 6.11 show that the properties of several attributes have changed.

6.3 Interaction Graphs

The analysis described in the previous section was limited to rendering the magnitude of interaction gains between attributes. Many interesting relationships are not visible in detail in the dendrogram. An interaction graph presents the interactions among a smaller number of interactions in more detail, focusing on individual interactions rather than on trying to include all the attributes.

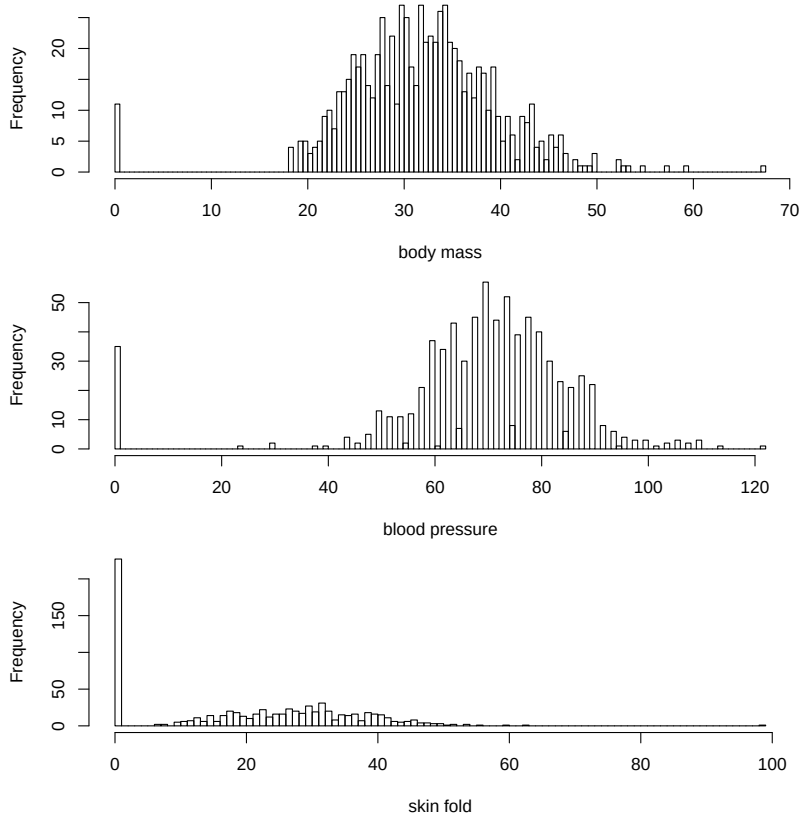


Figure 6.9: The histogram of attribute values for three attributes in the 'pima' data set indicates that there are frequent but unmarked missing values, coded as 0 or as 99. Furthermore, we can clearly see the influence of rounding and a few potential outliers.

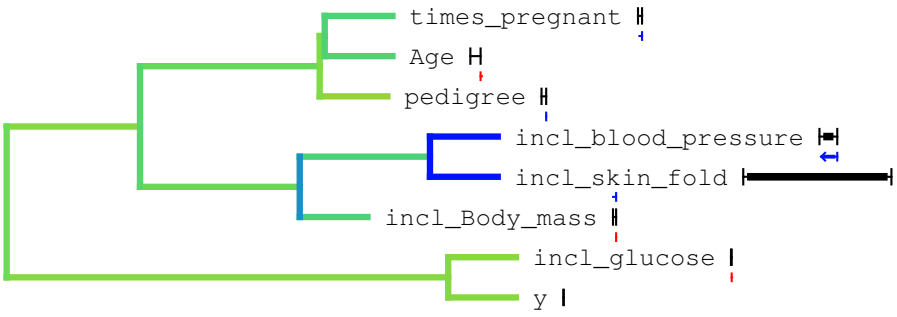


Figure 6.10: This interaction dendrogram was built with the inclusion attribute of *insulin* as the label. We can see that the inclusion of the *insulin* attribute is associated with the inclusion of the *skin fold* attribute. Fortunately, however, the association of the inclusion attribute with the main attributes, especially with the outcome y, is negligible. Therefore, we can safely assume that the values of *insulin* are missing at random.

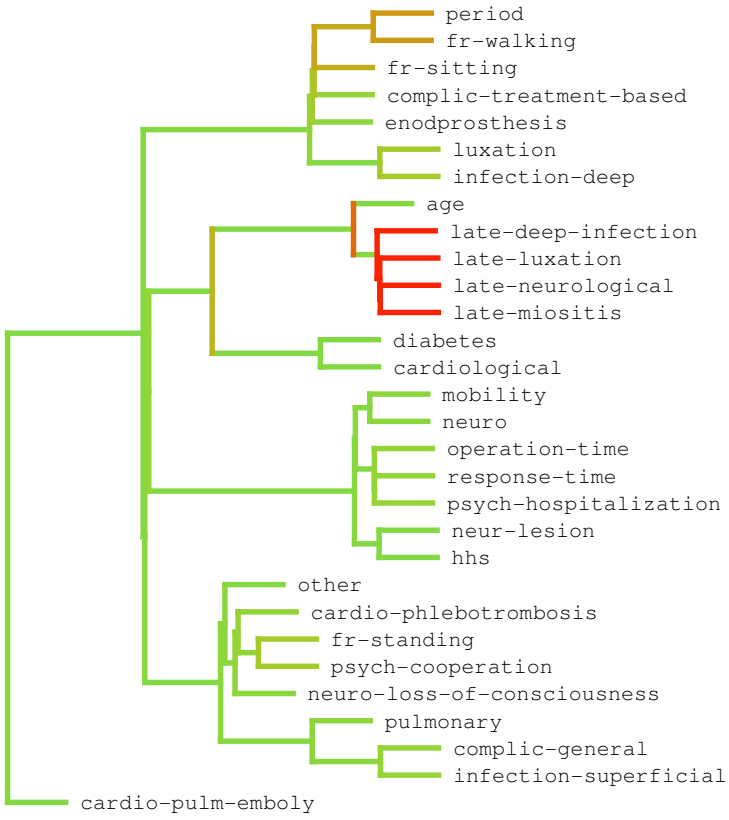


Figure 6.11: This unsupervised interaction dendrogram indicates which attributes have changed most with time. The *period* attribute appears in the topmost cluster accompanied by *fr-walking*, *fr-sitting*, *complications-treatment-based* and *endoprosthesis*. A more detailed analysis would reveal that in the second phase there were fewer complications, a certain endoprosthesis type was no longer used, and the evaluation of the walking and sitting ability changed considerably. The dendrogram also serves as a taxonomy of attributes that disregards the label.

To reduce clutter, only the strongest N interactions are shown, usually $5 \leq N \leq 20$. For classification, all visualized interactions should involve the label. This limitation is associated with the clarity of the visualization. It is extremely important to note that an interaction graph does not attempt to provide a predictive model, but merely to visualize the major interactions among the attributes, neglecting many weaker interactions and all potential higher-order interactions. Furthermore, with an interactive method for graph exploration, more interactions could be included. Fortunately, the distribution of interaction gains usually follows a bell-like distribution, with only a few interactions standing out from the crowd, either on the positive or on the negative side, so the presentation does capture the most distinct ones.

Each node in the interaction graph corresponds to an attribute. The information gain of each attribute is expressed as a percentage of the label entropy $H(Y)$, and written below the attribute name. There are two kinds of edges, bidirectional arrows and undirected dashed arcs. Arcs indicate negative interactions, implying that the two attributes provide partly the same information. The amount of shared information, as a percentage of the class entropy, labels the arc. Analogously, the amount of novel information labels the arrow, indicating a positive interaction between a pair of attributes. Figure 6.12 explains the interpretation of the interaction graph, while Figs. 6.13 and 6.14 illustrate two domains. We used the 'dot' utility (Koutsofios and North, 1996) for generating the graph.

Clearly, it is possible to apply interaction graphs on unsupervised learning problems as well. For an example, we again consider the US Senate in 2003. The nodes in the graph correspond to senators, and edges to their similarities. We only select a certain number of the strongest similarities to create a graph, using a threshold to discriminate between a strong similarity and the absence of it. Fig. 6.15 graphically illustrates the 20 pairs of senators with the highest Rajski's distance between their votes.

6.3.1 Interaction Graphs with P-Values

The decision about the strength of an interaction can be made either on the basis of interaction information, or based on the P-value that corresponds to the interaction. Usually we select some threshold that the P-values need to obey, which prevents spurious uncertain interactions. From those that remain, we visualize as many as we can, preferring to include those ones with high interaction information magnitude.

We have employed the significance testing approach to identify the significant interactions in a graph. We have used the Kirkwood superposition approximation to come up with the model, and have used the goodness-of-fit test to decide whether the model significantly deviates from the no-interaction estimate. This way we construct a model of the significant 2-way and 3-way interactions for a supervised learning domain. The resulting interaction graph is a map of the dependencies between the label and other attributes and is illustrated in Figs. 6.16 and 6.17. Kirkwood superposition approximation observed both negative and positive interactions. However, these interactions may sometimes be explained with a model assuming conditional independence: sometimes the loss of removing a negatively interacting attribute is lower than imperfectly modelling a 3-way dependence. Also, if two attributes are conditionally independent given the label, they will still appear redundant.

The interaction graph does not attempt to minimize any global fitness criterion, and should be seen as a very approximate guideline to what the model should look like. It

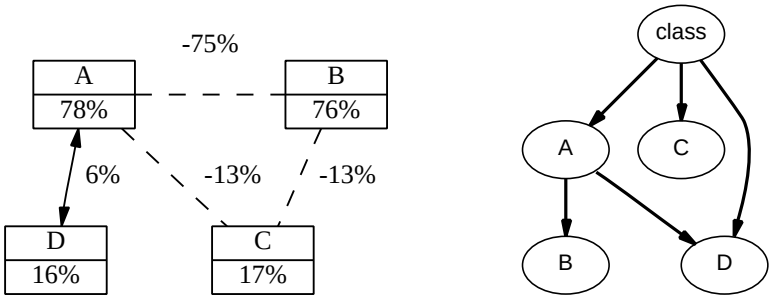


Figure 6.12: The four most informative attributes were selected from a real medical domain. In the interaction graph (left), the most important attribute A alone eliminates 78% of class entropy. The second most important attribute B alone eliminates 76% of class entropy, but A and B interact negatively (dashed arc), and share 75% of class entropy. So B reduces class entropy by only 76-75=1% of its truly own once we have accounted for A: but if we leave B out in feature subset selection, we are giving this information up. Similarly, C provides 4% of its own information, while the remaining 13% is contained in both, A and B. Attribute D provides 'only' 16% of information, but if we account for the positive interaction between A and D (solid bidirectional arrow), we provide for 78+16+6=100% of class entropy. Consequently, only attributes A and D are needed, and they should be treated as dependent. A Bayesian network (Myllymaki et al., 2002) learned from the domain data (right) is arguably less informative, as it only captures the strongest two interactions AB and AD, but not BC and AC.

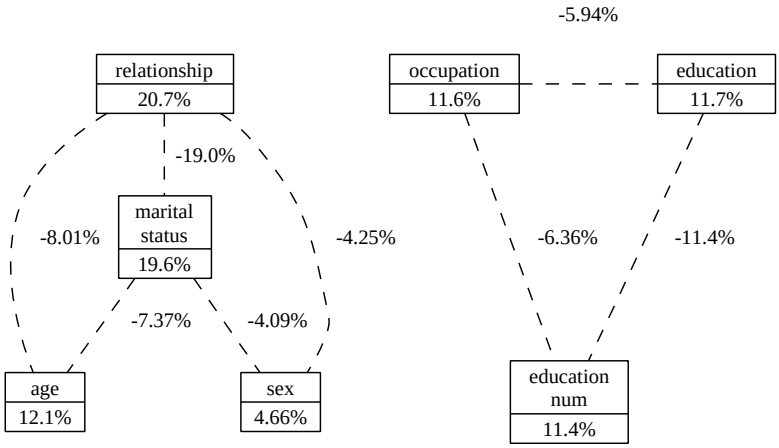


Figure 6.13: An interaction graph containing eight of the 3-way interactions with the largest interaction magnitude in the 'adult/census' domain, where the label is the income. The most informative attribute is *relationship* (describing the role of the individual in his family), and the mutual information between the label and *relationship* amounts to 20.7% of the label's entropy. All interactions in this graph are negative, but there are two clusters of them. The negative interaction between *relationship*, *marital status* and the label, *income*, comprises 19% of the income's entropy. If we wanted to know how much information we gained about the income from these two attributes, we would sum up the mutual information for both 2-way interactions and the 3-way interaction information: $20.7 + 19.6 - 19 = 21.3\%$ of entropy was eliminated using both attributes. Once we knew the relationship of a person, the marital status further eliminated only 0.6% of the income's entropy.

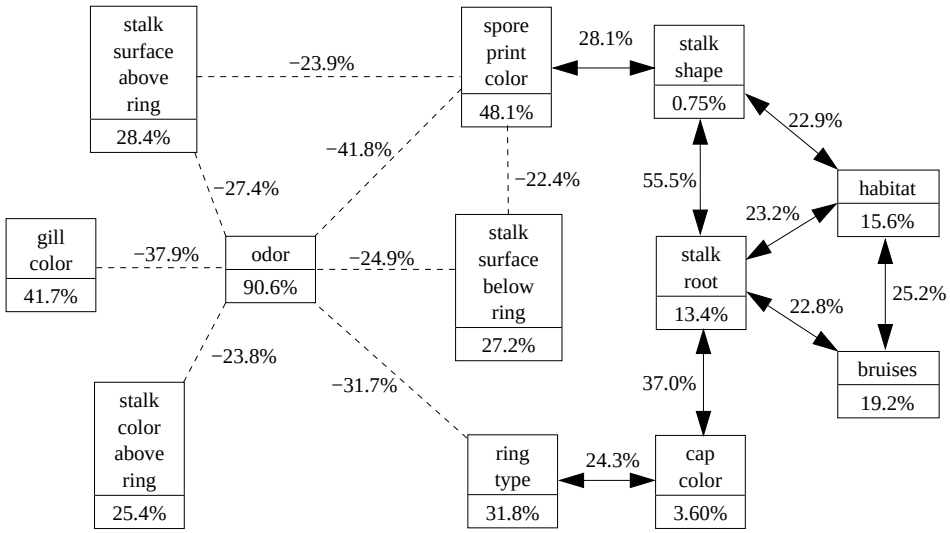


Figure 6.14: An interaction graph containing eight of the positive and eight of the negative 3-way interactions with the largest interaction magnitude in the 'mushroom' domain. The positive interactions are indicated by solid arrows. As an example, let us consider the positive interaction between *stalk* and *stalk root shape*. Individually, *stalk root shape* eliminates 13.4%, while *stalk shape* only 0.75% of the entropy of *edibility*. If we exploit the synergy, we gain additional 55.5% of entropy. Together, these two attributes eliminate almost 70% of our uncertainty about a mushroom's *edibility*.

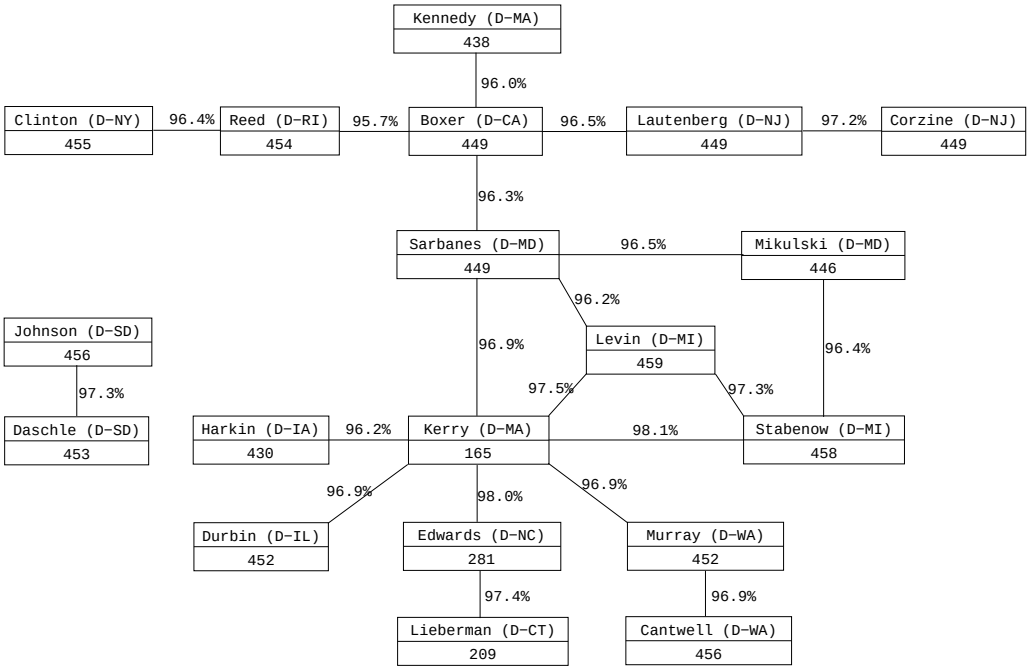


Figure 6.15: The nodes are labelled with the total number of votes cast, while the edges are marked with the percentage of those roll calls in which both senators voted and cast the same vote.

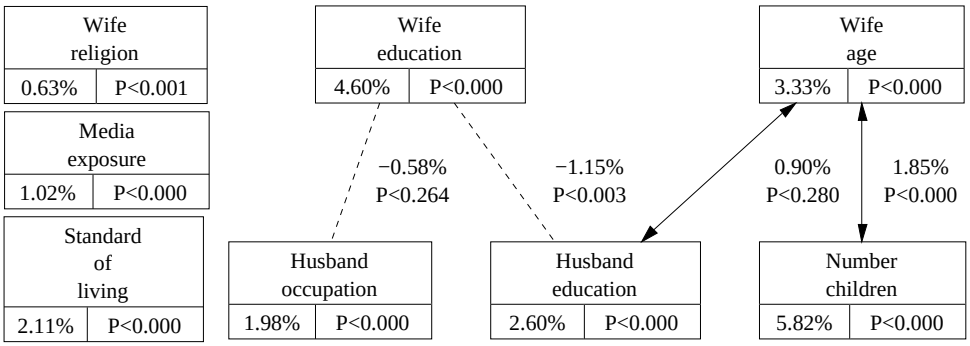


Figure 6.16: An interaction graph is illustrating interactions between the attributes and the label in the 'CMC' domain. The label in this domain is the *contraception method* used by a couple. The chosen P-value cutoff of 0.3 also eliminated one of the attributes (*wife working*). The strongest interaction between *wife age* and *number of children* among other things captures the pattern that wives without children do not use any contraception.

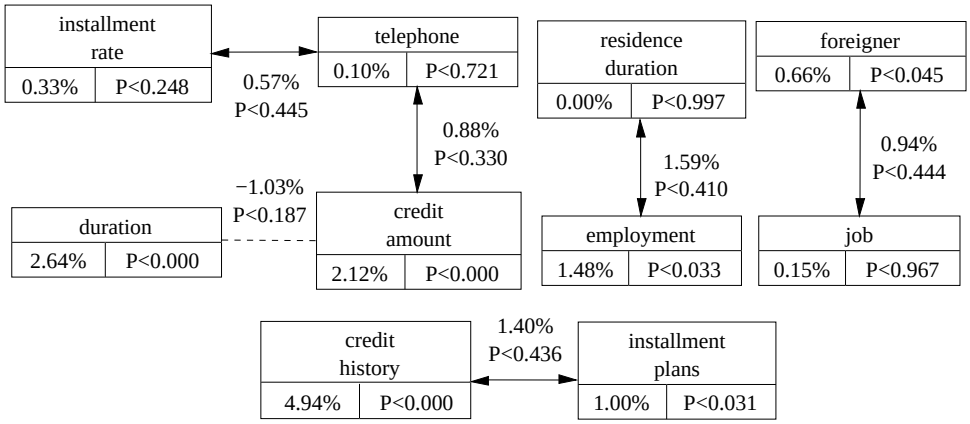


Figure 6.17: Only the significant interactions of the 'German credit' domain are shown in this graph, where the P-value cutoff is 0.5. The label in the domain is *credit risk*. Most notably, attributes *telephone*, *residence duration* and *job* are only useful as a part of a 3-way interaction, but not alone. We can consider them to be moderators.

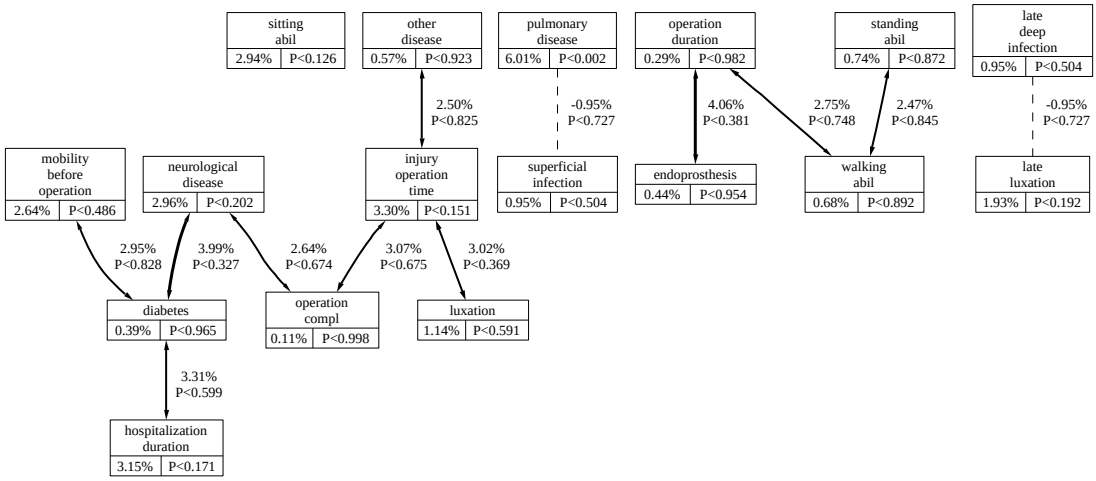


Figure 6.18: The interactions in the 'Harris hip score' data set are shown along with their P-values.

may also turn out that some attributes may be dropped. For example, the results from Sect. 4.4.2 indicate that the Kirkwood superposition approximation is not uniformly better than conditional independence models. So, one of the conditional independence models for a triplet of attributes could fit the data better than Kirkwood superposition approximation, and the interaction would no longer be considered significant. Nevertheless, it was already Freeman (1971) who proposed using the M tuples of attributes with the highest 2-way and positive 3-way interaction information as a heuristic to construct a model, inspired by and extending the work of Chow and Liu (1968)!

Another issue is that of multiple testing. As we have explained in Sect. 4.5.3, multiple testing corrections are based upon the underlying P-values along with assumptions about the dependencies between them. We provide a ranked list of the P-values in the scheme, and it is possible to employ one of the existing corrections to account for multiple testing: Streitberg (1999) recommends the Holm procedure. However, we do not make any kind of model-based decisions with the interaction graph. It is only a partial overview of certain dependencies in the data. Here, P-values are informative because they account for the complexity of the underlying interaction. Furthermore, they are grounded through an actual predictive model and its error, rather than through entropy decomposition which may fail to account for dependencies that yield zero interaction information.

We have also asked an expert to comment the interaction graph for the Harris hip score domain shown in Fig. 6.18 (Jakulin et al., 2003); the interactions surprised her (she would not immediately think about these if she would be required to name them), but could all justify them well. For instance, with her knowledge or knowledge obtained from the literature, specific (bipolar) type of endoprosthesis and short duration of operation significantly increases the chances of a good outcome. The presence of neurological disease is a high risk factor only in the presence of other complications during operation. It was harder for her to understand the concept of negative interactions, but she could confirm that the attributes related in this graph are indeed, as expected, correlated with one another. In general, she found the positive interactions more revealing and interesting.

It later turned out that only two negative interactions were at least somewhat sig-

nificant, so the expert was correct in disagreeing with the negative interactions. On the other hand, the positive interactions have distinctly higher significance, but still relatively low in the context of clinical research. For that reason, we would need to perform a decision-theoretic study to determine the clinical implications of taking those interactions into consideration.

6.3.2 Confidence Intervals and Attribute Importance

To demonstrate the application of mixture models to interaction analysis, we analyzed two UCI regression data sets, 'imports-85' and 'Boston housing'. Three kinds of models were learned: (1) the label alone, (2) each unlabelled attribute with the label, and (3) each pair of unlabelled attributes with the label. For each tuple, a five-component joint mixture model was estimated using the EM algorithm. Because both data sets are regression problems, the outcome was always included in the model.

The interaction information for each of these models was estimated along with its 95% confidence interval. This corresponds to an application of VaR (Sect. 4.5.4). For performance reasons we have employed the vagueness of loss method (Sect. 4.2.3), approximating the distribution of loss by computing the KL-divergence for each instance, and using the bootstrap replications over instances to compute the percentiles. The sample information gain was expressed as a proportion of the label sample entropy. The numbers below each attribute indicate the proportion of label entropy the attribute eliminates, with a bottom bound. The bottom bound corresponds to $\mathrm{VaR}_{0.025,\hat{P}}$, as interaction information is nothing but a change in utility. The interaction information was expressed as a percentage of the outcome entropy alone. The percentages are not always sensible for probability density functions, but with care they can nevertheless be more interpretable than bits of information.

Figure 6.19 shows the interactions between attributes applied to predicting the price of the car. Now consider this example of a greedily built regression model for car prices:

```
                Estimate  Std.Error   t-val  Pr(>|t|)
(Intercept) -32254.698  17385.307  -1.855   0.0651 .
curb.weight     13.126      1.406   9.333   <2e-16 ***
width          753.987    313.931   2.402   0.0173 *
height        -316.178    148.979  -2.122   0.0351 *
length        -119.198     64.586  -1.846   0.0665 .
```

The 'estimate' lists the combination of coefficient values that resulted in minimum loss. Often, the 'estimate' is considered to be a measure of importance of a particular attribute, but it does not account for the variance of that attribute. The 'std. error' is the standard error of the coefficient. The t-value is the t statistic, indicating the importance of the attribute. $\Pr(> |t|)$ expresses the t statistic relative to the null distribution. The asterisks indicate the importance of a particular attribute, based on the $\Pr(> |t|)$.

From the multiple regression model it would seem that the length, height, and width of the automobile are not particularly relevant about its price. This pitfall inherent to conditional models is avoided by constructing joint models, performing model comparisons, and by using information-theoretic examination of interactions in the models. This way, looking at Fig. 6.19, we would observe that *length* gives us very little additional information about the car price once we already know *curb.weight*, but in case the weight is not known, length alone is nevertheless quite a useful attribute.

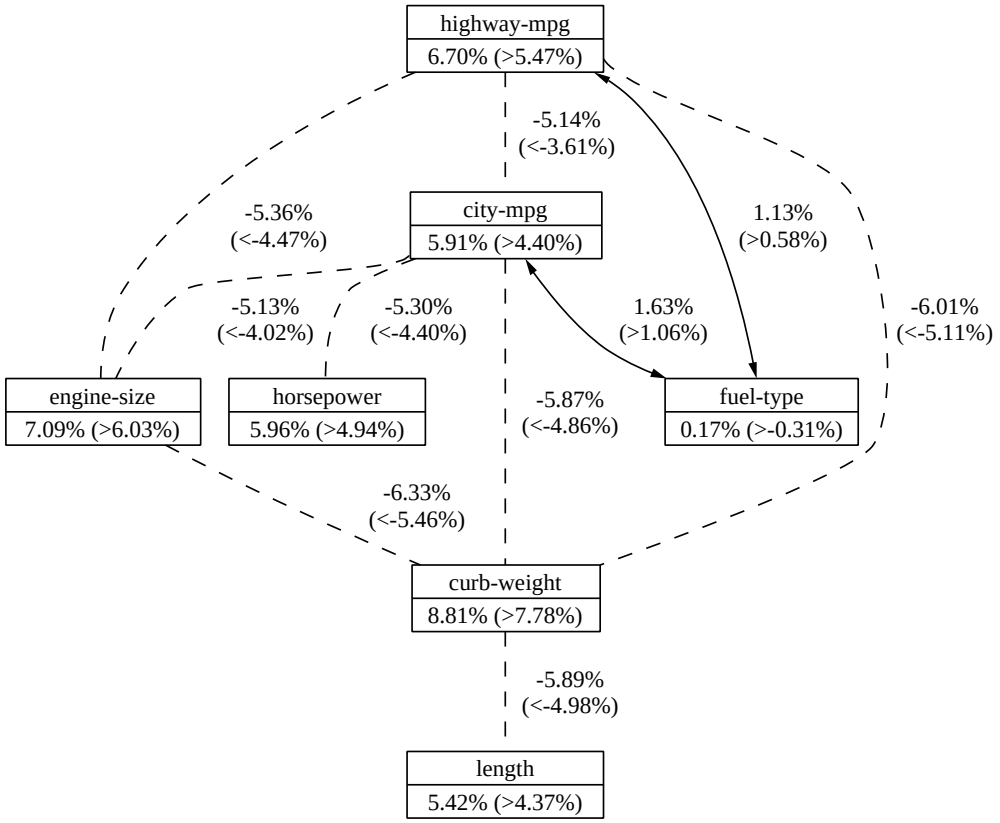


Figure 6.19: This interaction graph identifies the strongest 2-way and 3-way interactions in the 'imports-85' data set with the price of a car as the continuous label. For example, *highway mpg* alone eliminates 6.7% of uncertainty about the price on average, but in 97.5% of cases more than 5.5%. *fuel type* is apparently a useless attribute on its own, eliminating only 0.2% of entropy, but there is a positive interaction or a synergy between fuel type and the fuel consumption on the highway, eliminating an additional 1.13% of label entropy. Dashed edges indicate negative interactions or redundancies, where two attributes provide partly the same information about the label. For example, should we consider the fuel consumption both on highways and in the city, the total amount of label entropy eliminated would be $6.7 + 5.9 - 5.1$ percent, accounting for their overlap. Due to the imprecision of sample entropy and the unsupervised modelling criteria, apparent illogicalities may appear: the length of the automobile is hurting the predictions of the car's price in combination the car's weight.

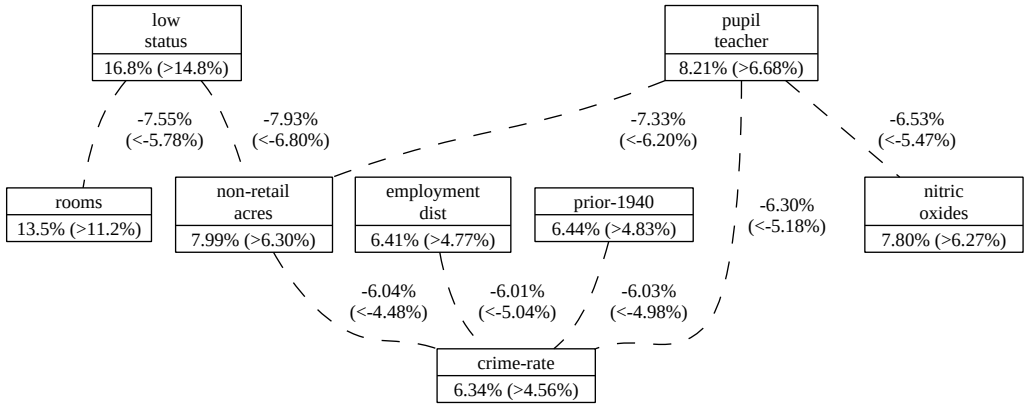


Figure 6.20: The strongest two-way and three-way interactions involving the label in the model of 'Boston housing'.

Figure 6.20 illustrates the result of interaction analysis of the 'Boston housing' data set. The outcome of interest is the median value of apartment in a certain area, as predicted by various properties of the area, such as unemployment, crime rate, pollution, etc. The most informative attribute is the proportion of lower status population. In the context of this attribute, *non-retail acres* becomes almost totally uninformative ($7.99 - 7.93 = 0.06$). Another useful attribute is *crime-rate*, which subsumes most of the information provided by *prior-1940* and *employment-dist*. Furthermore, a strong negative interaction between *pupil-teacher* and *nitric-oxides* must be noted. Although most negative interactions are due to correlations between attributes, these two are themselves not highly correlated, and the negative interaction is nonlinear in character. At low levels of pollution, the housing value is mostly independent of pollution given the pupil-teacher ratio. On the other hand, at higher levels of pollution, the pupil-teacher ratio does not vary.

Using the above interaction graph it is also possible to understand why *non-retail acres* and *prior 1940* prove to be insignificant (with P-values of 0.74 and 0.96, respectively) in a multiple regression model (R Development Core Team, 2004), even if they are significant on their own:

```
                 Estimate  Std.Error    t-val  Pr(>|t|)
(Intercept)       3.646e+01  5.103e+00   7.144 3.28e-12 ***
crime.rate       -1.080e-01  3.286e-02  -3.287 0.001087 **
zoned.lots        4.642e-02  1.373e-02   3.382 0.000778 ***
non.retail.acres  2.056e-02  6.150e-02   0.334 0.738288
Charles.River     2.687e+00  8.616e-01   3.118 0.001925 **
nitric.oxides    -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rooms             3.810e+00  4.179e-01   9.116  < 2e-16 ***
prior.1940        6.922e-04  1.321e-02   0.052 0.958230
employment.dist  -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
highways          3.060e-01  6.635e-02   4.613 5.07e-06 ***
property.tax     -1.233e-02  3.761e-03  -3.280 0.001112 **
pupil.teacher    -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
B                 9.312e-03  2.686e-03   3.467 0.000573 ***
low.status       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
```

These attributes are not irrelevant, they merely become insignificant in the context of

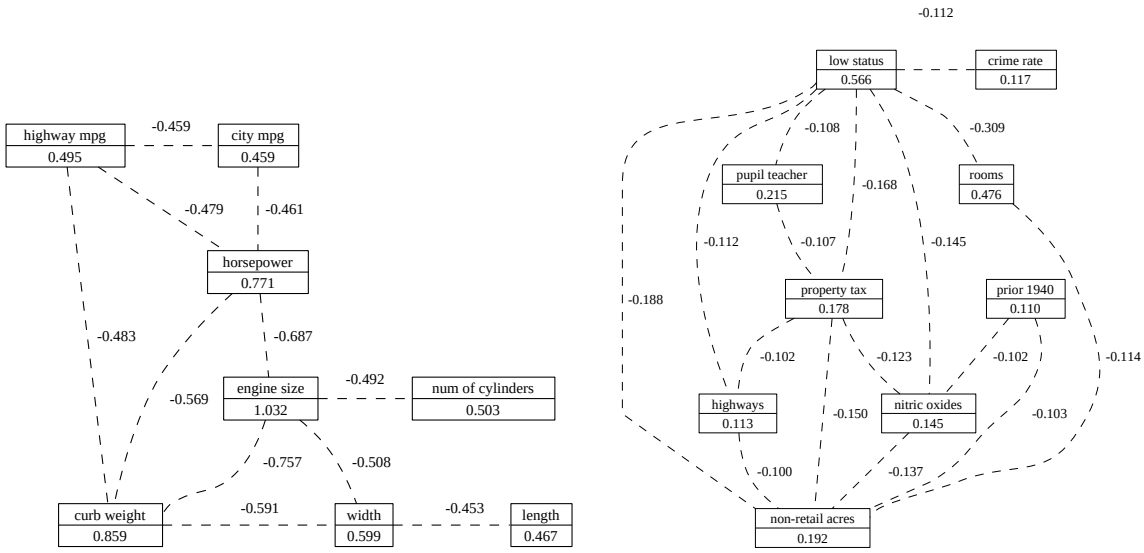


Figure 6.21: The strongest two-way and three-way interactions involving the label in the model of 'Boston housing' and 'imports-85' based on a multivariate normal model.

other attributes, such as *low status*. Of course, deciding which attribute should get the credit for predicting the outcome is often arbitrary: we may greedily credit just the best attribute, or we may be egalitarian in distributing the information credit among them all.

For comparison with the multivariate normal model that was described in Sect. 5.2, we show the interaction graph in Fig. 6.21. The units are bits, but the corresponding correlation coefficients can be looked up in Table. 5.1. It may interesting to compare these two graphs with the results obtained with the mixture model used for Figs. 6.19 and 6.20. There do not seem to be major differences, but the example should illuminate the fact that an interaction graph depends on the underlying probability model.

6.4 Interaction Drilling

The previous sections have focused on mapping the whole data set. The purpose of interaction dendrograms and matrices is to focus on a particular cluster of attributes, or to choose a representative attribute from each cluster. In the next phase, we use interaction graphs to evaluate the interactions in more detail. From the graph, we choose interactions that are significant, informative and interesting. We then focus on an interaction and examine it in more detail. This kind of analysis is performed for only a small set of attributes at once. This section will describe ways of performing this kind of localized analysis. The techniques are dependent on the model used. We will discuss nominal (unordered discrete) and continuous attributes.

6.4.1 Inside an Interaction

We have taken a record of all the marriages in Hawaii in 2002 (Hawaii State Department of Health, 2002). For each marriage, the nationality of the groom and of the bride were

recorded. We can interpret the pair of nationalities as attributes, groom's nationality G and bride's nationality B. The ranges of both attributes are the same. We can perform a test of the 2-way interaction, and find out that it is highly significant. However, the result of the test does not illuminate the nature of the interaction.

Recalling the interpretation of KL-divergence in Sect. 3.1.1, we can see that the loss is computed for every attribute value combination. The independence assuming model $P(G)P(B)$ implies that there is no pattern to intermarriage: all we control is the total number of grooms and brides of a particular nationality. Alternatively, we control for all possible intermarriage pairs in $P(G, B)$. But the error might not be the same for all pairs!

Again, we may resort to visualization. For each attribute value combination, such as (groom:Caucasian, bride:Causasian), we can calculate the probability of occurrence under the independence assumption $\hat{p} = P(G = \text{Caucasian})P(B = \text{Caucasian})$. Furthermore, we can calculate the probability under the dependence assumption $p = P(G = \text{Caucasian}, B = \text{Caucasian})$. We can compute a very simple error measure $p - \hat{p}$ that will reveal the extent of deviation from independence.

However, because of chance, the deviations for frequent attribute value combinations would swamp significant deviations for less frequent combinations. For that reason it is preferable to apply the *standardized Pearson residuals* (Agresti, 2002):

$$d(g, b) \triangleq \sqrt{n}\frac{P(G = g, B = b) - P(G = g)P(B = b)}{\sqrt{P(G = g)P(B = b)(1 - P(G = g))(1 - P(B = b))}} \tag{6.12}$$

If n is the data set size, $d(g, b)$ asymptotically has a standard normal distribution Normal$(0, 1)$. Therefore, under the null independence model it has the same scale for all value combinations. This makes it particularly appropriate for visualization. Fig. 6.22 demonstrates how the independence and dependence models can be visualized along with the standardized Pearson residuals between them.

6.4.2 Rules from Interactions

While we can add interactions one by one, finding the most salient one using the present visualization or, e.g., the mosaic plots (Theus and Lauer, 1999), it is often unnecessary to include the interaction as a whole: only a subset of situations is the source of the deviations from independence. The deviations in Fig. 6.22 can be captured by additional *rules* that account for the significant exceptions from independence. However, there is no single 'correct' rule when it comes to this. For example, there are two rules, the first specific and the second general:

1. Caucasian brides tend to marry Caucasian grooms.

2. Grooms and brides of the same nationality prefer to marry one another.

We can represent each rule as an additional attribute (Della Pietra et al., 1997). For example, for the first rule, we form an attribute R_1 with the range:

$$\Re_{R_1} = \{\text{bride} = \text{Caucasian} \wedge \text{groom} = \text{Caucasian}, \text{bride} \neq \text{Caucasian} \vee \text{groom} \neq \text{Caucasian}\}$$

In the second case, we form a *relational attribute* R_2 with the range:

$$\Re_{R_2} = \{\text{bride nationality} = \text{groom nationality}, \text{bride nationality} \neq \text{groom nationality}\}$$

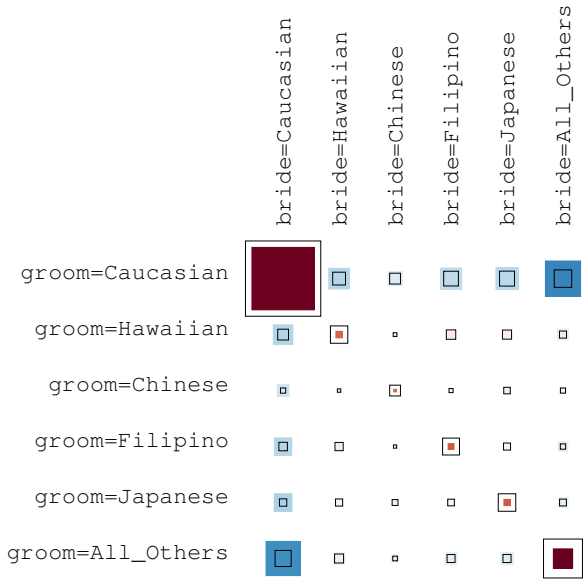


Figure 6.22: The marriages between brides and grooms in Hawaii are not independent of their nationalities. The colored rectangle indicates the prediction under the assumption of independence. The black rectangle indicates the truth. The area is proportional to the probability. The color reveals the standardized Pearson residual between the two models: blue means that the independence model overestimated, while red means that the independence model underestimated. The graph has a distinct red tone along the diagonal, which indicates the preference to marry a person of the same nationality.

Figure 6.23: If we account for the Caucasian tendency to intermarry with rule R_1, the errors decrease considerably. If we introduce exceptions in the independence-assuming model, renormalization introduces errors elsewhere (left). On the other hand, interpreting the rule R_1 as a binary attribute, and employing maximum entropy inference results in distinctly superior performance (right). The residuals are approximated by (6.12).

This relational attribute is closely associated with the notion of quasi-independence (Agresti, 2002). These two binary attributes can then be included into the model.

It is important to note, however, that they are definitely not independent of the previous two attributes, so we cannot exploit conditional and marginal independencies. Furthermore, in ascertaining the value of particular rules, we have to note that the importance of a particular rule depends on rules that are already in the model: there can be negative interactions between rules as well. Finally, the statistical model used in assessing the significance of rules should account for the fact that rules may be deterministically derived from the attribute values, so they should not be seen as an additional source of uncertainty.

It is possible, however, to merge certain attribute value combinations (Kononenko, 1991). It is clear simply replacing $P(g)P(b)$ with $P(g,b)$ would result in a non-normalized joint probability model. Re-normalization is simple enough:

$$\hat{P}(g,b) = \begin{cases} \frac{1-P(g,b)}{1-P(g)P(b)} P(g)P(b) & ;\ g \neq b \\ P(g,b) & ;\ g = b \end{cases} \tag{6.13}$$

If we employ this technique for R_1, the KL-divergence is reduced from 0.628 with the independence-assuming model, to 0.546 with this simple approach.

Unfortunately, the normalization causes the model to deviate from the true marginal statistics $P(G)$ and $P(B)$. The benefit from better approximating $P(g,b)$ may not exceed the losses caused by the normalization, as we have seen in Sect. 7.3.4. An alternative is to employ the iterative scaling algorithm to find the joint model with maximum entropy that still satisfies all the constraints: the marginals $P(G)$ and $P(B)$, and both rules $P(R_1)$ and $P(R_2)$. An additional benefit of iterative scaling is that R_2 is truly binary: we do not have to control for each individual nationality (e.g., Caucasian-Caucasian, Hawaiian-Hawaiian, etc.), just for the overall pattern. It is unclear how to include R_2 into consideration by working with exceptions.

It is interesting to examine the meaning of $P(R_1)$ and $P(R_2)$. For example, $P(R_1)$ indicates that 53% of marriages were between two Caucasians. By including just the R_1 into the model, and using the algorithm of Sect. 4.4.1, the KL-divergence is reduced dramatically to 0.247, even if the model has no more information than with (6.13). It is just that the iterative scaling algorithm tries to make sure that the joint model agrees both with the constraint of 53% marriages between Caucasians, and with the marginal statistics, such as that 63% of brides and 58% of grooms are Caucasian. The comparison is shown in Fig. 6.23.

If we introduce R_2, the KL-divergence falls to just 0.019. In general, R_2 is more powerful as it alone results in KL-divergence of 0.020. The meaning of $P(R_2)$ is that 77% of all marriages were between grooms and brides of the same origin. Again, iterative scaling makes sure that this constraint is fulfilled along with others.

Summary It is not necessary to assume an interaction between all attribute combinations. Each rule can be expressed as an attribute. Furthermore, we introduce relational rules into consideration. To include these attributes into consideration, methods such as maximum entropy may be applied to constrain the model with the rules. Instead of maximum entropy, we can apply closed-form methods, but they do not reach the same level of performance.

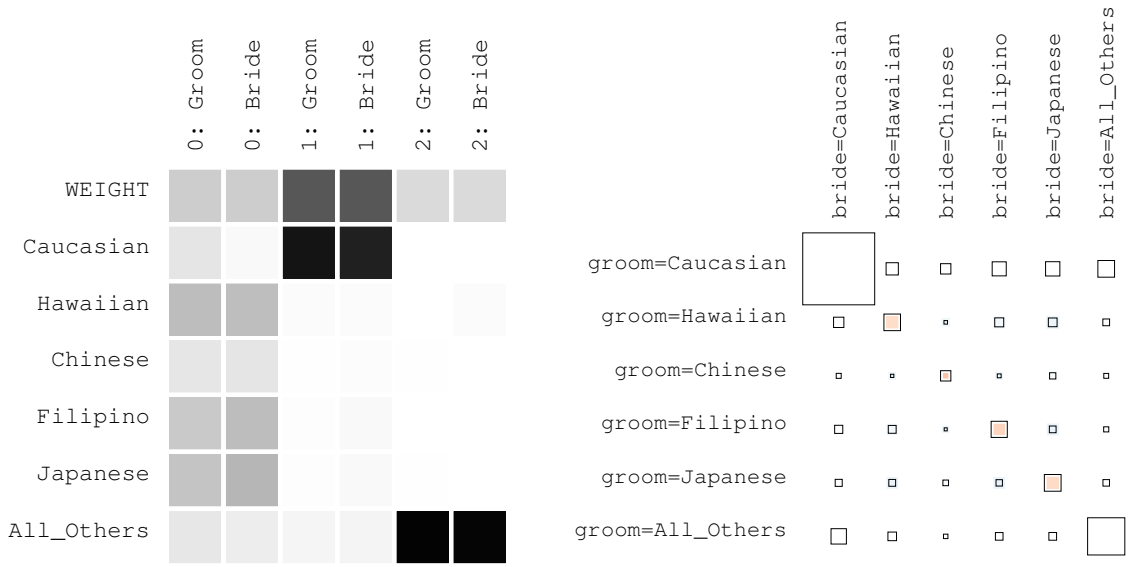


Figure 6.24: A latent class model may also be used to explain an interaction. 'Weight' indicates the number of instances covered by the class, essentially the marginal model of $P(C)$. The lines below show the distribution of attribute values corresponding to the component indicated in the column, essentially the model of $P(B|C)$ and $P(G|C)$ (left). The KL-divergence of the resulting model is relatively low at 0.076, but fails to capture the quasi-independence within class 0 (right).

6.4.3 Mixture Models

Discrete Attributes

It is possible to effectively visualize mixture models, which were defined in Sect. 5.3. First we will focus on discrete attributes by re-address the problem of Hawaiian marriages from Sect. 6.4.2. To do this, we will apply the *latent class* model, which is a special case of a mixture model (Agresti, 2002).

With the latent class model, we can identify groups of instances that contain no interactions. The interaction between the attributes is thus captured by the new class attribute C. The bride and groom attributes are assumed to be conditionally independent given C:

$$\hat{P}(B,G) = \sum_{c \in \Re_C} P(c)P(B|c)P(G|c) \tag{6.14}$$

The class C is essentially a constructed attribute that attempts to explain the interaction between B and G.

To obtain the model, we have performed 20 random initializations of C, followed by the EM algorithm to maximize the model's likelihood. From these attempts the one with the lowest KL-divergence was selected. This KL-divergence was 0.076: not as competitive as the quasi-independence model of the previous section, but considerably more competitive than the independence-assuming model.

The model can be shown visually, as we have done in Fig. 6.24. The first component covers 20% of the population and is an Asian/Pacific mixture, but with an excess of Caucasian grooms. The second component covers 66% of the population and is mostly

Caucasian, with a small excess of Japanese and Filipino brides. The third component describes 14% of population of unspecified nationality with a tiny surplus of Hawaiian brides. As we could have seen, the interpretability of the latent class model is quite good. Similar analysis is instinctively performed by humans: the latent class 'dog' can be seen as explaining the interactions between barking sounds, fur, sharp teeth and a wagging tail.

Continuous Attributes

Fig. 6.25 demonstrates the difference between a locally independent mixture of Gaussians and a general mixture of Gaussians. Generally, there is a trade-off between a larger number of components in a locally independent mixture, or a smaller number of components in a general finite mixture.

Another useful discovery that can be made about the data is evidence for multiple groups in data. Generally, the decision-theoretic value of structure is the reduction in entropy achieved by using K instead of K' components in a finite mixture model, $K > K'$. Structure allows a relatively simple model to capture complex non-linear relationships in data, not just multimodality. Through local analysis, we may investigate the structure aspect in small subsets of attributes. We do this by examining two models, a locally independent mixture model p that allows for structure, and a multivariate normal model q that only accounts for correlation:

$$p: \qquad \mathbf{x} \sim \sum_{k=1}^{5} \pi_k \prod_{i}^{d} \text{Normal}(\mu_{k,i}, \sigma_{k,i}) \tag{6.15}$$

$$\boldsymbol{\pi} \sim \text{Multinomial}(\boldsymbol{\lambda}, 1),\ \sum_{k} \lambda_k = 1 \tag{6.16}$$

$$q: \qquad \mathbf{x} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{6.17}$$

When the KL-divergence $D(p\|q)$ is large, we have gained information through the assumption of structure, and this is what often makes a projection interesting. The results of such analysis are illustrated in Fig. 6.26 for the 'Boston housing' data set shows the pair of attributes with the maximum and the minimum $D(p\|q)$.

Each component can also be viewed as a separate rule, as a leaf in a classification tree, as a prototypical instance, or as a support vector. For example, the component identifying living people can be described with $temp = 37℃ \pm 10$, while the component identifying healthy people is $temp = 37℃ \pm 2$.

3-way interactions involving continuous attributes are not as easily visualized in two dimensions. Let us focus on the interaction involving pollution, pupil/teacher ratio and the housing prices. This is a negative interaction, which is slightly weaker than others in the data set, so it does not appear in Fig. 6.20. Nevertheless, it is an interaction, and it can be visualized.

It is possible to give up the information on the variability in the labelled attribute, housing price, by only providing the mean. To obtain the mean, we employ a regression model. For this task, we employed a support vector machine with a polynomial kernel (Chang and Lin, 2005). Because the distribution of pupil/teacher ratio and the pollution are not independent, it is helpful to indicate the projections of instances so that we nevertheless see where the regression model is interpolating (either in the areas of low pollution and a low pupil/teacher ratio or in the areas of high pollution and a high pupil/teacher

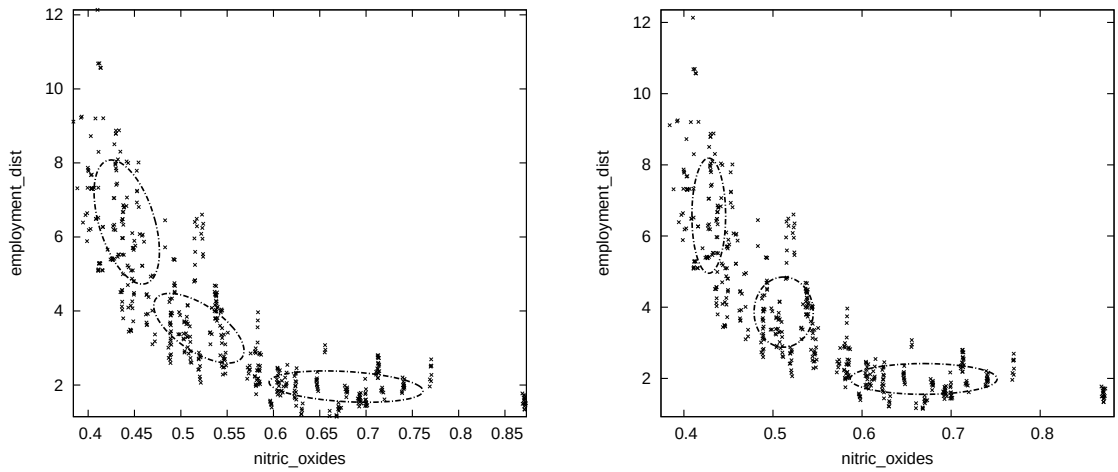


Figure 6.25: A multivariate normal mixture may unravel nonlinearities in the data. In this example, the pollution as measured by the concentration of nitric oxides is non-linearly decreasing with distance. The ellipses depict the circumference of each component at one standard deviation in the reference mixture model. Each component captures localized linearity in an area of the attribute space (left), achieving the logarithmic loss of 0.561 bits. A locally independent mixture model is not as efficient, but nevertheless achieves a relatively good result in terms of logarithmic loss (0.581 bits) (right). This can be compared to 2.05 bits achieved by the the locally independent model with a single component, and 1.41 bits by the multivariate normal model: both mixtures are distinctly better.

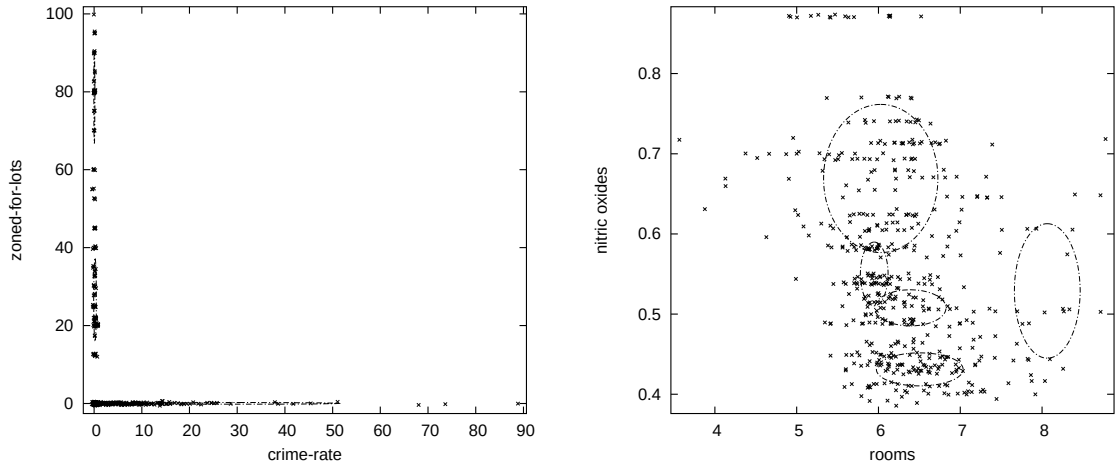


Figure 6.26: For the 'Boston housing' data set, the scatter plot on the top illustrates the nonlinear dependence between *crime rate* and *zoned for lots*, which has the highest amount of structure among all attribute pairs. On the other hand, structure is not of considerable utility to the model of *nitric oxides* and *rooms* (bottom).

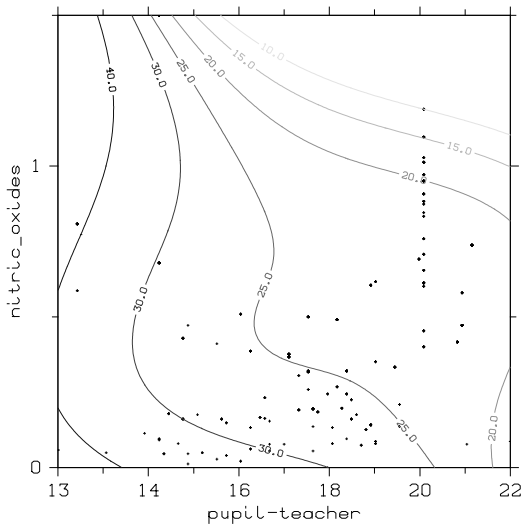


Figure 6.27: The nonlinear dependence between housing values, pollution, and quality of education is modelled using SVM regression with a polynomial kernel, and visualized with a contour plot.

ratio) and where it is extrapolating (in the areas of high pollution and a low pupil/teacher ratio – such areas do not exist). From Fig. 6.27 we can see that the gradient of average housing prices is distinctly non-linear.

6.4.4 Attribute Value Proximity Measures

In Sect. 6.2 we have seen that the concept of similarity needs not be seen as primary. Instead, similarity can be derived from a loss function, a probability model, and data. Our earlier discussion focused on the similarities between whole attributes. Yet, inside each attribute there are many values, and we might also discuss similarities between the values themselves.

Assume an attribute A with a range $\{a_1, a_2, \ldots, a_k\}$. This attribute can be transformed into a set of k binary attributes $\{A_1, A_2, \ldots, A_k\}$, where $A_i \triangleq \{A = a_i\}$. It would be meaningless trying to investigate mutual information between A_i and A_j: there is a distinct pattern of dependence, as only one of these attributes can take the value of 1 for a given instance.

A better definition of similarity would be based on whether other attributes can be predicted by distinguishing between a_i and a_j. Or, similarly, whether other attributes can be used to distinguish between a_i and a_j. Let us define a binary attribute $\dot{A}_{i,j}$ that distinguishes between a_i or a_j, and all other values. It is used to select only those instances where A takes the value of either a_i or a_j. The value $\dot{a}_{i,j}$ is therefore defined as:

$$\dot{a}_{i,j} \triangleq \begin{cases} 1 & ; \; A = a_i \vee A = a_j \\ 0 & ; \; \text{otherwise.} \end{cases}$$

Assuming other attributes to be $\mathcal{V}$, we can now define the *attribute-value distance*

using information theory:

$$\langle a_i, a_j \rangle_V \triangleq 1 - \frac{I(A; \mathcal{V} | \dot{A}_{i,j} = 1)}{H(A, \mathcal{V} | \dot{A}_{i,j} = 1)} \tag{6.18}$$

Because this is merely an application of Rajski's distance, it is a metric. Furthermore, because of the uncertainty regarding the probability model, it is an uncertain quantity. Finally, while it would in general be very rare for the attribute-value distance to be zero, interpreting interaction information as a model comparison might sometimes result in a considerable probability being assigned to the possibility that the two attribute values should not be distinguished. Distinguishing the two attribute values would result in a greater loss than not distinguishing them. In such a case, a negative attribute-value distance would correspond to the indistinguishability of the attribute values.

It is clear that if $\mathcal{V}$ contains many attributes, the attribute-value distance is not tractable: we face the usual assortment of problems associated with the curse of dimensionality. However, we can assume that higher-order interactions do not exist. If we assume only 2-way interactions between each attribute $X \in \mathcal{V}$ and the focal attribute $\check{A}_{i,j}$, the resulting Rajski's distance is as follows:

$$\langle a_i, a_j \rangle_V = 1 - \frac{\sum_{X \in \mathcal{V}} I(A; X | \dot{A}_{i,j} = 1)}{(1 - |\mathcal{V}|) H(A | \dot{A}_{i,j} = 1) + \sum_{X \in \mathcal{V}} H(A, X | \dot{A}_{i,j} = 1)} \tag{6.19}$$

For computing $\hat{H}$ we subtract the term corresponding to A that would otherwise be counted several times using the Bethe approximation to free energy (Yedidia et al. (2004), also see Sect. 8.2.4). Some care is required, as these approximations may not have all of the properties of Rajski's distance. For example, the value of (6.19) can be negative, because we are not accounting for the overlap between the 2-way interactions. If some interactions are to be assumed, we can include them in (6.19), and the distance $\langle \cdot \rangle_V$ will be affected.

We have used the 'census/adult' data set to examine the performance of this approach. The results of attribute-value clustering are shown in Fig. 6.28. Perhaps most notable is the taxonomy of occupations into the blue and white collar, with the service occupations in the blue collar group. By their education, we can divide individuals into those with elementary school, those with high school, those with university degrees, and those with junior college degrees. Finally, by their nationality, we can separate the Latin American, Asian and the European countries. There are some exceptions, especially among the less frequent groups. Among the European nations, we can clearly identify the Mediterranean cluster.

Instead of forming hierarchies, we can also convert nominal attributes into continuous attributes with meaningful distances between values. To do this, we can employ the multidimensional scaling algorithms (Borg and Groenen, 1997), we have employed the SMACOF algorithm (de Leeuw, 1977). The result for the *occupation* attribute is shown in Fig. 6.29, and can be easily paralleled with the clustering above. Such mapping of attributes is appropriate for adapting nominal attributes for various metric classifiers, such as support vector machines or the nearest neighbor algorithm.

Our definition of similarity based on dependence and independence differs from other definitions that have appeared in the literature. One of the first contributions on this

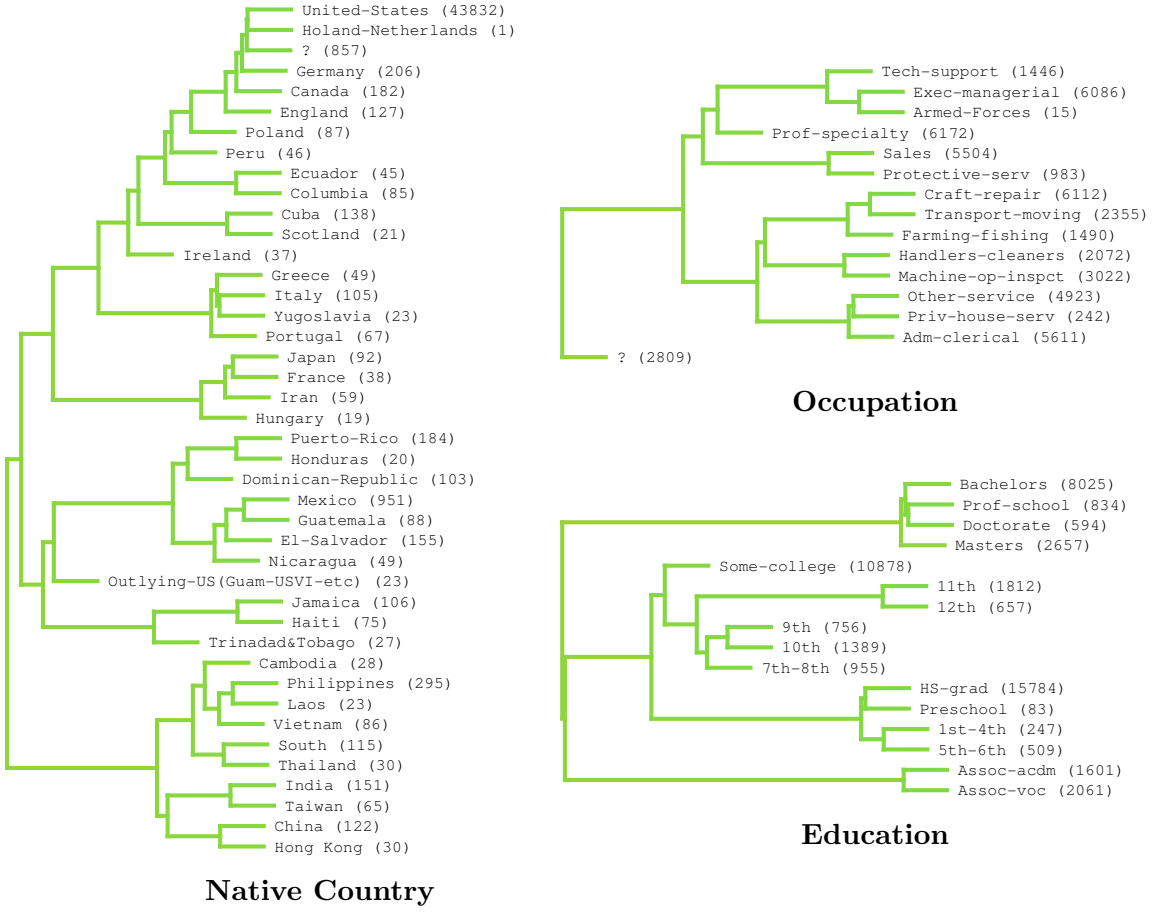


Figure 6.28: Hierarchical clustering based on attribute-value proximities clearly captures the similarities between values as derived from the data. The number in the brackets indicates the number of instances with the particular attribute value; for several values this number is too low to allow reliable positioning, noticeable especially for the native-country attribute. This can be seen as automated taxonomy construction.

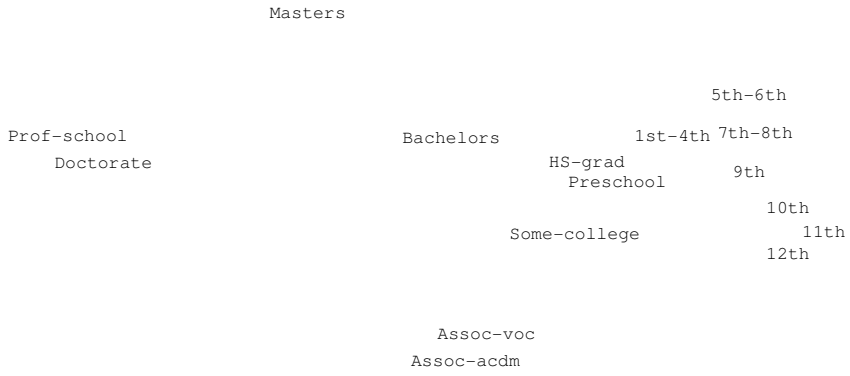


Figure 6.29: Applying multidimensional scaling to embed the attribute-value proximities into a euclidean space can help replace a nominal attribute by one, two or more continuous attributes. The euclidean distances between the positions of attribute values correspond to their information-theoretic proximities. This can be seen as continuous-valued constructive induction.

topic was the concept of *value difference metric* (VDM) (Stanfill and Waltz, 1986), where the difference between two nominal values x_1 and x_2 corresponds to the squared difference between $P(Y|x_1)$ and $P(Y|x_2)$, Y being the labelled attribute. Generalizing this idea, Baxter (1997) defined similarity between two attribute value combinations $\mathbf{x}$ and $\mathbf{x}'$ through a *canonical distortion measure* based on the relevant probabilities in the two contexts that correspond to $\mathbf{x}$ and $\mathbf{x}'$. The same definition was put in the context of supervised learning, Fisher information and KL-divergence by Kaski and Sinkkonen (2000). Recently, Kang et al. (2004) have employed this definition for attribute value clustering.

Our attribute-value distance should be distinguished from general distances that can defined directly on the attribute values. For example, each instance that enters support vector machine classification (Schölkopf and Smola, 2002) can be seen as described with a vector of distances to a specified set of support vectors. The support vectors comprise an informative subset of the training data set. The distance from the instance to a particular support vector corresponds to a distinct attribute in the corresponding model. This then allows the model to be essentially linear, as the distances and support vectors allow capturing the nonlinearity. Furthermore, the model is linear in the reproducing kernel Hilbert space that linearizes the non-linear kernel.

The choice of the kernel determines how the distance is computed. In that sense, we can redefine the kernel and the resulting distance in order to improve the classification performance, at the same time learning the metric. Learning the metric then corresponds to maximizing classification performance. This has been done in the context of support vector machines (Lanckriet et al., 2004), by maximizing the alignment between label similarities and the inferred similarities between instances. The similarities can be adjusted directly to maximize classification accuracy, as it has been done for nearest-neighbor classifiers (Wettschereck et al., 1997, Hastie and Tibshirani, 1996).

Our scheme goes in the opposite direction: the distance $\langle a_i, a_j \rangle_V$ is determined by the choice of the loss function, the model and the data. Furthermore, it is general in the sense that the similarity can be defined both between attribute values and between instances. The distance between instances is merely a special case of attribute-value proximity. We can define a new attribute *id*, and assign each instance a unique value. The resulting attribute-value proximity will serve as an constructed measure for assessing the distances between instances under a particular set of assumptions about the pattern of interactions between attributes.

If attribute-value clustering is performed on the labelled attribute, we may use classification algorithms that can benefit from this representation (Frank and Kramer, 2004). Alternatively, the attribute-value taxonomies for non-labelled attributes have also been found to benefit the classification performance (Kang et al., 2004). Of course, the attribute-value distance should reflect the predictive relevance with respect to predicting the label.

6.4.5 Latent Attributes

There are very many interactions of low order in some data sets. An example of such a data set are the roll call votes. As shown in Fig. 6.1, there is a tremendous number of 2-way interactions between the attributes. It is usually quite difficult to work with such models, and many assumptions are required. Although we have discussed the finite mixture models, latent attributes provide more versatility.

A solution people often intuitively employ is to seek latent factors that account for

the interactions. For example, people often use the concept of a *bloc* to explain the correlations between strategic or political preferences of senators. Alternatively, people infer a *dimension* and arrange the senators on an ideological axis. We will now explain how these concepts may be obtained automatically on the example of the US Senate votes in 2003.

There has been a tremendous amount of work on inference of latent structure, some of which we have already addressed in Ch. 4 and will revisit in Ch. 7. Some of them are based on information-theoretic criteria, like the information bottleneck (Tishby et al., 1999). Our intention is not to reinvent or replace, but to induce both continuous and discrete latent attributes on the same data, and compare their implications.

Blocs: Discrete Latent Attributes

Many clustering algorithms assign an instance wholly to a single cluster. For example, we have to decide whether a whale is a fish or a mammal: it cannot be a half-fish and a half-mammal. On the other hand, probabilistic clustering algorithms assign each instance a probability of membership in a particular cluster. Of course, the probabilities have to sum up to 1, so the clusters should be seen as mutually exclusive. A practical statistical model, *discrete PCA*, and the algorithm are described by Buntine and Jakulin (2004) as applied to the analysis of text.

Blocs are not clusters of instances. Instead, blocs are clusters both of attributes and of instances, of senators and of issues. In the latent class models, an instance is described by a probability distribution over classes, but each class covers all the attributes. In discrete PCA, however, a particular component covers both a distribution of instances and a distribution of attributes. For our example, we will refer to a group of attributes belonging to the same component as a *bloc*.

A bloc can be seen a fictional senator that has an opinion on every issue. A bloc may also be seen as a fixed ideological position. Again, the opinion may be probabilistic: the bloc may be uncertain about how to vote. On the other hand, each true senator can be seen as having a probability distribution of belonging to a particular bloc or another. This membership is assumed to be constant for the whole duration of the analysis. The votes in each issue are to be described solely using the blocs.

We can now pursue the parameters that agree with the data. We seek the opinions of the blocs across the issues, and we seek the memberships of senators in blocs. There is an important issue of the number of blocs. We can interpret this as a nuisance parameter and avoid making an arbitrary decision about the 'best' number of blocs (Neal, 2000). On the other hand, the information about blocs is informative to a human analyst, even if other configurations of memberships could also agree with the data.

We provide a novel visualization that combines the 2-way interaction dendrogram and the 5-valued bloc attribute in Fig. 6.30. We have examined multiple setting of the bloc number, and the setting of 5 proved distinctly likelier under our prior expectations for the parameter values. The setting of 4 blocs lost the nuances in data, and there was not enough data to distinguish a specific configuration of bloc membership when assuming more than 5 blocs. We can see that the the blocs have well-defined meanings: the Republican party is split into three blocs, the majority (A), the extreme minority (B) and the moderate minority (C); the Democratic party is split into the majority (E) and a moderate minority (D). It turns out that the outcome is primarily a member of A and E: we can explain the

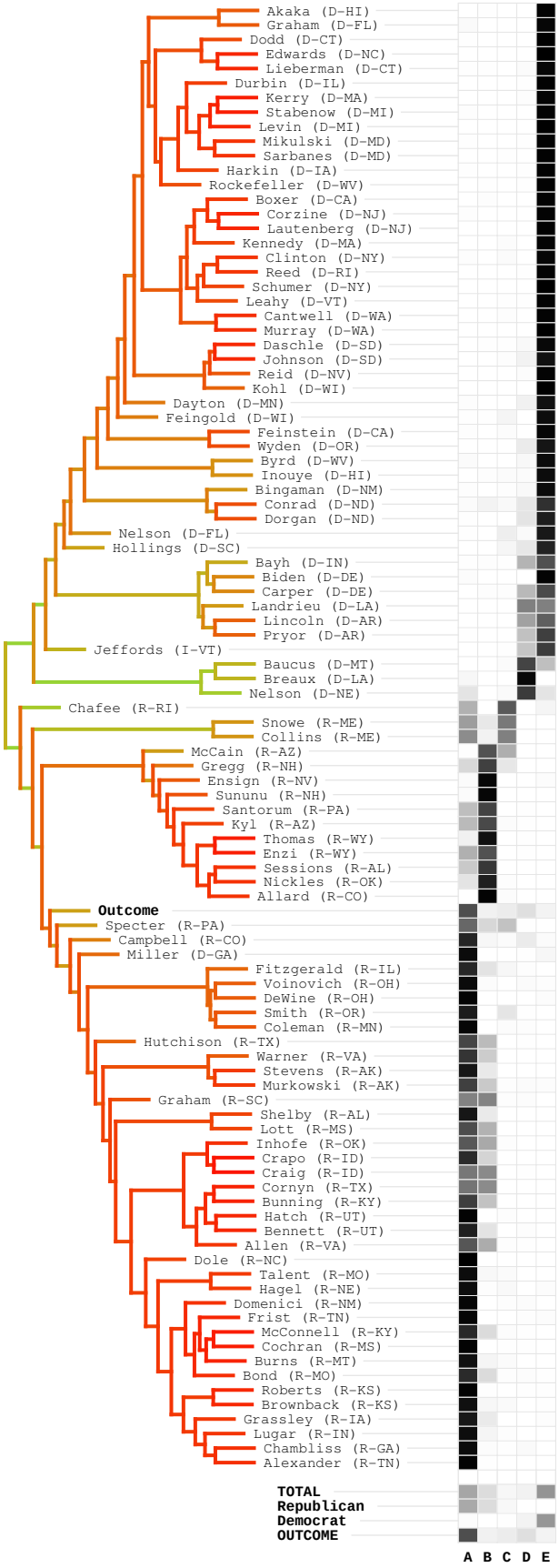


Figure 6.30: In the hierarchical clustering of senators based on their pair-wise Rajski's distance, we can identify the two major clusters: the Republican and the Democratic. Both the cluster color and the cluster height indicate the compactness of the cluster: green clusters are weakly connected, while red clusters are strongly connected. The bars on the right hand side depict the five blocs resulting from the discrete latent attribute analysis, the dark blocks indicating a high degree of membership.

outcome purely through through the interaction of these two blocs.

Of course, our analysis is just a particular model using which we are able to interpret the voting choices of the senators. While it may hopefully be clear and informative, it is not true or optimal in any way.

Dimensions: Continuous Latent Attributes

The task of the ubiquitous principal component analysis (PCA) or Karhunen-Loeve transformation (Press et al., 1992) is to reduce the number of dimensions, while retaining the variance of the data. The dimension reduction tries not to crush different points together, but to remove correlations. The remaining subset of dimensions are a compact summary of variation in the original data. The reduction can be denoted as $\mathbf{u} = \mathsf{W}(\mathbf{x} - \boldsymbol{\mu})$, where $\mathbf{u}$ is a 2-dimensional 'position' of a senator in a synthetic vote space obtained by a linear projection W from the V-dimensional representation of a senator.

The roll call data can be is represented as a $J \times V$ matrix $\mathsf{P} = \{p_{j,v}\}$. The J rows are senators, and the V columns are roll calls. If $p_{j,v}$ is 1, the j-th senator voted 'Yea' in the v-th roll call, and if it is -1, the vote was 'Nay'. If the senator did not vote, some value needs to be imputed, and we have used simply the outcome of the vote. The transformation W by applying the SVD algorithm to the centered matrix P: the centering is performed for each vote, by columns. The SVD represents the centered matrix $\mathsf{P} - \boldsymbol{\mu}$ as a product of three matrices: UDV^T, where U is a column-orthogonal matrix, V a square and orthogonal matrix, and D a diagonal matrix containing the singular values. The dimensionality-reduced 'locations' of senators are those columns of U that correspond to the two highest singular values, but they must be multiplied with the corresponding singular values. These two columns can be understood as uncorrelated latent votes that identify the ideological position of the senator. The position 'explains' the votes cast by a senator in roll calls, and similarities between positions 'explain' the associations between the votes.

We employed the ordinary SVD algorithm, and the results are shown in Fig. 6.31. We can see the distinct left-right axis that corresponds to the intuitive liberal-conservative dimension in perceiving politicians. Of course, SVD is not the best choice, and we use it merely for illustration: there are other scaling algorithms developed especially for roll call analysis (Clinton et al., 2004, de Leeuw, 2003, Poole, 2000), but the results are close.

6.5 Text Mining and the Curse of Dimensionality

In a typical text mining data set, the documents are represented with a bag of words. Each document is an instance, and the number of occurrences of each word is an attribute. In the popular Reuters-21578 benchmark data set there are over 21000 documents with over 38000 individual words. The complexity of even 2-way interaction analysis would be overwhelming with approximately 1.5 billion combinations. We will now describe two practical approaches that can be used to cope with such an onslaught of attributes.

A Probability Model for Interacting Words in Text

Before we can discuss interaction analysis, we need to define a probability model to be used for modelling text. Namely, how can we characterize an interaction between two

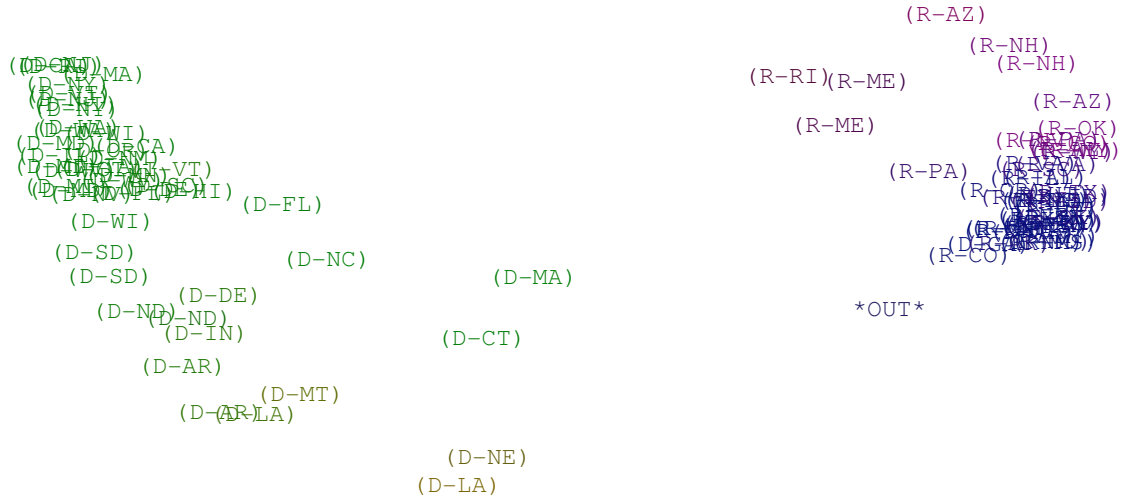


Figure 6.31: Ordinary principal component analysis manifests a distinct clustering of Republican and Democrat senators with some dispersion. 'OUT' denotes the outcome. It is possible to see all the clusters that appeared in earlier analysis: bloc A is the distinct dense cluster to the right, bloc B appears to the top of it, bloc C is left of B. The Democrat bloc D is in the bottom, the Democrat core is to the left. Senators Kerry (D-MA) and Lieberman (D-CT) appear in the center due to their infrequent voting and the imputation method we used. The coloring of individual senators is based on their membership in individual blocs: A is blue, B is purple, C is violet, D is yellow, and E is green.

words? The number of appearances of the word in a document can be seen as a discrete attribute, but it is often undesirable to treat the count as an attribute directly.

Instead, let us define a binary attribute A' for a word A with the following range $\Re_{A'} = \{1 : A \text{ appears}, 0 : A \text{ does not appear}\}$. Thus, k appearances of a word can be seen as k instances with $A' = 1$, and a document without the word can be seen as an instance with $A' = 0$. This model is quite suitable when we are classifying documents: the type of each document is another nominal attribute, and is the context in which a word can appear multiple times. This approach is referred to as the multinomial model in contrast with the multivariate Bernoulli model where we only consider whether the word appeared in a document or not, disregarding the count of appearances. The multinomial model was found to be superior (McCallum and Nigam, 1998).

However, when there are two or more words we have to account for the pattern of their co-appearance. Although it is possible to use complex probability models (Buntine and Jakulin, 2004), we had reasonable success with the following two-attribute combination representing the words A and B:

	$B' = 1$	$B' = 0$
$A' = 1$	A appears along with B; B appears along with A.	A appears without B.
$A' = 0$	B appears without A.	The document contains neither A nor B.

Except with $A' = B' = 0$, a single document generates a number of instances that corresponds to the sum of occurrences of A and B. It is possible to work quite efficiently with such a model, and it is possible to extend it to several words. It was used both for

Sect. 3.3.2 and for the present one.

Dynamic Context

Let us focus on a particular word, A. It is reasonably cheap to examine all words that co-appear with A and compute the mutual information $I(A'; B')$. The number of co-occurrences can be seen as a heuristic that indicates dependence (Goldenberg and Moore, 2004), but we have to note that it can also be the co-absence of frequently appearing words or the mutual exclusion of words that yields high mutual information.

It is possible to extend the search further: is there a word C that co-appears with A, yet does not co-appear with B, even if B co-appears with A? Such triples are examples of positive 3-way interactions. This approach to search yielded the unexpected polysemous words discussed in Sect. 3.3.2. In general, we can always examine the interactions that include the context in which we are placed. Frequently used contexts justify more detailed analysis, but it would not be tractable to perform such detailed analysis beforehand.

On-Line Clustering

The agglomerative clustering approach is based on the dissimilarity matrix. But such a dissimilarity matrix cannot be tractably computed for such a tremendous number of attributes as are present in text mining. Instead of working purely with words as attributes, we can form new attributes that combine multiple words. Consider a set of words $\mathcal{S} = \{A, B, \ldots\}$. The corresponding binary attribute S' for the set is defined as:

$$S' \triangleq \begin{cases} 0 & ; \text{ if the document } \mathbf{d} \text{ does not contain any of the words in } \mathcal{S}; \\ 1 & ; \text{ if the document } \mathbf{d} \text{ at least one of the words in } \mathcal{S}. \end{cases} \tag{6.20}$$

The appearance of any word in $\mathcal{S}$ corresponds to a single event. For example, if A appeared 5 times, and B 6 times in a particular document, there would be 11 events involving $S' = 1$ if $\mathcal{S} = \{A, B\}$.

The set-based attributes thus attempt to represent the whole cluster of words. We can approximate the result of analyzing interactions between a particular word Y and a whole set of words $\mathcal{S}$, by simply examining the interaction between Y' and S'.

However, we do not usually have clusters of attributes in the data. The clustering of attributes needs to be performed during the analysis itself. To achieve this, we can employ the notion of ϵ-clustering: if the mutual information between a word and the set-based attribute corresponding to any of the existing clusters exceeds ϵ, we form a new cluster. On the other hand, if there is at least one set-based attribute that has a mutual information with the word greater than ϵ, we include the word in the corresponding cluster. The full algorithm is disclosed in Fig. 6.32.

The number of clusters is proportional to ϵ but is not known in advance: the higher the ϵ, the more likely it is that a word will form a new cluster. The computational complexity of the algorithm is linear in the best case (when ϵ is sufficiently high that at most k clusters will be formed) and quadratic in the worst case (when ϵ is too low, each word forms its own cluster). Therefore, the computational complexity depends on ϵ and the characteristics of the data.

Over several attempts and a constant number of clusters, we can evaluate the clustering quality with the following simple measure that deems all clusters and all words equally

$\mathcal{C} \leftarrow \emptyset$ {Initial clusters}
$\mathcal{A} \leftarrow \{A_1, A_2, \ldots, A_m\}$
while $i \in \{1, 2, \ldots, m\}$ **do**
  $\mathcal{S} \leftarrow \arg\max_{\mathcal{X} \in \mathcal{C}} I(A_i; X')$ {The most informative cluster about A_i}
  if $I(A_i; S') < \epsilon$ **then** {Not informative enough}
    $\mathcal{C} \leftarrow \mathcal{C} \cup \{\{A_i\}\}$ {Form a new cluster consisting solely of A_i}
  else {Informative enough}
    $\mathcal{S} \leftarrow \mathcal{S} \cup \{A_i\}$ {Include the word in the cluster}
  end if
end while

Figure 6.32: In the incremental ϵ-clustering algorithm, the number of clusters is not known in advance.

important:

$$q = \sum_{\mathcal{X} \in \mathcal{C}} \sum_{A \in \mathcal{X}} I(A; X') \tag{6.21}$$

A clustering thus has a higher quality if the set-based attributes have higher mutual information about individual words.

Because the formation of the clusters is sequential, different assignments to clusters may result depending on the order of introduction of the attributes. It is possible to execute the construction with several different permutations and pick the assignment to clusters that maximizes some measure of assignment utility. In a somewhat different context, this strategy was used by (Slonim et al., 2002, Peltonen et al., 2004). However, they start with the instances already being assigned to K clusters, and then they shuffle instances among clusters if this increases the utility.

In our application, we simply sorted the words from the most to the least frequent, and started forming the clusters with the most frequent words. The frequent words serve as contexts and bridges, associating the contextually related but infrequent words that do not always co-appear. The clustering of the 38800 words over 19000 documents with $\epsilon = 0.001$ takes only 2 seconds on a contemporary notebook computer, and results in the structure in Table 6.1.

This clustering of words is distinctly different from clustering of documents, which results in sets of words that are all associated with a particular topic. Instead, word clusters seem to indicate the tone of a particular word, its grammatical properties, and the characteristics of the writer. For example, clusters 0-2 capture the frequently used grammatical words. Cluster 8 captures grammatically incorrect or infrequent words. On the other hand, cluster 4 refer to financial reports, cluster 13 refers to policies, and cluster 15 to politics.

Every clustering is an approximation. However, with a sufficiently low ϵ, we can assume that there are no associations between clusters, and can perform interaction analysis purely with the words within a certain cluster, digging deeper. On the other hands, interpreting clusters as topics, we can analyze the interactions between topics, exploring the breadth. Finally, we can cluster the clusters, rising above the representation. Clusters serve as emergent levels of detail.

The clustering algorithm we have examined is not intended to replace other methods:

idx	#w	#e	typical words
0	45	568885	the (0.359), of (0.312), said (0.300), to (0.259), and (0.255)
1	307	189482	it (0.176), is (0.137), mln (0.115), will (0.115), company (0.064)
2	364	170231	its (0.126), was (0.112), an (0.106), as (0.097), not (0.094)
3	1997	139369	from (0.128), dlrs (0.118), year (0.107), pct (0.096), has (0.073)
4	564	128181	banks (0.036), japan (0.035), such (0.035), markets (0.034), most (0.032)
5	743	112457	budget (0.023), deficit (0.022), committee (0.021), way (0.020), must (0.020)
6	2404	111506	reforms (0.007), tough (0.007), increasingly (0.007), success (0.007), often (0.006)
7	990	105839	they (0.056), stock (0.053), more (0.053), than (0.050), bank (0.049)
8	20585	99502	reuter (0.462), urbaine (0.036), underwritng (0.036), suborindated (0.036)
9	1702	97534	coffee (0.009), consumers (0.009), exporters (0.009), economist (0.008)
10	646	90099	trade (0.051), we (0.049), there (0.049), under (0.046), all (0.046)
11	952	88385	price (0.040), prices (0.039), when (0.037), because (0.037), dlr (0.037)
12	1002	87310	bond (0.026), manager (0.025), selling (0.022), bonds (0.022), chief (0.021)
13	1114	84737	political (0.015), marks (0.015), mark (0.015), policies (0.014), rather (0.013)
14	1340	84643	economists (0.012), lending (0.012), china (0.012), overseas (0.011), hard (0.010)
15	2301	80157	opposition (0.011), won (0.010), motors (0.010), majority (0.009)
16	875	79894	economy (0.026), policy (0.026), many (0.025), what (0.024), demand (0.024)
17	882	77599	offer (0.035), tax (0.034), profit (0.031), general (0.029), statement (0.029)

Table 6.1: The ϵ-clustering of words in the 'Reuters' data set, using $\epsilon = 0.001$. In the listing '#w' denotes the number of words in the cluster and '#e' the number of all events the cluster generates. The listed words have the highest mutual information with the cluster itself (and is noted in the bracket behind the word), and are therefore the most representative.

we have not performed any rigorous experiments and validations. The intention is merely to show that interaction analysis can be done extremely efficiently in an incremental online fashion for very large data sets. But how to do it best remains an open problem.

CHAPTER 7

Attribute Selection and Construction

7.1 Interactions in Classification Trees and Discretization

The machine learning community has long been aware of interactions, and many methods have been developed to deal with them. There are two problems that may arise from incorrect treatment of interactions: *myopia* is the consequence of assuming that interactions do not exist, even if they do exist; *fragmentation* is the consequence of acting as if interactions existed, when they are not significant.

We will briefly survey several popular learning techniques in the light of the role they have with respect to interactions. We will show how mutual and conditional mutual information are routinely used for machine learning applications. The novel aspect of the present section is that it points at the importance of negative interactions, and that it interprets the concepts of myopia and fragmentation in terms of information theory.

7.1.1 Myopia

Greedy attribute selection and split selection heuristics are often based on various quantifications of 2-way interactions between the label Y and an attribute A. The frequently used information gain heuristic in decision tree learning is a simple example of how interaction magnitude has been used for evaluating attribute importance. With more than a single attribute, information gain is no longer a reliable measure. First, with positive interactions, such as the exclusive or problem, information gain may underestimate the actual importance of attributes, since $I(A, B; Y) > I(A; Y) + I(B; Y)$. Second, in negative interactions, information gain will overestimate the importance of attributes, because some of the information is duplicated, as can be seen from $I(A, B; D) < I(A; D) + I(B; D)$.

These problems with positive interactions are known as *myopia* (Kononenko et al., 1997). Myopic attribute selection evaluates an attribute's importance independently of other attributes, and it is unable to appreciate their synergistic effect. The inability of myopic attribute selection algorithms to appreciate interactions can be remedied with algorithms such as Relief (e.g. Kira and Rendell, 1992, Robnik-Šikonja and Kononenko, 2003), which increase the estimated quality of positively interacting attributes, and reduce the estimated worth of negatively interacting attributes.

Ignoring negative interactions may cause several problems in machine learning and statistics. We may end up with attributes providing the same information multiple times, hence biasing the predictions. For example, assume that the attribute A is a predictor of the outcome y_0, whereas the attribute B predicts the outcome y_1. If we duplicate A into another attribute A' but retain a sole copy of B, naïve Bayesian classifier trained on $\{A, A', B\}$ will be biased towards the outcome y_0. Hence, negative interactions offer opportunity for eliminating redundant attributes, even if these attributes are informative on their own. An attribute A' would then be a conditionally irrelevant source of information about the label Y given the attribute A when $I(A'; Y|A) = 0$, assuming that there are no other attributes positively interacting with the disposed attribute (Koller and Sahami, 1996). Indirectly, we could minimize the mutual information among the selected attributes, via eliminating A' if $I(A; A')$ is large (Hall, 2000). Finally, attribute weighting, either explicit (by assigning weights to attributes) or implicit (such as fitting logistic regression models or support vector machines), helps remedy some examples of negative interactions. Not all examples of negative interactions are problematic, however, since conditional independence between two attributes given the label may result in a negative interaction information among all three.

Attribute selection algorithms are not the only algorithms in machine learning that suffer from myopia. Most supervised discretization algorithms (e.g. Fayyad and Irani, 1993) are local and discretize one attribute at a time, determining the number of intervals with respect to the ability to predict the label. Such algorithms may underestimate the number of intervals for positively interacting attributes (Nguyen and Nguyen, 1998, Bay, 2001). For example, in a domain with two continuous attributes A and B, labelled with classes y_1 when $A > 0$, $B > 0$ or $A < 0$, $B < 0$, and with class y_0 when $A > 0$, $B < 0$ or $A < 0$, $B > 0$ (the continuous version of the binary exclusive or problem), all univariate splits are uninformative. On the other hand, for negatively interacting attributes, the total number of intervals may be larger than necessary, causing fragmentation of the data. Hence, in case of positive and negative interactions, multivariate or global discretization algorithms may be preferred.

7.1.2 Fragmentation

To both take advantage of synergies and prevent redundancies, we may use a different set of more powerful methods. We may assume dependencies between attributes by employing dependence modelling (Kononenko, 1991, Friedman et al., 1997), create new attributes with structured induction methods (Shapiro, 1987, Pazzani, 1996, Zupan et al., 1999), or create new classes via class decomposition (Vilalta and Rish, 2003). The most frequently used methods, however, are the classification tree and rule induction algorithms. In fact, classification trees were originally designed also for detection of interactions among attributes in data: one of the first classification tree induction systems was named Automatic Interaction Detector (AID) (Morgan and Sonquist, 1963).

Classification trees are an incremental approach to modelling the joint probability distribution $P(Y|A, B, C)$. The information gain split selection heuristic (e.g. Quinlan, 1986) seeks the attribute A with the highest mutual information with the label Y: $A = \arg\max_X I(Y; X)$. In the second step, we pursue the attribute B, which will maximize the mutual information with the label Y, but in the context of the attribute A selected earlier: $B = \arg\max_X I(Y; X|A)$.

In case of negative interactions between A and B, the classification tree learning method will correctly reduce B's usefulness in the context of A, because $I(B;Y|A) < I(B;Y)$. If A and B interact positively, B and Y will have a larger amount of mutual information in the context of A than otherwise, $I(B;Y|A) > I(B;Y)$. Classification trees enable proper treatment of positive interactions between the currently evaluated attribute and the other attributes already in the context. However, if the other positively interacting attribute has not been included in the tree already, then this positive 2-way interaction may be overlooked. To assure that positive interactions are not omitted, we may construct the classification tree with look-ahead (Norton, 1989, Ragavan and Rendell, 1993), or we may seek interactions directly (Pérez, 1997).

Instead of performing a full search with look-ahead, Esmeir and Markovitch (2004) simply evaluate tuples of attributes for split selection in classification tree induction. For example, the attribute A from the set of attributes outside the model $\mathcal{A}$ should be chosen that results in the greatest improvement in some context of k attributes from $\mathcal{A}$:

$$\arg\max_{A\in\mathcal{A}} \left(\max_{\mathcal{C}\subseteq\mathcal{A},|\mathcal{C}|\leq k} I(A;Y|\mathcal{C},X) \right) \tag{7.1}$$

The lookahead is of size k in this case, while X is the current context of the model, e.g., the nodes in the classification tree that are above the current position. This procedure will assure that those attributes that are involved in positive l-way interactions with the label, $l \leq k+2$, will enter the model quickly, even if they provide no information on their own.

The classification tree learning approach does handle interactions, but it is not able to take all the advantage of mutually and conditionally independent attributes. Assuming dependence increases the complexity of the model because the dimensionality of the probability distributions estimated from the data is increased. A consequence of this is known as *fragmentation* (Vilalta et al., 1997), because the available mutual information between an attribute B and the label Y is not assessed on all the data, but merely on fragments of it. Fragmenting is harmful if the context A is independent of the interaction between B and Y. For example, if $I(B;Y) = I(B;Y|A)$, the information provided by B about Y should be gathered from all the instances, and not separately in each subgroup of instances with a particular value of the attribute A. This is especially important when the training data is scarce. Although we used classification trees as an example of a model that may induce fragmentation, other methods too are subject to fragmentation by assuming dependence unnecessarily.

Three approaches may be used to remedy fragmentation. One approach is based on ensembles: aggregations of simpler trees, each specializing in a specific interaction. For example, random forests (Breiman, 1999) aggregate the votes from a large number of small trees, where each tree can be imagined to be focusing on a single interaction. One can use hybrid methods that employ both classification trees and linear models that assume conditional independence, such as the naïve Bayesian classifier (Kononenko et al., 1988, Kohavi, 1996) or logistic regression (Abu-Hanna and de Keizer, 2003, Landwehr et al., 2003). Finally, feature construction algorithms may be employed in the context of classification tree induction (e.g. Pagallo and Haussler, 1990, Setiono and Liu, 1998).

7.2 Interactions in Naïve Bayesian Classification Models

The *naïve Bayesian classifier* (NBC) in its general form is used to predict the conditional probability of the label Y given the values of the attributes $A_1, A_2, \ldots, A_k$. It is assumed that the attributes are conditionally independent given the label:

$$P(A_1, A_2, \ldots, A_k|Y) = \prod_{i=1}^{k} P(A_i|Y) \tag{7.2}$$

This assumption is frequently violated, and several approaches attempt to improve the prediction performance by joining the attributes that are not conditionally independent.

From the conditional independence assumption (7.2), we can develop the conditional probability model using the Bayes rule:

$$P(Y|A_1, A_2, \ldots, A_k) = P(Y)\frac{\prod_{i=1}^{k} P(A_i|Y)}{P(A_1, A_2, \ldots, A_k)} \tag{7.3}$$

Because $P(A_1, A_2, \ldots, A_k)$ is not modelled, we may assume that it is constant for all values of Y and integrate it out. The naïve Bayesian classifier does have the 'Bayesian' in it, but it does not involve priors as described in Sect. 2.2.4, the epitome of Bayesian statistics. NBC is normally used in a perfectly non-Bayesian context.

The NBC is a special case of a more general class of Bayesian networks (Pearl, 1988). It is possible to use Bayesian networks structured similarly as the NBC, but solving certain problems of the conditional independence assumption. Of course, the structure has to satisfy the demand that the Bayesian network must be a directed acyclic graph. Whenever two attributes are no longer conditionally independent given the label, we can say that the model accounts for a particular dependence. There are two specific ways of extending the naïve Bayesian model, illustrated in Fig. 7.1. Furthermore, as shown in Fig. 7.2, latent attribute models and tree-augmented naïve Bayesian classifiers cannot be reduced to one another (Ling and Zhang, 2002): the TAN cannot model the 3-parity problem, while latent attributes cannot perform a non-disjunct decomposition into two groups of attributes, but require a full collapse into a single group.

The key problem for machine learning is how to identify these dependencies. There are three main directions of approaching this problem:

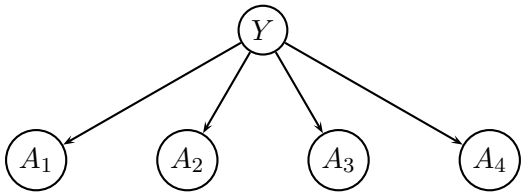
- Algorithmic identification of the dependence structure using heuristics (Chow and Liu, 1968).
- Utility-driven search in the model space (Pazzani, 1996).
- Constructive induction of latent attributes that capture the dependencies (Monti and Cooper, 1999).

We will now investigate these approaches, and report on our experiments that employed interaction information as a heuristic.

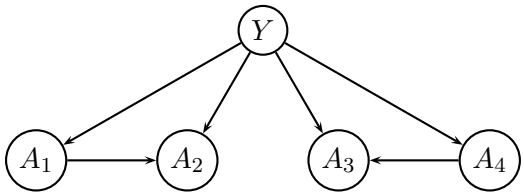
7.2.1 Heuristic Models

The importance of tree-structured models lies in the ability to reconstruct the joint model in closed form using the chain rule. For example, the expansion from Fig. 7.2(b) would

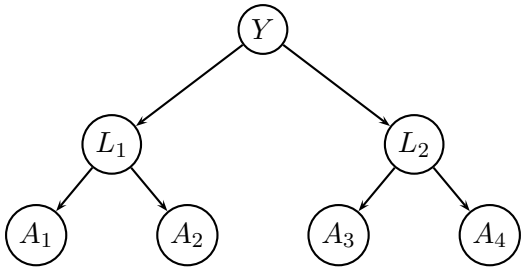
(a) Basic Naïve Bayesian Classifier
$\{(Y), (A_1|Y), (A_2|Y), (A_3|Y), (A_4|Y)\}$



(b) Tree-Augmented Naïve Bayes (TAN)
$\{(Y), (A_1|Y), (A_2|Y, A_1), (A_3|Y, A_4), (A_4|Y)\}$



(c) Naïve Bayes with Merged Attributes (ENB)
$\{(Y), (L_1|Y), (L_2|Y), (A_1|L_1), (A_2|L_1), (A_3|L_2), (A_4|L_2)\}$



(d) Naïve Bayes Augmented with a Finite Mixture Model (FAN)
$\{(Y), (L), (A_1|Y, L), (A_2|Y, L), (A_3|Y, L), (A_4|Y, L)\}$

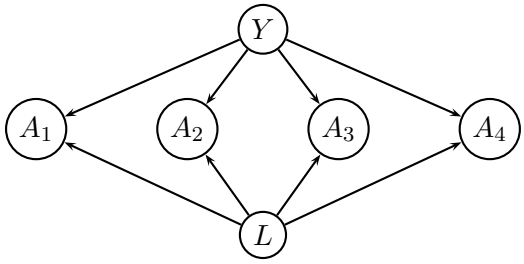


Figure 7.1: The naïve Bayesian classifier and its generalizations. All are special cases of Bayesian networks, but intended for classification. The illustrated TAN and ENB models can be equivalent. Theoretically, FAN can capture any pattern of dependence, but might not be the most efficient approach.

be simply:

$$P(Y|A_1, A_2, A_3) \propto P(Y)P(A_1|Y)P(A_2|Y, A_1)P(A_3|Y, A_1) \tag{7.4}$$

The Chow-Liu algorithm (Chow and Liu, 1968) is a fast algorithm for finding the maximum likelihood tree-structured joint model for a given joint probability model $P(A_1, A_2, \ldots, A_k)$. The algorithm first computes mutual information $I(A_i; A_j)$ for each pair of attributes, and then builds the maximum spanning tree with Kruskal's algorithm. Effectively, the pairs of attributes are sorted by their mutual information, and consecutively introduced into the model as arcs. A pair is skipped if it would result in a cycle. Because the arcs are directed in a Bayesian network, we can pick any attribute and make sure that all the arcs are directed outwards as to prevent cycles.

However, using the Chow-Liu algorithm would have the same problem as other Bayesian network structure learning algorithms: the maximum likelihood criterion does not seek to improve the label prediction performance. Friedman et al. (1997) combined the Chow-Liu algorithm with the naïve Bayesian classifier. Instead of using the mutual information Friedman et al. employ the conditional mutual information $I(A_i; A_j|Y)$. Conditional mutual information is a frequently used heuristic for using maximum likelihood techniques for classification tasks. While most approaches seek to draw dependencies between whole attributes, Kononenko (1991) performs this for individual attribute values. This approach, however, will be discussed in later sections.

Bayesian Model Averaging With the ideas about model uncertainty from Sect. 2.2.5, using maximum likelihood models may be inappropriate. Because the TAN models are more complex than the original NBC ones, there is an increased likelihood of overfitting. One way of remedying overfitting is by assuming Bayesian priors and averaging over them. Friedman et al. used the Laplace probability estimate, and observed an improvement in classification performance. The Laplace probability estimate uses Bayesian inference using the uniform Dirichlet prior to obtain the conditional probabilities for each node and arc. It is possible, however, to be Bayesian over the structure, too: Meilă and Jaakkola (2000) define a tractable prior over the structures. Cerquides and López de Màntaras (2003) have developed a simplified implementation of Bayesian model averaging that resulted in improved performance on standard machine learning benchmarks.

7.2.2 Search Algorithms

As in Sect. 3.2.2, it has been observed (Domingos and Pazzani, 1997, Rish et al., 2001) that conditional mutual information is not always a reliable heuristic when we are concerned about classification accuracy. For that reason, several researchers have considered the actual classification performance as a heuristic. Unlike the methods of Sect. 7.2.1, the methods of this section do not attempt to maximize the (conditional) likelihood of the model. Instead, the methods maximize the predictive performance directly.

Pazzani (1996) formulates learning as a search in the state space of the models. There are two formulations, the *forward sequential selection and joining* (FSSJ) and the *backward sequential elimination and joining* (BSEJ). The FSSJ starts with an empty model, and can perform the following types of operations:

- Add a new attribute into the model assuming its independence of other attributes.

- Add a new attribute joined with another attribute that is already in the model.

All possible actions are analyzed using internal leave-one-out validation, and the action that resulted in the best performance is selected. The procedure continues until an improvement can be made. On the other hand, BSEJ starts with the complete model of independent attributes, and then performs the identical procedure, but with the following two kinds of operations:

- Remove an attribute from the model.
- Merge a pair of attributes already in the model.

Pazzani found that BSEJ worked somewhat better in most cases, but resulted in more complex models.

The disadvantage of both FSSJ and BSEJ is that they are restricted to joining attributes through latent attribute approach. The more general Bayesian networks can tolerate more flexible arrangements, such as the one shown in Fig. 7.2. However, Friedman et al. (1997) argued that more general Bayesian networks are often outperformed in classification tasks by the simple naïve Bayesian classifier. Namely, the structure of Bayesian networks is usually selected through their performance as generative models, not as discriminative ones. The structure learning algorithms generally do not use the conditional loss functions, such as (3.3), that are appropriate for supervised learning and for classification, but generative ones, such as (3.2).

It is simple to modify the loss function for discriminative learning. Such an optimization procedure for learning Bayesian networks intended specifically for classification was developed by Keogh and Pazzani (2002) and by Grossman and Domingos (2004). Grossman and Domingos (2004) perform a hill-climbing search with the following operations:

- Add a new arc into the model.
- Remove an existing arc from the model.
- Reverse an existing arc in the model.

The search space is quite large and widely branching, and various simplifications were required to keep it tractable. They also found that it was useful to restrict the complexity of Bayesian networks because high-order dependencies resulted in unreliable performance: the best results were achieved by restricting the maximum number of parents of each node to 2. Unlike other researchers, Grossman and Domingos have used p-loss functions, not just classification accuracy.

Keogh and Pazzani (2002) investigated a restricted form of the above search space: arcs may only be added, not removed and not reversed. Even so, their algorithm seems to quite consistently outperform the procedure of Friedman et al. (1997). In order to reduce the learning time, the introduced a more specialized algorithm, *SuperParent*, which restricts the above search space into a unidimensional sequence of a single operation:

1. Find the best 'super parent' attribute P that improves the performance the most by putting all unconnected attributes into the context of P by creating an arc from P to each of them.

2. Among all unconnected attributes, find the 'favorite child' attribute A that improved the model's classification performance the most in the context of P. If the improvement is positive, the arc $P \rightarrow A$ is made permanent in the model.

Thereby, the single operation made by SuperParent is: find the super parent, and connect it to its favorite child.

These algorithms are all greedy search algorithms. As such, the algorithms tend to be slow, require internal cross-validation or some other kind of regularization to prevent overfitting, and can be improved by non-greedy search methods. But because they directly seek to optimize the same criterion that is used in the final evaluation, they very often perform distinctly better than any of the alternatives.

7.2.3 Latent Attribute Induction Algorithms

Dependencies between pairs of attributes capture isolated sources of correlation. In many data sets, however, there may be a global source of variation. The idea of latent attribute induction is to get rid of dependencies by postulating a global source of variation represented with a hidden attribute L that influences both labelled and unlabelled attributes. The approach was applied to classification by Monti and Cooper (1999), who have experimented with two variants: in FM, they performed the classification purely with a mixture model of the type $P(L)P(Y|L)\prod_i P(X_i|L)$. In FAN, they augmented the naïve Bayesian classifier with the mixture model, as in Fig. 7.1(d). They find that FAN is generally better for classification, whereas FM is better for estimation of the label's probability.

There is no need for the hidden attribute L to be relevant to the class. This has been addressed by Vilalta and Rish (2003), who infer a separate L for each label. The resulting model can be seen as being of the form $P(Y)P(L|Y)\prod_i P(X_i|L)$ and is shown in Fig. 7.3. In some cases, a single latent attribute is inappropriate. Zhang et al. (2004) have generalized the above approaches so that a hierarchy of latent attributes is inferred, quite similar to the deterministic scheme of Zupan et al. (1999).

A somewhat different approach to decomposing than latent attributes are *Bayesian multinets* (Geiger and Heckerman, 1996, Peña et al., 2002). A practical application of multinets would be to create a separate TAN for each class, as the pattern of dependencies may differ for each class. The conditional mutual information $I(A_i; A_j|Y)$ is actually the average over all values of Y:

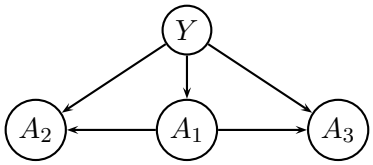
$$I(A_i; A_j|Y) = \sum_{y \in \Re_Y} P(y) I(A_i; A_j|Y = y) \tag{7.5}$$

For that reason, a separate TAN can be constructed for each labelled attribute value y, using the conditional mutual information in the context of that particular value.

7.3 Case Studies of Interactions and the NBC

We will now examine how interaction information can help us understand the reasons for low or high naïve Bayesian classifier performance. In Sect. 7.3.1 we will examine whether interaction information is a useful heuristic for merging the attributes over a number of data sets, repeating some material from (Jakulin and Bratko, 2003). In Sect. 7.3.4 will show that it is often the negative interaction information and not overfitting that explains

(a) A Tree-Augmented Naïve Bayes Model
that cannot be represented with the merged attributes.
$\{(Y), (A_1|Y), (A_2|Y, A_1), (A_3|Y, A_1)\}$



(b) A Naïve Bayes Model with Merged Attributes
that cannot be represented with the TAN model.
$\{(Y), (L|Y), (A_1|L), (A_2|L), (A_3|L), (A_4|Y)\}$

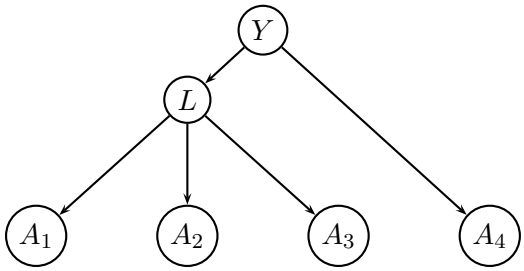


Figure 7.2: The TAN and ENB models are not subsets of one another.

Class-Decomposed Naïve Bayesian Classifier
$\{(Y), (L|Y), (A_1|L), (A_2|L), (A_3|L), (A_4|L)\}$

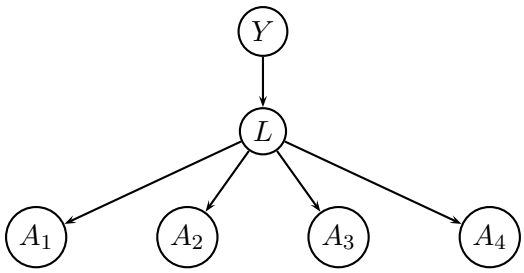


Figure 7.3: Class decomposition extends the range of the labelled attribute by creating pseudo-classes.

the benefits of attribute selection for the naïve Bayesian classifier. In Sect. 7.3.3 we will study an application of interaction information to guiding attribute construction, using some material from (Jakulin et al., 2003). In Sect. 7.3.4 we will show that certain problems with interactions are noticeable already on the training set, and they are better referred to as approximation than generalization errors. Finally, we will present a context-dependent attribute selection heuristic based on interaction information that successfully compares with exhaustive greedy search.

7.3.1 Positive and Negative Interactions

In each experimental data set, we select the most interacting pair of attributes according to (a) conditional mutual information (CI), (b) positive interaction information (PIG), and (c) negative interaction information (NIG). We build naïve Bayesian classifiers (NBC) in which the selected interactions are "resolved." That is, the selected pair of most interacting attributes is replaced in NBC by its Cartesian product. This interaction resolution is done for the result of each of the three interaction detection heuristics (CI, PIG and NIG), and the performance of the three resulting classifiers is compared.

We chose to measure the performance of a classifier with Brier loss (described below). We avoided classification accuracy as a performance measure for the following reasons. Classification accuracy is not very sensitive in the context of probabilistic classification: it usually does not matter for classification accuracy whether a classifier predicted the true label with the probability of 1 or with the probability of, e.g., 0.51. To account for the precision of probabilistic predictions, we employed the *Brier loss*. Given two probability distributions, the predicted label probability distribution $\hat{P}$, and the actual label probability distribution P, the (half-)Brier loss of the prediction is (Brier, 1950):

$$b(P, \hat{P}) \triangleq \frac{1}{2} \sum_{y \in \Re_Y} \left(P(y) - \hat{P}(y) \right)^2 \tag{7.6}$$

Brier loss is usually referred to as Brier score. Since the prediction is getting worse the increasing Brier loss, it would be misleading to refer to it as 'score', a word that carries a positive connotation. Error rate is a special case of Brier loss for deterministic classifiers, while Brier loss could additionally reward a probabilistic classifier for better estimating the probability. In a practical evaluation of a classifier given a particular testing instance, we approximate the actual class distribution by assigning a probability of 1 to the true class of the testing instance. For multiple testing instances, we compute the average Brier loss.

As the basic learning algorithm, we have used the naïve Bayesian classifier. After the most important interaction was determined outside the context of other attributes, we modified the NBC model created with all the domain's attributes by taking the single most interacting pair of attributes and replacing them with their Cartesian product, thus eliminating that particular dependence. All the numerical attributes in the domains were discretized beforehand, and missing values represented as special values. Evaluations of the default NBC model and of its modifications with different guiding heuristics were performed with 10-fold cross-validation. For each fold, we computed the average loss. For each domain, we computed the loss mean and the standard error over the 10 folds. We performed all our experiments with the Orange toolkit (Demšar and Zupan, 2004).

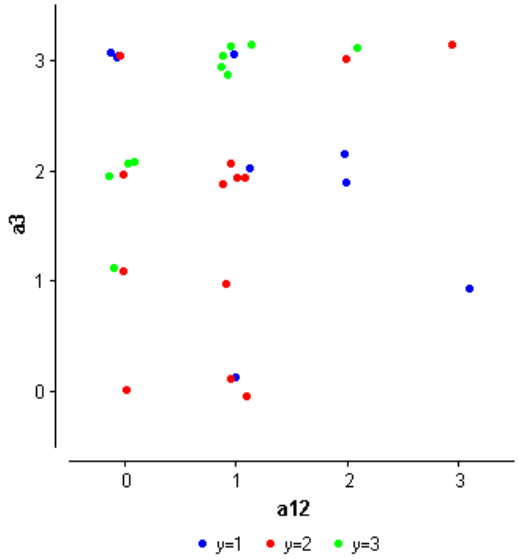


Figure 7.4: The 3-way interaction between $A12$, $A3$ and the label y in the 'lung' data set yields the P-value of 0.023. Cross-validation also deems this interaction as significant. However, there are very many degrees of freedom involved, and it might be unclear whether this example counts as an interaction or not.

In Table 7.1 we sorted 26 of the UCI KDD archive (Hettich and Bay, 1999) domains according to the number of instances in the domain, from the smallest on the top to the largest on the bottom, along with the results obtained in the above manner. The most distinct observation can be made is the bad performance of conditional mutual information. Only in one data set (adult) joining the pair of attributes with maximum conditional mutual information gave the best result. But even in that data set, the improvement was not significantly better than the original. On the other hand, CI resulted in significantly *worse* performance in 14 out of 26 experiments.

Positive interactions or synergies are frequent in artificial data sets (monk1, monk2, KRKP), when they are often introduced intentionally to test the myopia of a learning algorithm. Somewhat surprisingly, they are also frequent in small data sets (lung, soy-small, wine): it is unclear whether their existence here truly indicates significant interactions, or merely demonstrates the deficiencies of using cross-validation to penalize overfitting on small data sets. For example, the two interacting attributes in the lung-cancer data set, $A12$ and $A3$, each have four values, and the label three. The conditional probability model under the assumption of dependence thus estimates $4 \times 4 \times 3 = 48$ different probability values. At the same time, there are only 32 instances, shown in Fig. 7.4. It is known that statistical methods for determining the goodness-of-fit are unreliable with so many degrees of freedom and so few instances (Agresti, 2002), so the cross-validation result may be misleading.

We can summarize the findings as:

- Conditional mutual information is not a good heuristic for determining whether two attributes should be assumed as dependent in classification with the naïve Bayesian classifier.

	Brier Loss			
	NBC	PIG	NIG	CI
lung	0.382 ± 0.045	**0.313 ± 0.049**	0.380 ± 0.046	0.382 ± 0.045
soy-small	**0.032 ± 0.013**	**0.032 ± 0.013**	**0.032 ± 0.013**	**0.032 ± 0.013**
zoo	0.062 ± 0.012 √	**0.061 ± 0.013**	0.062 ± 0.013 √	0.069 ± 0.015 √
lymphography	**0.181 ± 0.020**	0.207 ± 0.027	0.182 ± 0.021 √	0.198 ± 0.029 √
wine	0.015 ± 0.006 √	**0.014 ± 0.006**	0.020 ± 0.008 √	0.024 ± 0.008
glass	**0.224 ± 0.017**	0.227 ± 0.021 √	0.239 ± 0.014 √	0.227 ± 0.023 √
breast	**0.221 ± 0.015**	0.267 ± 0.011	0.222 ± 0.023 √	0.232 ± 0.015 √
ecoli	**0.141 ± 0.013**	0.155 ± 0.015	0.233 ± 0.023	0.233 ± 0.023
horse-colic	0.222 ± 0.014 √	0.275 ± 0.021	**0.219 ± 0.013**	0.257 ± 0.013
voting	0.090 ± 0.010	0.095 ± 0.010	**0.070 ± 0.010**	0.098 ± 0.012
monk3 †	0.042 ± 0.004	**0.030 ± 0.006**	0.043 ± 0.005	**0.030 ± 0.006**
monk1 †	0.175 ± 0.008	**0.002 ± 0.000**	0.186 ± 0.011	**0.002 ± 0.000**
monk2 †	**0.229 ± 0.006**	0.254 ± 0.008	0.235 ± 0.007 √	0.254 ± 0.008
soy-large	0.080 ± 0.009 √	**0.073 ± 0.008**	0.080 ± 0.009 √	0.076 ± 0.009 √
wisc-cancer	**0.024 ± 0.004**	0.025 ± 0.005 √	0.026 ± 0.004 √	0.026 ± 0.004 √
australian	0.112 ± 0.007 √	0.123 ± 0.009	**0.111 ± 0.008**	0.112 ± 0.008 √
credit/crx	0.113 ± 0.008 √	0.127 ± 0.007	**0.110 ± 0.008**	0.125 ± 0.009
pima	**0.160 ± 0.006**	0.176 ± 0.009	0.168 ± 0.007	0.169 ± 0.006
vehicle	0.289 ± 0.010	**0.267 ± 0.010**	0.290 ± 0.007	0.290 ± 0.007
heart	**0.285 ± 0.010**	0.307 ± 0.011	0.294 ± 0.010 √	0.296 ± 0.010
german	**0.173 ± 0.007**	0.190 ± 0.006	0.179 ± 0.008 √	0.181 ± 0.009
cmc	0.297 ± 0.007 √	**0.297 ± 0.008**	0.312 ± 0.007	0.309 ± 0.007
segment †	**0.057 ± 0.004**	0.061 ± 0.004 √	0.060 ± 0.004 √	0.058 ± 0.004 √
krkp †	0.092 ± 0.004	**0.078 ± 0.003**	0.098 ± 0.005	0.110 ± 0.005
mushroom	0.002 ± 0.000	0.009 ± 0.001	**0.000 ± 0.000**	0.001 ± 0.000
adult	0.119 ± 0.002 √	0.128 ± 0.002	0.121 ± 0.002 √	**0.119 ± 0.002**

Table 7.1: A comparison of heuristics for merging attributes. The table lists Brier losses obtained with 10-fold cross validation after resolving the most important interaction, as assessed with different methods. A result is set in bold face if it is the best for the domain, and checked if it is within the standard error of the best result for the domain. We marked the artificial† domains.

- Cross-validation may not detect overfitting through joining of attributes on small data sets.
- According to empirical results in real-world domains, strong positive interactions are quite rare, while negative interactions are plentiful.
- Positive interactions often appear in artificial data sets.

7.3.2 Interactions, Classifiers and Loss Functions

We have seen earlier that interaction information can be interpreted as a KL-divergence between a model that assumes dependence and another one that does not. The independence-assuming model resembles the naïve Bayesian classifier (Sect. 3.2.2):

$$I(A;B;Y) = D\left(P(Y|A,B)\middle\|P(Y)\frac{P(A|Y)P(B|Y)}{P(A)P(B)}\right) \tag{7.7}$$

times	NBC	PIG	NIG	CI
best	11	10	6	4
good $\surd$	8	3	12	8
bad	7	13	8	14

Table 7.2: No heuristic for joining attributes is consistently better than the default naïve Bayesian classifier. However, joining through maximum conditional independence is bad.

However, our learning algorithm is naïve Bayesian classifier proper, and we use Brier loss and not KL-divergence. Therefore, it makes sense to derive a measure that will make use of the same loss function and the same classifier in the heuristic as are used for the final evaluations. Such a measure is $BS(A, B)$:

$$BS(A,B) \triangleq \sum_{a\in\Re_A, b\in\Re_B, y\in\Re_Y} P(a,b,y)\left(P(y|a,b) - \frac{P(y)P(a|y)P(b|y)}{\sum_{y'\in\Re_y} P(y')P(a|y')P(b|y')}\right)^2 \quad (7.8)$$

It computes the expected Brier loss between the true conditional probability of the label Y given the attribute values, and the naïve Bayesian prediction. The larger the value of BS, the more we gain by assuming dependence between the attributes. From the results in Table 7.3, the following conclusions can be made:

- Generally, it is slightly better for the heuristics to make use of the same utility functions and of the same classification model we use to score the final results. There are some exceptions, such as classification error, that are not desirable as they lack discrimination power and are not proper.
- Even so, our heuristic did not result in consistent improvements across the data sets. The decision about whether to join two attributes is not independent of other attributes in the model.

7.3.3 Using Interaction Information to Guide Attribute Merging

We have examined attribute interactions and the effect they have on performance of the naïve Bayesian classifier in the domain of predicting the patient's long term clinical status after hip arthroplasty. The data we have considered was gathered at Department of Traumatology of University Clinical Center in Ljubljana from January 1988 to December 1996. For each of the 112 patients, 28 attributes were observed at the time of or immediately after the operation. All attributes are nominal and most, but not all, are binary (e.g., presence or absence of a complication). Patient's long-term clinical status was assessed in terms of Harris hip score (Harris, 1969) at least 18 months after the operation. Harris hip score gives an overall assessment of the patient's condition and is evaluated by a physician who considers, for example, patient's ability to walk and climb stairs, patient's overall mobility and activity, presence of pain, etc. The numerical Harris hip score in scale from 0 to 100 was discretized into three classes: *bad* (up to 70, for 43 patients), *good* (between 70 and 90, for 34 patients) and *excellent* (above 90, for 35 patients).

	Brier Loss		
	NB	PIG	BS
lung	0.382 ± 0.045	**0.313 ± 0.049**	0.375 ± 0.056
soy-small	**0.032 ± 0.013**	**0.032 ± 0.013**	**0.032 ± 0.013**
zoo	0.062 ± 0.012 √	**0.061 ± 0.013**	0.061 ± 0.013 √
lymphography	0.181 ± 0.020 √	0.207 ± 0.027	**0.175 ± 0.022**
wine	0.015 ± 0.006 √	**0.014 ± 0.006**	0.021 ± 0.007
glass	**0.224 ± 0.017**	0.227 ± 0.021 √	0.242 ± 0.017 √
breast	**0.221 ± 0.015**	0.267 ± 0.011	0.246 ± 0.020
ecoli	**0.141 ± 0.013**	0.155 ± 0.015	0.300 ± 0.022
horse-colic	**0.222 ± 0.014**	0.275 ± 0.021	0.230 ± 0.014 √
voting	0.090 ± 0.010	0.095 ± 0.010	**0.057 ± 0.014**
monk3	0.042 ± 0.004	**0.030 ± 0.006**	**0.030 ± 0.006**
monk1	0.175 ± 0.008	**0.002 ± 0.000**	**0.002 ± 0.000**
monk2	**0.229 ± 0.006**	0.254 ± 0.008	0.256 ± 0.008
soy-large	0.080 ± 0.009 √	**0.073 ± 0.008**	0.081 ± 0.010
wisc-cancer	**0.024 ± 0.004**	0.025 ± 0.005 √	0.025 ± 0.004 √
australian	0.112 ± 0.007 √	0.123 ± 0.009	**0.107 ± 0.009**
credit	0.113 ± 0.008 √	0.127 ± 0.007	**0.106 ± 0.008**
pima	**0.160 ± 0.006**	0.176 ± 0.009	0.174 ± 0.007
vehicle	0.289 ± 0.010	0.267 ± 0.010	**0.254 ± 0.010**
heart	**0.285 ± 0.010**	0.307 ± 0.011	0.303 ± 0.009
german	**0.173 ± 0.007**	0.190 ± 0.006	0.191 ± 0.008
cmc	0.297 ± 0.007 √	**0.297 ± 0.008**	0.314 ± 0.008
segment	0.057 ± 0.004	0.061 ± 0.004	**0.051 ± 0.004**
krkp	0.092 ± 0.004	**0.078 ± 0.003**	0.080 ± 0.004 √
mushroom	0.002 ± 0.000	0.009 ± 0.001	**0.000 ± 0.000**
adult	**0.119 ± 0.002**	0.128 ± 0.002	0.133 ± 0.002

times	**NBC**	**PIG**	**BS**
best	11	9	10
good √	7	2	5
bad	8	15	11

Table 7.3: Generally, heuristics aligned with the method and the loss function (BS) are slightly better than generic ones (PIG).

In our experimental evaluation, interaction information scores were obtained from considering the complete data set, new attributes were created and added into the data set. Here is the outline of the particular implementation of this procedure as used in the test case:

1. Consider all pairs of attributes and estimate their interaction information with the label.

2. Select N attribute pairs with the highest interaction information, and for each pair:

   - Construct a new joined attribute, such that each distinct pair of values of original attributes maps to a distinct value of a new attribute.
   - Add the constructed attribute to the attribute set.

3. Obtain a new data set with the original and the N new attributes. Order these attributes by mutual information, and build a classifier from the n best ranked attributes, $n = 1, 2, 3, \ldots$.

In the second phase, the naïve Bayesian classifier was built using the altered data set and evaluated at different sizes of the selected attribute subset. The ordering of the attributes for attribute subset selection using information gain and modelling using the subset were both performed on the learning data set, but evaluated on the test set. The evaluation was performed using the leave-one-out scheme: for the data set containing l instances, we performed l iterations, $j = 1, 2, \ldots, l$, in which all instances except j-th were used for training, and the resulting predictive model was tested on the j-th instance. We report average performance statistics over all l iterations. All the experiments were performed with the Orange toolkit (Demšar and Zupan, 2004). To measure the performance of classification models we have used two error measures, error rate and Brier loss. Brier loss has recently gained attention in medicine (Margolis et al., 1998), because it is better suited for evaluating probabilistic classifiers.

We have assessed how the inclusion of different number of newly constructed and original attributes affects the prediction performance. Figure 7.5 illustrates the search space for our domain, where the number n of attributes selected is plotted on the horizontal and the number N of interactions resolved on the vertical axis. The best choice of n and N can be determined with a wrapper mechanism for model selection. We can observe several phenomena: increasing the number of attributes in the attribute subset does not increase the error rate as much as it hurts the precision of probability estimates, as measured by the Brier loss. Furthermore, there are diminishing returns to resolving an increasing number of interactions, as illustrated in the contour diagrams in Fig. 7.5. Unnecessary interactions merely burden the attribute subset selection mechanisms with additional negative interactions. Figure 7.6 presents the results in terms of Brier loss and error rate with four resolved interactions.

There are several islands of improved predictive accuracy, but the best appears to be the area with approximately 4 resolved interactions and 4 selected attributes. Classification accuracy reaches its peak of 60% at the same number of attributes used. This accuracy improves upon the accuracy of 56% obtained in our previous study, where manually crafted attributes as proposed by domain experts were used in the naïve Bayesian classifier (Zupan et al., 2001). Both are a substantial improvement over models constructed from the

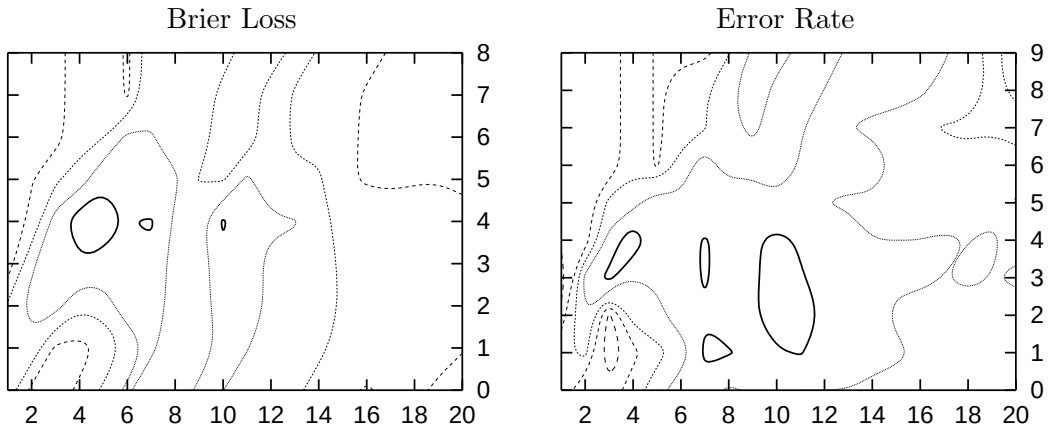


Figure 7.5: Dependence of the Brier loss and error rate on the attribute subset size, n (horizontal axis) and on the number of interactions resolved, N (vertical axis). Emphasized are the areas of the best predictive accuracy, where Brier loss is less than 0.3 and the error rate less than 0.45.

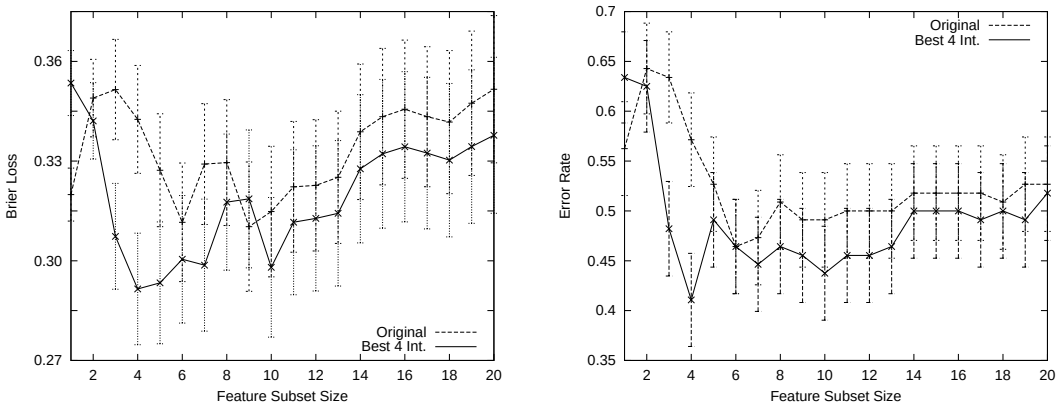
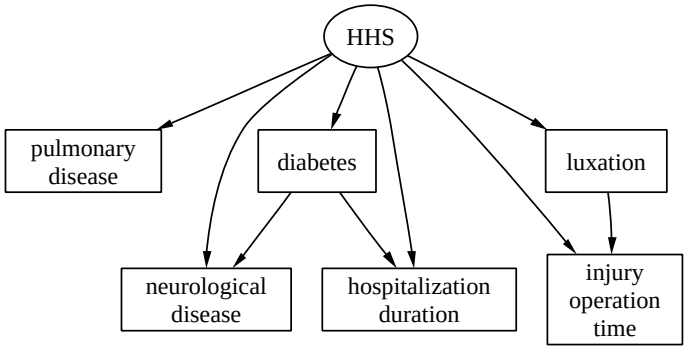


Figure 7.6: Average Brier loss and error rate as computed by leave-one-out and its dependence on the number of attributes used in the model for $N = 4$ (solid line) and $N = 0$ (dashed). For all measurements, the standard error is shown.

original set of attributes, where the accuracy of NBC with the original 28 attributes is 45%, and does not rise beyond 54% even with use of attribute subset selection. The TAN model on all attributes gets 55% accuracy. Attribute selection helps TAN models too, as the TAN on the 9 best attributes achieved 57% accuracy. The general Bayesian network model obtained via B-Course achieved only 47% accuracy, only slightly better than the naïve Bayes. Overall, our simple method proved to be quite competitive.

The results in Fig. 7.7 show that three of the four constructed attributes were chosen in building of the model. The table provides the set of important interactions in the data, where an important increase in predictive accuracy can be seen as an assessment of the interaction importance itself, given the data. Because the attribute 'diabetes' appears twice among the merged attributes, we can avoid double-counting the attribute by using a Bayesian network as shown at the top of Fig. 7.7. The model can be easily interpreted using the concepts of Sect. 3.3.2. The 'diabetes' attribute *moderates* the influence



Mutual Information	Attribute
0.118	**luxation + injury_operation_time**
0.116	**diabetes + neurological_disease**
0.109	**hospitalization_duration + diabetes**
0.094	pulmonary_disease

Figure 7.7: The best predictive model for the HHS dataset for the case $N = 4, n = 4$ is shown as a Bayesian network (above). We also list the average mutual information between the label and an attribute or a pair of them in the table (below). The attribute pairs are typeset in bold.

of neurological disease and hospitalization duration on the outcome. This means that the influence of the neurological disease and of hospitalization duration on the outcome should be considered separately for diabetics and non-diabetics. For example, there were more bad outcomes among diabetics with a long hospitalization time. There is another moderation effect between luxation and injury operation time. Similarly, the influence of injury operation time should be considered separately, depending on whether there was a luxation or not. Unfortunately, there were relatively few luxations and relatively few diabetics in our sample, and it is difficult to remove the possible influence of chance. Hence, our findings about moderation effects are not conclusive. It known that mediation and moderation effects are more difficult to confirm (McClelland and Judd, 1993).

We have compared the results obtained with the greedy method with global search-based attribute subset selection as implemented by B-Course (Myllymaki et al., 2002). The model without interactions achieved classification accuracy of 59% and Brier loss of 0.30 with 7 selected attributes. If the 10 interactions with the highest interaction information were added, the model achieved classification accuracy of 62% and Brier loss of 0.28 with a model consisting of 8 attributes. B-Course's model is quite complex and includes all the attributes from Table 7.7, in addition to two of the original attributes and two interactions.

In summary, we have used interaction information to identify a small group of potential interactions, which restricts the space of possible models. These interactions are represented as ordinary attributes. We then perform attribute selection to come up with the model. The resulting model can usually be expressed as a Bayesian network with-

out latent attributes. Greedy attribute selection yields much smaller models with slightly worse performance than does search-based attribute selection using B-course. Of course, this case study is not systematic in terms of scientific conclusions.

7.3.4 Approximation and Generalization Error

The conclusion to be taken from the previous section is that heuristics are not generally reliable. Furthermore, there are many dependencies between attributes: the importance of assuming the dependence between any pair of attributes cannot be decided without knowing what other attributes are already in the model.

What some may find somewhat surprising is that the problems associated with dependencies can be identified already on the training data. With the naïve Bayesian classifier, adding an attribute that is perfectly useful on its own may deteriorate the results because of overcounting of evidence (Langley and Sage, 1994). For example, if attributes are sorted by their information gain, and introduced into the naïve Bayesian classification model one by one, the performance already on the training set is not monotonically decreasing, as shown in Fig. 7.8. On the other hand, logistic regression generally does improve almost monotonically with the addition of new attributes, as shown in Fig. 7.9.

Therefore, an introduction of an attribute into a model may result in *approximation error*. Approximation error is noticeable already on the training data. On the other hand, *generalization error* results from the difference between the training and the test set. It is quite well-known that logistic regression can overfit. However, with attribute selection the naïve Bayesian classifier can also overfit: we can build very many functions if we perform a selection among 10000 randomly tossed coins, yet none of them is going to work in future.

7.3.5 Contextual Attribute Selection in Naïve Bayes

The main realization we can draw from Sect. 7.3.4 is that adding an attribute can corrupt the naïve Bayesian classifiers' performance. The cause of the loss is the disagreement with the assumptions, the approximation error. The deterioration in performance is already noticeable on the training data. This contrasts with classification trees that tend to uniformly gain in performance with increasing complexity on the training data.

We can follow the approach of sequential Bayes (Michie and Al Attar, 1991) in picking the best attribute at each step *in the context of the attributes already in the model.* This kind of attribute selection is contextual: we are looking for the attribute that will contribute the most in the company of other attributes. Only the first attribute is chosen by its solo performance. Formally, if $D(P\|Q)$ is our loss function, Y is our label, and $\mathcal{A}$ the existing set of attributes in the model, we should pick the attribute X among the attributes $\bar{\mathcal{A}}$ outside the model that results minimum expected loss:

$$\arg\min_{X\in\bar{\mathcal{A}}} D\left(P(Y|\mathcal{A},\bar{\mathcal{A}})\middle\|\frac{1}{Z}P(Y)P(X|Y)\prod_{A\in\mathcal{A}}P(A|Y)\right) \tag{7.9}$$

Here, Z is a normalization constant, needed to assure that the probabilistic predictions of the naïve Bayesian classifier sum up to 1. Because $P(Y|\mathcal{A},\bar{\mathcal{A}})$ is unknown, we approximate it with the data set, by examining the distribution of the values of Y at each possible combination of the values of the attributes in $\mathcal{A}\cup\bar{\mathcal{A}}$.

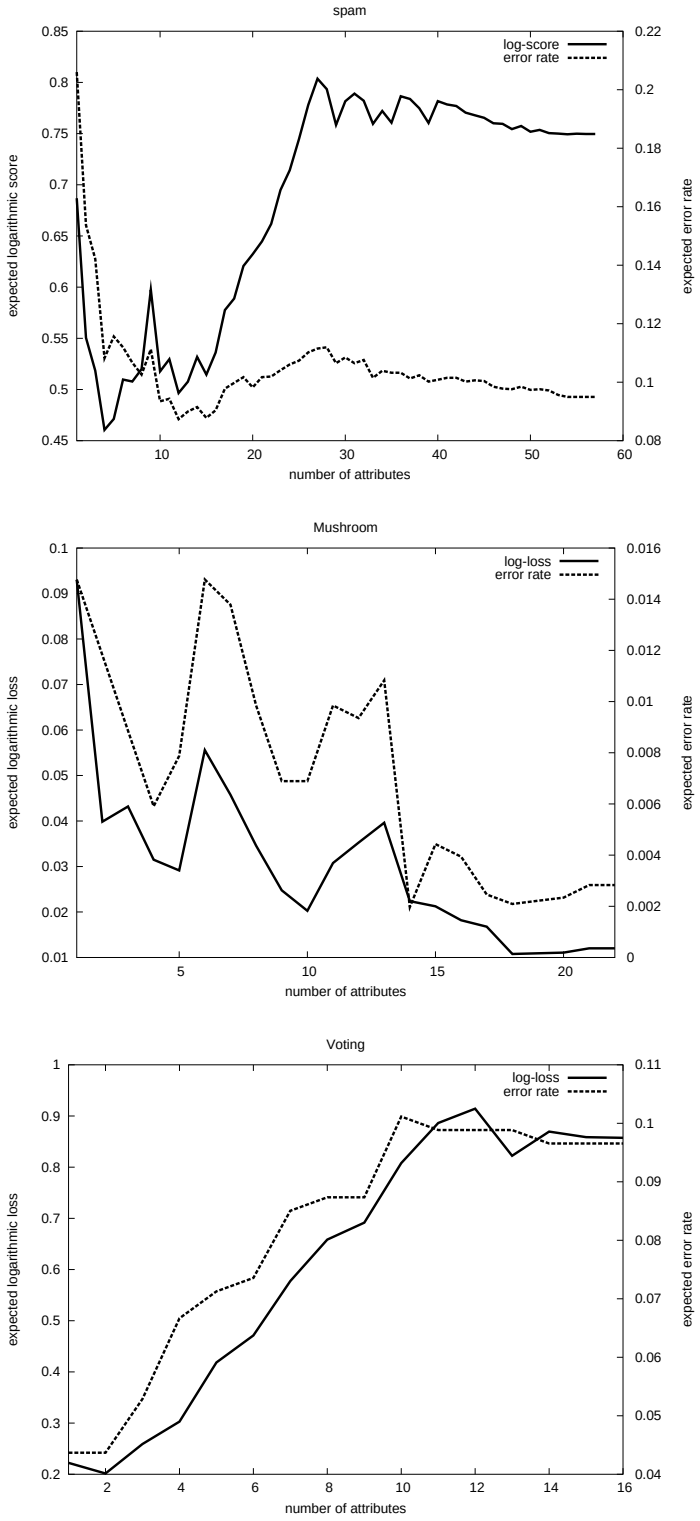


Figure 7.8: Naïve Bayesian classifier's performance loss on the training data does not improve monotonically with the addition of new attributes.

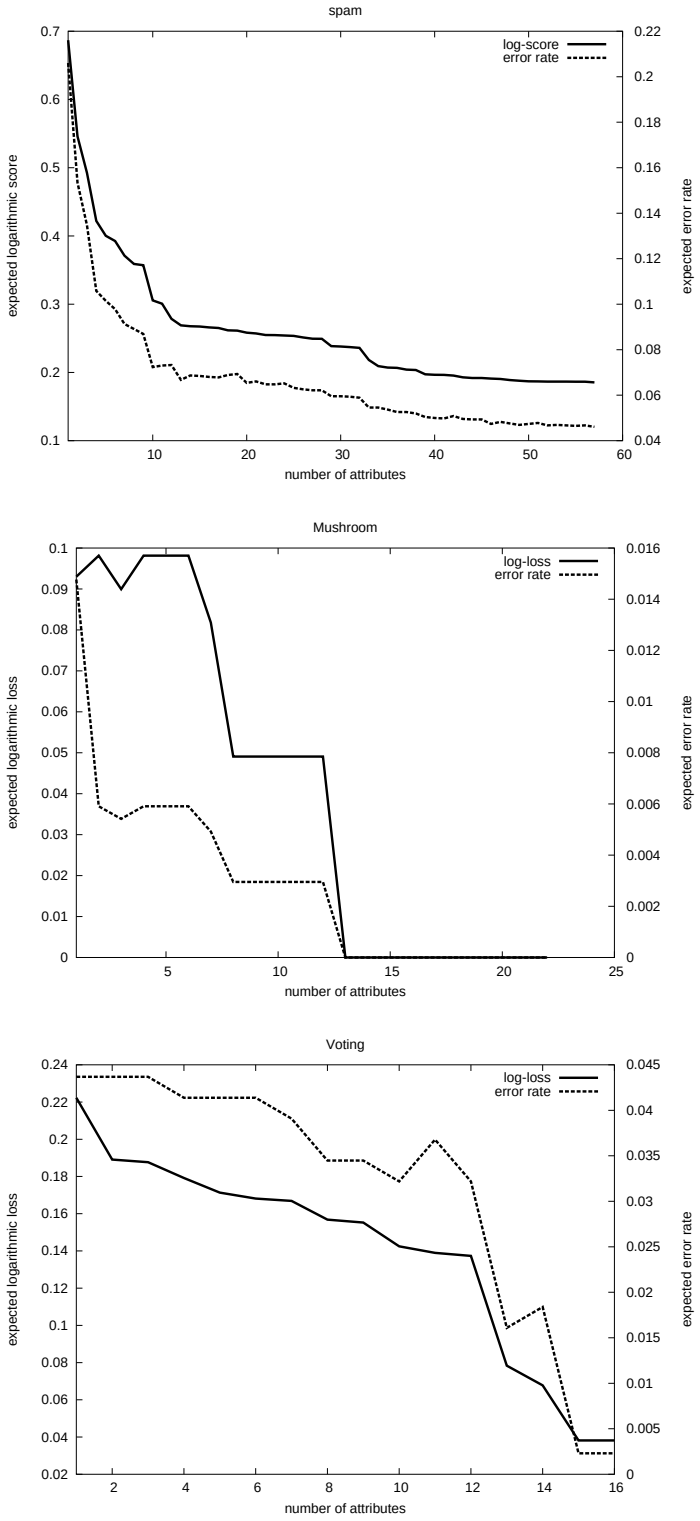


Figure 7.9: The loss of logistic regression, on the other hand, does improve almost monotonically. However, we must account for a larger amount of overfitting.

An interesting information-theoretic heuristic is described by Fleuret (2004), where a single attribute already in the model is used as the context for a candidate attribute. The basic idea is to consider each candidate X among the attributes outside the model. Attribute X alone provides $I(X;Y)$ bits of information about the label Y. When we have the context attributes $\mathcal{A}$, however, it would seem that $I(X;Y|\mathcal{A})$ would be a good heuristic. This is not the case, as conditional mutual information is never negative, but above we saw that the contribution of an attribute may be indeed negative. The problem is that mutual information refers to the 'true' probability model, while the naïve Bayesian classifier uses a simplified approximation to this unknown 'true' model. Fleuret (2004) suggests that we should find a context attribute A in $\mathcal{A}$ with the most distinct negative interaction with the candidate X. Such attribute A is the worst possible context for X, and the information gain of X should be evaluated in the context of A: $I(X;Y|A)$. We can express this criterion formally as:

$$\arg\max_{X\in\bar{\mathcal{A}}}\left(\min_{A\in\mathcal{A}} I(X;Y|A)\right) \tag{7.10}$$

Another benefit of this criterion is that conditional mutual information can often be assessed more reliably than the loss on the full space of attributes, so it might be more robust to overfitting. The conditional mutual information can be cached for a more efficient execution of the algorithm. Fleuret (2004) found that this approach is highly competitive; in his experiments, NBC with such feature selection outperformed the support vector machine (SVM) classifiers and AdaBoost in the test set performance.

The disadvantage of the approach is that it only considers a single attribute of the context. The second disadvantage is that it will reward an attribute that is involved in positive interactions with other attribute in the model, although the naïve Bayesian classifier is unable to make use of positive interactions. Using interaction information, we can account for all the negative interactions the attribute is involved with. However, since the naïve Bayesian classifier in unable to take advantage of positive interactions, the interaction information can be capped at zero. The selection procedure in the style of previously listed ones picks the attribute X according to the following criterion:

$$\arg\max_{X\in\bar{\mathcal{A}}}\left(I(X;Y)+\sum_{A\in\mathcal{A}}\min\{0,I(X;Y;A)\}\right) \tag{7.11}$$

The capping is denoted by $\min\{0, I(X;Y;A)\}$, where we prevent positive interactions from improving the performance of a particular attribute. The computational cost of this procedure is equal to the one of Fleuret (2004). This heuristic may consider all attributes in the context, but it still neglects higher-order interactions: there may be positive interactions between negative interactions.

We have compared these three context-aware attribute selection algorithms on a random selection of UCI data sets. In the training set experiments in Fig. 7.10, it is clear that the brute force algorithm maintains an advantage, but that the interaction information-based algorithm (7.11) seems to hold a slight advantage over the Fleuret algorithm (7.10).

The advantage is even more pronounced in the generalization error on the training data, illustrated in Fig. 7.11, where we employed twice replicated 5-fold cross-validation: the sum of interaction information is less sensitive to the choice of the training data. This effect can be easily seen in the 'Mushroom' and 'Voting' data sets, which are smaller than

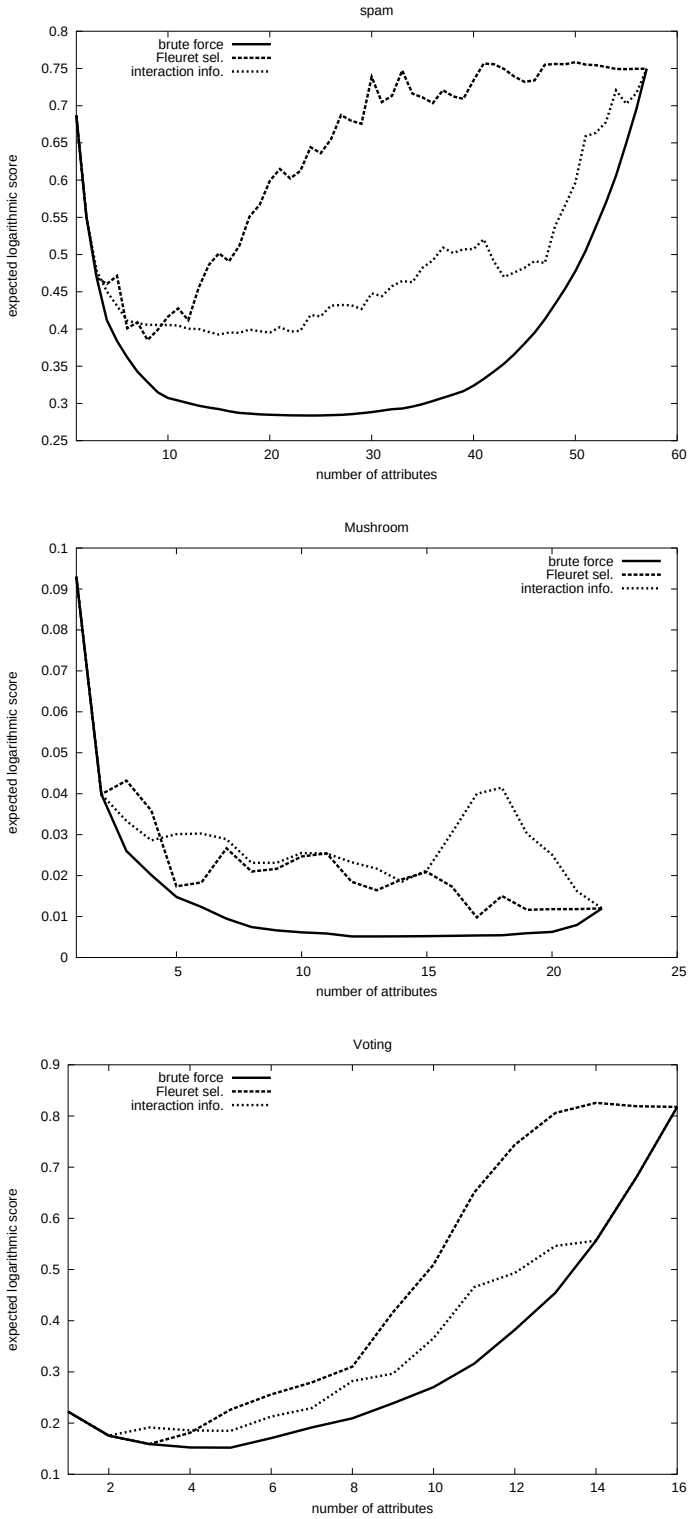


Figure 7.10: A comparison of context-sensitive attribute selection algorithms on the training data manifests the approximation error.

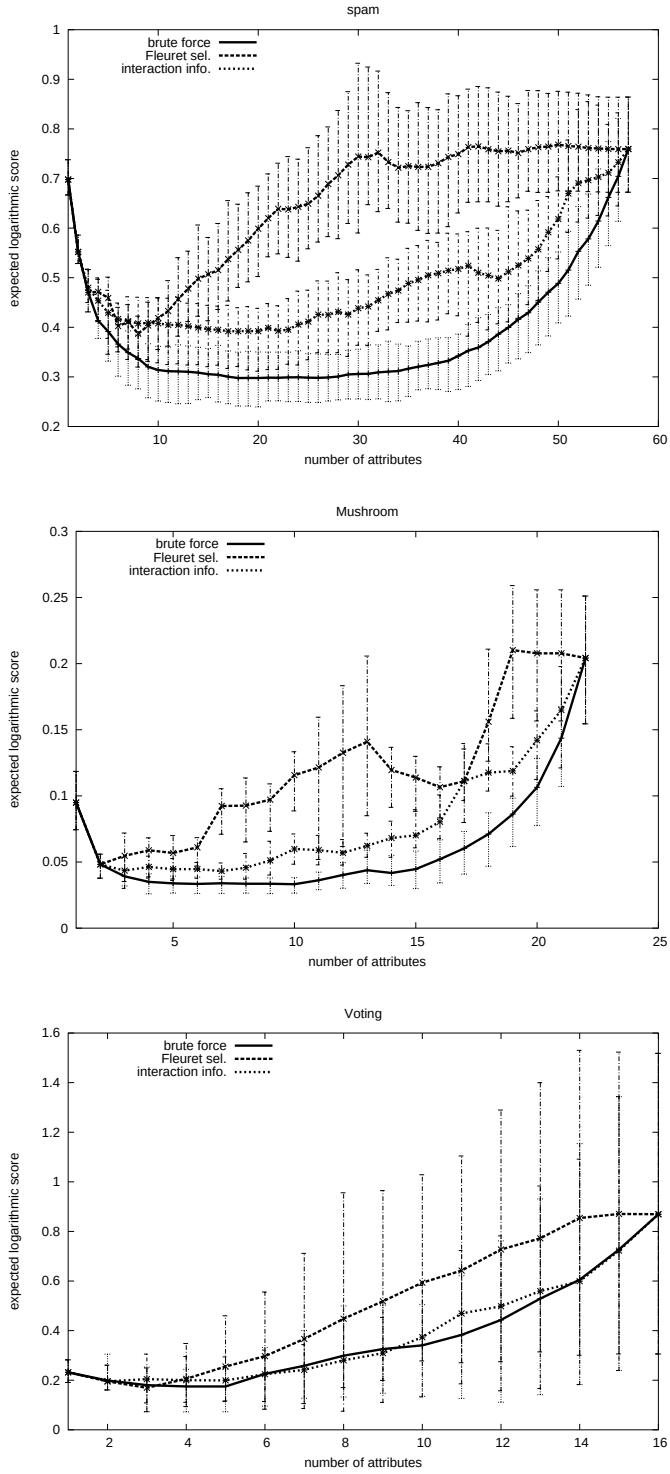


Figure 7.11: A comparison of context-sensitive attribute selection algorithms on the test data indicates that aggregate interaction information is a more reliable indicator of approximation error than the maximum conditional mutual information.

	Brier Loss of the Naïve Bayes Classifier with:				
	all attributes	brute force	mutual info.	Fleuret sel.	interaction info.
lung	0.575 ± 0.085 √	**0.562 ± 0.079**	0.581 ± 0.085 √	0.567 ± 0.088 √	0.634 ± 0.069 √
soy-small	0.000 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
zoo	0.125 ± 0.047 √	0.133 ± 0.046 √	**0.106 ± 0.038**	0.123 ± 0.046 √	0.125 ± 0.047 √
lymph	0.537 ± 0.056	**0.326 ± 0.033**	0.360 ± 0.027	0.368 ± 0.028	0.330 ± 0.018 √
wine	0.013 ± 0.008	0.008 ± 0.006 √	0.020 ± 0.011	0.017 ± 0.010	**0.006 ± 0.003**
glass	**0.618 ± 0.047**	0.630 ± 0.051 √	**0.618 ± 0.047**	**0.618 ± 0.047**	0.635 ± 0.053 √
breast	0.208 ± 0.016	0.198 ± 0.011 √	**0.192 ± 0.010**	0.194 ± 0.010 √	0.194 ± 0.010 √
ecoli	1.070 ± 0.056	**0.916 ± 0.042**	1.056 ± 0.054	1.070 ± 0.056	1.044 ± 0.049
horse-colic	1.013 ± 0.074	**0.423 ± 0.016**	0.426 ± 0.012 √	0.431 ± 0.012 √	0.431 ± 0.012 √
voting	0.091 ± 0.010	**0.037 ± 0.008**	0.044 ± 0.010 √	0.037 ± 0.009 √	0.040 ± 0.010 √
monk3	0.043 ± 0.004 √	**0.043 ± 0.004**	0.043 ± 0.004 √	0.043 ± 0.004 √	0.043 ± 0.004 √
monk1	0.174 ± 0.008 √	0.175 ± 0.008 √	**0.174 ± 0.008**	**0.174 ± 0.008**	**0.174 ± 0.008**
monk2	**0.229 ± 0.006**	0.229 ± 0.006 √	0.229 ± 0.006 √	0.229 ± 0.006 √	0.229 ± 0.006 √
soy-large	0.832 ± 0.095	**0.696 ± 0.088**	1.566 ± 0.103	1.039 ± 0.087	0.803 ± 0.082
wisc-cancer	**0.023 ± 0.004**	0.028 ± 0.005	0.031 ± 0.005	0.027 ± 0.005	0.026 ± 0.005 √
australian	0.112 ± 0.008 √	0.104 ± 0.008 √	0.111 ± 0.010 √	0.104 ± 0.008 √	**0.103 ± 0.010**
credit	0.111 ± 0.007	0.104 ± 0.007 √	0.110 ± 0.007	**0.100 ± 0.007**	0.102 ± 0.007 √
pima	0.160 ± 0.006	0.154 ± 0.005 √	0.154 ± 0.005 √	0.155 ± 0.005 √	**0.151 ± 0.005**
vehicle	0.589 ± 0.021	**0.446 ± 0.020**	0.584 ± 0.013	0.494 ± 0.019	0.487 ± 0.021
heart	0.713 ± 0.024	**0.664 ± 0.019**	0.696 ± 0.021	0.691 ± 0.021	0.670 ± 0.019 √
german	0.174 ± 0.007 √	**0.172 ± 0.005**	0.172 ± 0.007 √	0.174 ± 0.008 √	0.176 ± 0.008 √
cmc	0.445 ± 0.010	0.417 ± 0.006 √	**0.416 ± 0.006**	**0.416 ± 0.006**	**0.416 ± 0.006**
segment	0.262 ± 0.015	**0.150 ± 0.008**	0.287 ± 0.018	0.161 ± 0.009	0.200 ± 0.014
krkp	0.092 ± 0.004	**0.074 ± 0.002**	0.085 ± 0.003	0.078 ± 0.003	0.081 ± 0.002
mushroom	0.034 ± 0.003	**0.005 ± 0.001**	0.008 ± 0.000	0.008 ± 0.000	0.008 ± 0.001
adult	0.120 ± 0.002	**0.098 ± 0.001**	0.117 ± 0.002	0.101 ± 0.001	0.101 ± 0.001

times	**NB**	**B**	**MI**	**CMI**	**IG**
best	3	15	5	4	5
good √	6	10	8	10	13
bad	17	1	13	12	8

Table 7.4: Naïve Bayesian classifier almost always benefits from attribute selection. The greedy context-dependent brute force selection algorithm distinctly outperforms other algorithms. The interaction information heuristic is the most competitive of the heuristics, but not considerably better than the conditional mutual information heuristic.

'Spam'. The minimum and maximum performance over all 10 experiments is illustrated with error bars. A more complete evaluation with the same methodology as in previous sections is shown in Table 7.4: we used the model that resulted in the best performance on the training data.

Even if it would seem that the brute force algorithm is more prone to overfitting, its direct reference to the naïve Bayesian model maintains the advantage in comparison to the Fleuret and interaction information heuristics. This is no longer true, however, if we employ logistic regression instead of the naïve Bayes. As shown in Fig. 7.12, applying the brute force algorithm to logistic regression results in overfitting. Here, it might even be wiser to use information-theoretic attribute selection algorithms.

We can summarize our findings as:

- In the Sect. 7.3.4 we have seen that adding an attribute into a naïve Bayesian classifier may deteriorate the classification performance already on the training set. The reason for this is approximation error, caused by deviation from the conditional independence assumptions. This is generally not the case for logistic regression.
- A solution for this problem is to regard an attribute, which is a candidate for inclusion in the model, in the context of the attributes that are already in the model.
- The brute force heuristic (7.9) that examines the addition of all the available attributes and computes the performance of the resulting classifier on the test set is the best greedy heuristic we have found.
- The second best heuristic (7.11) accounted for the negative interaction information of the attribute with the attributes already in the model.

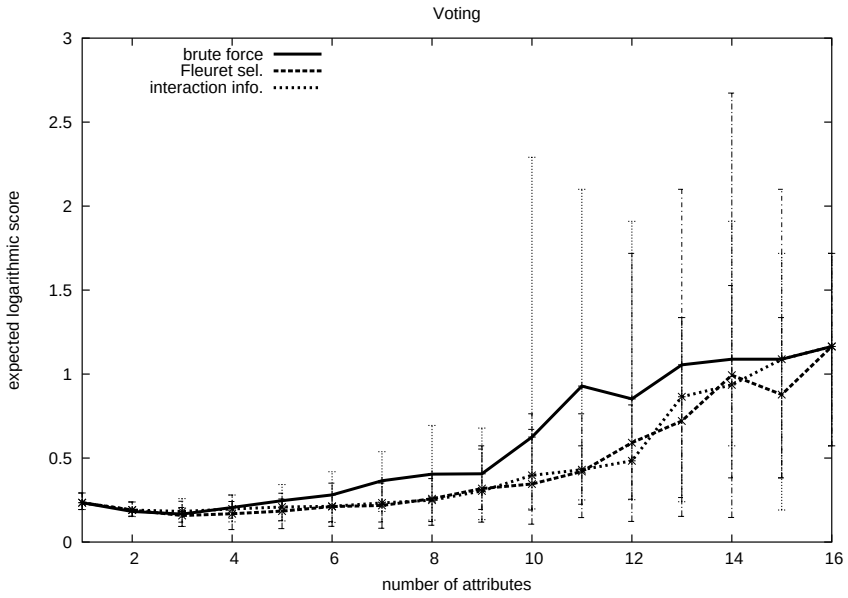


Figure 7.12: Brute force context-sensitive attribute selection for logistic regression may be prone to overfitting.

CHAPTER 8

Prediction with Interaction Models

We have already discussed how interactions may help in attribute selection in Ch. 7. Furthermore, we have also used interaction information as a heuristic to construct graphical models in Sect. 7.3.3, but the procedure was only half-automated. In Sect. 6.3.1, we mentioned that interaction graphs based on the significance tests of interactions are heuristical graphical models, but significance tests do not directly address the requirements for classification. In this chapter, we will discuss a fully automated approach specifically targeted towards supervised learning.

Most probability modelling work in machine learning has been based on graphical models (Darroch et al., 1980). In this chapter we will provide a practical approach to learning a considerably more general family of hierarchical models for classification. But because the term of hierarchical models has become roughly synonymous with multilevel models (Gelman et al., 2004b), we will try to avert confusion and refer to non-multilevel hierarchical models as *interaction models*. While graphical models are restricted with respect to different interactions that can exist in the same model, interaction models allow any k-tuple of attributes, labelled or unlabelled, to engage in a k-way interaction.

Because interaction models cannot be estimated in closed form in general, we will make use of methods that were developed in statistical physics for the purpose of approximating the joint and conditional probability models from a collection of marginal parts. We will fuse several local models, *submodels*, into a global joint model. Individual submodels act as constraints upon the joint or conditional models. With an approximate fusion method, *Kikuchi-Bayes*, we never have a globally normalized predictive model, but we renormalize it within a given classification context.

Although we work with discrete attributes and classification problems, the framework of prediction translates to continuous attributes and to regression problems. The notion of submodels as constraints is not trivial to be solved in an exact fashion, but the approximate fusion approach of Kikuchi-Bayes avoids the problem with ease.

8.1 Interaction Models

We will begin with the definition. First assume a set of attributes $\mathcal{V} = \{X_1, X_2, \ldots, X_m\}$. The attributes need not all be 'natural' but can also be constructed, hidden or latent. Each attribute may be labelled or unlabelled, and we identify the subset of labelled ones with $\mathcal{Y} \subseteq \mathcal{V}$, and the unlabelled ones with $\mathcal{X} = \mathcal{V} \setminus \mathcal{Y}$. We will refer to the value assignment to $\mathcal{X}$ as $\mathbf{X} = \mathbf{x}$, and to the value assignment to $\mathcal{Y}$ as $Y = y$. The whole value assignment will be $\mathbf{V} = \mathbf{X} \oplus Y$, where $\oplus$ denotes a concatenation of vectors. The model is specified with a set of marginal probability submodels $\mathcal{M} = \{P(\mathcal{S}|\boldsymbol{\theta}_{\mathcal{S}});\ \mathcal{S} \subseteq \mathcal{V}\}$. We can refer to a particular submodel $P(\mathcal{S})$ as a k-way interaction, where k denotes the cardinality of $\mathcal{S}$, $k = |\mathcal{S}|$. The submodel is defined with its parameters $\boldsymbol{\theta}_{\mathcal{S}}$.

While general models place no constraints on the set $\mathcal{M}$, the condition for interaction models is that there is no probability submodel $P(\mathcal{S}_1)$ that could be obtained by marginalizing another submodel $P(\mathcal{S}_2)$: such a submodel $P(\mathcal{S}_1)$ would be redundant. This happens when $\mathcal{S}_1$ is wholly contained in $\mathcal{S}_2$, and can be expressed by the non-redundancy condition: $\forall P(\mathcal{S}_1), P(\mathcal{S}_2) \in \mathcal{M} : \mathcal{S}_1 \cap \mathcal{S}_2 \notin \{\mathcal{S}_1, \mathcal{S}_2\}$. The non-redundancy condition keeps reduce the total number of hierarchical models, but neither by giving up the flexibility of the model, nor by increasing the complexity of the models.

The concept of an interaction model is not novel. An interaction model can also be represented with a factor graph (Kschischang et al., 2001), where attributes correspond to variable nodes, and submodels to factor nodes. It can also be represented as a log-linear model, or as a 'probabilistic database with a hypergraph scheme' (Badsberg and Malvestuto, 2001). It is important to distinguish graphs and hypergraphs. In a graphical model a k-clique corresponds to a single k-way interaction. In a hypergraph model, a k-clique corresponds to $k(k-1)$ 2-way interactions, and a k-hyperedge identifies a k-way interaction. The part-to-whole model corresponds to the complete hypergraph (Sect. 4.4).

Although the submodels are joint, it is possible to account for conditional submodels: If there is a conditional submodel $P(\mathcal{S}_1|\mathcal{S}_2)$, it has to be collapsed into a joint submodel $\hat{P}(\mathcal{S}_1, \mathcal{S}_2|\mathcal{M})$ by using $\mathcal{M}$ to build the model for $\hat{P}(\mathcal{S}_2|\mathcal{M})$. After this, $\hat{P}(\mathcal{S}_1, \mathcal{S}_2|\mathcal{M}) = P(\mathcal{S}_1|\mathcal{S}_2, \mathcal{M})\hat{P}(\mathcal{S}_2|\mathcal{M})$ can be included among the submodels in $\mathcal{M}$. It is possible to express Bayesian networks as interaction models: for every conditional probability $P(X|Y)$, we insert $P(X, Y)$ into $\mathcal{M}$.

8.2 Fusion Algorithms for Interaction Models

Let us now assume that an interaction model $\mathcal{M}$ is given for the attributes in $\mathcal{V}$. We will denote all the attribute values at once as a vector $\mathbf{V}$. $\mathcal{M}$ is composed of a number of submodels, and the goal of fusion is to form a single joint $P(\mathbf{V})$ based on the set of submodels $\mathcal{M} = \{P(\mathbf{V}_{\mathcal{S}_1}), \ldots, P(\mathbf{V}_{\mathcal{S}_\ell})\}$. To this aim, we will employ a *fusion algorithm*.

The usual approach is to parameterize $P(\mathbf{V}|\boldsymbol{\Theta})$ and view the submodels as constraints upon $\boldsymbol{\Theta}$. There are many values of $\boldsymbol{\Theta}$ that are consistent with their marginal constraints, and the question is which individual $\boldsymbol{\Theta} = \boldsymbol{\theta}$ to pick. The maximum entropy principle (Sect. 2.2.4) states that among the several consistent models, one should pick the one that has the highest entropy. If we interpret entropy as expected loss, the MaxEnt model is worst-case optimal. The downside to the direct maximum entropy approach is that the model has to be specified in its exponentially increasing entirety (Wallach, 2004). For that

purpose, it is helpful to seek approximations. But approximations are feasible, even for partially overlapping submodels, and they are examined in Sect. 8.2.3.

8.2.1 Maximum Entropy Fusion

Depending on our purpose, we may choose to maximize the joint entropy or the conditional entropy. Maximum joint entropy assures that the joint model will be as unpretentious as possible for joint modelling given the submodel constraints. Maximum conditional entropy assures that the conditional model will be as careful as possible for prediction of the conditioned attributes with the ones conditioned with. For a more detailed discussion see Sect. 5.3.2.

Generally, maximizing entropy directly is problematic. Most present methods take a number of assumptions regarding the form of the joint model, usually focusing on the exponential family models on binary attributes. Instead of maximizing the Shannon entropy, we may maximize the empirical entropy $\hat{H}(\mathcal{V}|\theta, \mathcal{D})$ in the context of the data $\mathcal{D}$. It is also possible to employ Rényi entropy instead of Shannon entropy for maximization (Jaynes, 1957, Zitnick and Kanade, 2004).

A different strategy is to employ a constraint satisfaction algorithm from the iterative scaling family (Darroch and Ratcliff, 1972). It can be shown that if the initial approximation is a uniform distribution, the constraint-satisfied solution will try to remain as close as possible to the initial approximation. Generalized iterative scaling is not a particularly efficient optimization algorithm, and more sophisticated gradient descent algorithms may be used instead (Malouf, 2002).

Entropy maximization is a constrained primal optimization problem in the space of consistent distributions, but we may tackle the unconstrained dual problem of maximizing the log-likelihood with gradient descent in the space of maximum entropy distributions (Berger et al., 1996, Della Pietra et al., 1997). Namely, entropy maximization given the constraints can be modelled with the *Boltzmann distribution* (Jaynes, 2003).

The Boltzmann distribution (also referred to as a Gibbs state with potentials ψ, as Gibbs or Boltzmann-Gibbs distribution, or as a loglinear model in statistics) is expressed as (Darroch et al., 1980):

$$P(\mathbf{v}|\boldsymbol{\beta}) \triangleq \frac{1}{Z} \prod_{P(\mathcal{S})\in\mathcal{M}} \psi_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}) = \exp\left\{ \sum_{P(\mathcal{S})\in\mathcal{M}} \log \psi_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}) - \log Z \right\} \tag{8.1}$$

We can see that instead of a marginal $P(\mathcal{S})$, we have an *potential* $\psi(\mathcal{S})$, a nonnegative function from $\Re_{\mathcal{S}}$. The parametrization $\boldsymbol{\beta}$ defines the potentials. Z is referred to as the *partition function* and it assures that the global model is normalized:

$$Z \triangleq \int_{\Re_{\mathcal{V}}} \prod_{P(\mathcal{S})\in\mathcal{M}} \psi_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}) d\mathbf{v} \tag{8.2}$$

8.2.2 Region-Based Approximation to Free Energy

We will now describe a few terms from statistical mechanics, associating them with the terms that have already been used in this text. For a particular global configuration $\mathbf{v}$

we can define the *energy*, also referred to as the Hamiltonian, by the following (Yedidia et al., 2004):[1]

$$E(\mathbf{v}) \triangleq -\sum_{P(\mathcal{S})\in\mathcal{M}} \log \psi_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}) \tag{8.3}$$

The Helmholtz free energy of the system is $-\log Z$ (Yedidia et al., 2004). Often the Boltzmann distribution is written as $\exp\{-E(\mathbf{v})/T\}/Z(T)$, where T stands for temperature, and E for energy. We will henceforth assume that $T = 1/k_B$, but remember why Z is referred to as a function.

$P(\mathbf{v}|\boldsymbol{\beta})$ and Z are generally intractable. For that reason, we would prefer to make use of a hypothetical $\hat{P}$. The methods that make use of this are referred to as *variational*. We can now define the *variational average energy* $U(\mathcal{V}|\hat{P})$ and the *variational entropy* $H(\mathcal{V}|\hat{P})$ (Yedidia et al., 2004):

$$U(\mathcal{V}|\hat{P}) \triangleq \int_{\Re_{\mathcal{V}}} \hat{P}(\mathbf{v})E(\mathbf{v})d\mathbf{v} \tag{8.4}$$

$$H(\mathcal{V}|\hat{P}) \triangleq -\int_{\Re_{\mathcal{V}}} \hat{P}(\mathbf{v})\log \hat{P}(\mathbf{v})d\mathbf{v} \tag{8.5}$$

The *variational free energy* $F(\mathcal{V}|\hat{P})$ can now be defined as:

$$F(\mathcal{V}|\hat{P}) \triangleq U(\mathcal{V}|\hat{P}) - H(\mathcal{V}|\hat{P}) = D(\hat{P}\|P) + \log Z \tag{8.6}$$

Clearly, the variational free energy is at the minimum precisely when $\hat{P}$ and P are equal.

We employ the variational free energy minimization to seek a tractable $\hat{P}$ that approximates the true P well. We can ignore the partition function in performing this variational optimization, because it is constant. We have not gained much unless $\hat{P}$ is in some way simpler than P. We can employ the *mean-field* form for $\hat{P}$ where $\hat{P}_{MF} = \prod_{V\in\mathcal{V}} \hat{P}_{MF}(V)$. Given, say, a Bayesian network model $\hat{P}_{BN}$, we could then minimize the variational free energy $D(\hat{P}_{MF}\|\hat{P}_{BN})$ to obtain $\hat{P}_{MF}(V)$ for each $V \in \mathcal{V}$.

Another path, suggested by Yedidia et al. (2004) is to redefine the variational free energy expression. Now assume that there are several *regions* that can possibly overlap. A region is defined as a subset of the potentials in (8.1), but we will interpret it as a set of attributes. We represent the regions with a *weighted set of regions* $\mathcal{G}_R = \{\langle \mathcal{R}, c_{\mathcal{R}}\rangle\}$, where $c_{\mathcal{R}}$ denotes the *counting number* of the region $\mathcal{R}$. The importance of counting numbers is to assure that each attribute and each interaction[2] is counted exactly once.

Based on the weighted set of regions, we can define the *region-based approximate entropy*, which can be computed on the whole $\mathcal{V}$:

$$\hat{H}_{\mathcal{G}_R}(\mathcal{V}|\hat{P}) \triangleq -\sum_{\langle \mathcal{R}, c_{\mathcal{R}}\rangle\in\mathcal{G}_R} c_{\mathcal{R}} \int_{\Re_{\mathcal{R}}} \hat{P}(\mathbf{v}_{\mathcal{R}})\log \hat{P}(\mathbf{v}_{\mathcal{R}})d\mathbf{v}_{\mathcal{R}} \tag{8.7}$$

Within a particular region, Yedidia et al. (2004) define the *region energy* based on the Boltzmann probability model:

$$E_{\mathcal{R}}(\mathbf{v}_{\mathcal{R}}) \triangleq -\sum_{P(\mathbf{v}_{\mathcal{S}})\in\mathcal{M},\, \mathcal{S}\subseteq\mathcal{R}} \log \psi_{\mathcal{S}}(\mathbf{v}_{\mathcal{S}}) \tag{8.8}$$

[1] Neal and Hinton (1998) refer to $E(\mathbf{v}) + \log Z = -\log P(\mathbf{v}|\boldsymbol{\beta})$ as energy.

[2] Here we interpret each interaction $\mathcal{S}$ as a factor in the factor graph terminology of Yedidia et al. (2004). Namely, each $\mathcal{S}$ corresponds to a hyperedge, and a factor graph is just a way of denoting hypergraphs with bigraphs.

In the context of generalized belief propagation, there is a definition of region-based average energy, based on the Boltzmann probability model within the region, and of the resulting region-based free energy. It is then possible to seek the such marginals $\hat{P}(\mathbf{v}_\mathcal{R})$ for each region $\mathcal{R}$ so that the *region-based free energy* is minimized:

$$\hat{F}_\mathcal{R}(\mathcal{V}|\hat{P}) \triangleq \sum_{\langle \mathcal{R}, c_\mathcal{R} \rangle \in \mathcal{G}_R} c_\mathcal{R} \int_{\Re_{\mathcal{V}_\mathcal{R}}} \hat{P}(\mathbf{v}_\mathcal{R}) \log \frac{\hat{P}(\mathbf{v}_\mathcal{R})}{\prod_{\mathcal{S} \subseteq \mathcal{R}} \psi_\mathcal{S}(\mathbf{v}_\mathcal{R})} d\mathbf{v}_\mathcal{R} \tag{8.9}$$

The fixed points of the generalized belief propagation algorithm then correspond to the stationary points of the region-based free energy.

8.2.3 Region-Based Approximation to Probability Models

Before using the region-based free energy $\hat{F}_\mathcal{R}(\mathcal{V}|\hat{P})$ as a criterion in optimization, let us first question whether region-based approximate entropy on the true marginals on regions is a faithful approximation to the true entropy. We will interpret the difference in entropies as a Kullback-Leibler divergence and infer an approximate joint probability model that corresponds to a given set of regions.

If there is no region in $\mathcal{G}_R$ that would not be contained within some submodel $P(\mathcal{S}) \in \mathcal{M}$, we can compute the non-approximate *region-based entropy* with respect to the true joint probability model $\hat{H}_{\mathcal{G}_R}(\mathcal{V}|P)$. Namely, if the marginals in region-based approximate entropy are indeed correct, as it is sought by the free energy minimization, the region-based approximate entropy will be equal to the region-based entropy.

Calculating the region-based entropy for an interaction model can be done in closed form without actually having the 'true' joint model P, just the set of its 'true' marginals. These are already given as the submodels. The reliability of the approximation to entropy can be evaluated by the following:

$$\hat{H}_{\mathcal{G}_R}(\mathcal{V}|P) - H(\mathcal{V}|P) = D(P \| \hat{P}'_{\mathcal{G}_R}), \quad \hat{P}'_{\mathcal{G}_R}(\mathcal{V}) \triangleq \prod_{\langle \mathcal{R}, c_\mathcal{R} \rangle \in \mathcal{G}_R} P(\mathcal{R})^{c_\mathcal{R}} \tag{8.10}$$

Unfortunately, as we have seen in Sect. 4.4.1, there are valid weighted sets of regions that may result in negative region-based entropy: the underlying $\hat{P}_{\mathcal{G}_R}$ is non-normalized. Nevertheless, in many circumstances the subsequently normalized *region-based probability approximation* $\hat{P}_{\mathcal{G}_R}(\mathcal{V}) \propto \hat{P}'_{\mathcal{G}_R}(\mathcal{V})$ is a reasonable approximation to the Boltzmann model $P(\mathcal{V}|\boldsymbol{\beta})$, and in some cases it is exact. Region-based probability approximation generalizes the Kirkwood superposition approximation from Sect. 4.4.1.

8.2.4 Constructing Region-Based Representations

When we approach the problem of obtaining the region-based probability approximation for a given interaction model, we have to decide which regions to pick. Generally, each interaction should be assigned an *initial region*. Clearly, each region should be a subset of some submodel $\mathcal{S}$, as otherwise there would not be enough information to obtain the marginal. However, the initial regions may overlap, so heuristics are needed to account for this. Yedidia et al. (2004) describe a few properties of a region-based representation.

Validity The weighted set of regions is a *valid* approximation of the interaction model $\mathcal{M}$ iff:

$$\forall V \in \mathcal{V} : \sum_{\langle \mathcal{R}, c_{\mathcal{R}} \rangle \in \mathcal{G}_R, V \in \mathcal{R}} c_{\mathcal{R}} = 1 \tag{8.11}$$

$$\forall P(\mathcal{S}) \in \mathcal{M} : \sum_{\langle \mathcal{R}, c_{\mathcal{R}} \rangle \in \mathcal{G}_R, \mathcal{S} \subseteq \mathcal{R}} c_{\mathcal{R}} = 1 \tag{8.12}$$

This means that each attribute and each interaction is counted exactly once. Yedidia et al. (2004) prove that if the region-based representation is valid, the region-based average energy will be exact. They also prove that if the true joint probability model is a multivariate uniform distribution, and if the marginals are correct, the region-based entropy will be exactly the true entropy.

MaxEnt-Normality A weighted set of regions is MaxEnt-normal if it is valid and if the corresponding region-based entropy for some set of marginals $\hat{P}(\mathcal{R})$ achieves its maximum when all the marginals of $\hat{P}$ are uniform.

In addition to this, Yedidia et al. (2004) mention an additional property that results in good performance: $\sum_{\mathcal{R}} c_{\mathcal{R}} = 1$. We will now address the methods for obtaining the region graph given a set of interactions or initial regions, summarizing Yedidia et al. (2004). We will unfairly neglect to discuss other approximations, such as the *hypertree factorization* (Wainwright and Jordan, 2003) that can also be used to construct region graphs of higher quality than the Kikuchi approximation. It is not necessary that the counting numbers are integers, namely.

Exact Special Cases

For a few configurations of interactions, the resulting $\hat{P}'_{\mathcal{G}_{\mathcal{R}}}$ requires no normalization and is exact. In reality, the configurations are chosen as to facilitate exactness, and not by the actual properties of the data.

Mean Field / Fully Factorized The mean field approximation is applicable when the interactions are disjunct (non-overlapping). In such a case, each submodel $P(\mathcal{S})$ has a corresponding region $\langle \mathcal{S}, 1 \rangle$.

Junction Tree Some arrangements of interactions can be represented with *junction trees*. A junction tree is a tree whose nodes are subsets of attributes, and obeys the *junction tree property*: if a node $\mathcal{S}_2$ is on the path between $\mathcal{S}_1$ and $\mathcal{S}_3$, then $\mathcal{S}_1 \cap \mathcal{S}_3 \subseteq \mathcal{S}_2$. Furthermore, if $\mathcal{S}_1 \cap \mathcal{S}_3 \neq \emptyset$, there must exist a path between $\mathcal{S}_1$ and $\mathcal{S}_3$. We form a region-based representation of the junction tree by assigning $\langle \mathcal{S}, 1 \rangle$ to each node $\mathcal{S}$, and $\langle \mathcal{S}_1 \cap \mathcal{S}_2, -1 \rangle$ to each edge $\langle \mathcal{S}_1, \mathcal{S}_2 \rangle$ in the junction tree.

Bethe Approximation

The Bethe approximation starts with the initial regions, and each of which is assigned the weight of 1. So, for each $P(\mathcal{S}) \in \mathcal{M}$ there is a $\langle \mathcal{S}, 1 \rangle \in \mathcal{G}_R$. Because the same attribute $V \in \mathcal{V}$ may appear in multiple regions, we introduce a *small region* for every $V \in \mathcal{V}$, where

$\mathcal{R}_0 \leftarrow \{\emptyset\}$ {Redundancy-free set of initial regions.}
for all $\mathcal{S} \in \mathcal{M}$ **do** {for each initial region}
  if $\forall \mathcal{S}' \in \mathcal{R}_0 : \mathcal{S} \not\subseteq \mathcal{S}'$ **then**
    $\mathcal{R}_0 \leftarrow \mathcal{R}_0 \cup \{\mathcal{S}\}$ {$\mathcal{S}$ is not redundant}
  end if
end for
$\mathcal{R} \leftarrow \{\langle \mathcal{S}, 1\rangle;\ \mathcal{S} \in \mathcal{R}_0\}$
$k \leftarrow 1$
while $|\mathcal{R}_{k-1}| > 2$ **do** {there are feasible subsets}
  $\mathcal{R}_k \leftarrow \{\emptyset\}$
  for all $\mathcal{I} = \mathcal{S}^\dagger \cap \mathcal{S}^\ddagger : \mathcal{S}^\dagger, \mathcal{S}^\ddagger \in \mathcal{R}_{k-1}, \mathcal{I} \notin \mathcal{R}_k$ **do** {feasible intersections}
    $c \leftarrow 1$ {the counting number}
    for all $\langle \mathcal{S}', c' \rangle \in \mathcal{R}, \mathcal{I} \subseteq \mathcal{S}'$ **do**
      $c \leftarrow c - c'$ {consider the counting numbers of all regions containing the intersection}
    end for
    if $c \neq 0$ **then**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{\langle \mathcal{I}, c\rangle\}$
    end if
    $\mathcal{R}_k \leftarrow \mathcal{R}_k \cup \{\mathcal{I}\}$
  end for
end while
return $\{\langle R, c\rangle \in \mathcal{R}; c \neq 0\}$ {Region graph with the counting numbers.}

Algorithm 8.1: This algorithm yields the Kikuchi approximation when given the initial set of regions $\mathcal{M} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\ell\}$.

$c_V = 1 - |P(\mathcal{S}) \in \mathcal{M}; V \in \mathcal{S}|$ is 1 less the number of interactions in which V is involved. Based on this region-based approximation, we may define the *Bethe entropy* as

$$\sum_{P(\mathcal{S})\in\mathcal{M}} H(\mathcal{S}) + \sum_{V\in\mathcal{V}} H(V)(1 - |\{P(\mathcal{S}) \in \mathcal{M}|V \in \mathcal{S}\}|). \tag{8.13}$$

We have already employed Bethe entropy in Sect. 6.4.4. Our joint model included all interactions of the type $\{A, X_i\}$, and we used it to compute the Bethe entropy (denominator), and the difference between the mean field entropy and Bethe entropy (nominator) in (6.19).

Kikuchi Approximation

The Kikuchi approximation, also referred to as the cluster variation method, is perhaps the most flexible approach to constructing the region-based representation of a given interaction model. It is based on assigning each initial region the counting number of 1. To account for the overlap, it considers all the intersections between pairs of initial regions and forms the first new layer of regions that correct for the overlap. All the overlap might not be accounted for yet, so it continues by considering all intersections among the first layer of regions and forms the second layer of regions. The full procedure is shown as Algorithm 8.1.

When the initial regions are disjunct, the Kikuchi approximation is identical to the mean field. It is not difficult to see that the Kikuchi approximation is identical to the Bethe approximation when no pair of initial regions shares more than a single attribute. Furthermore, the Kikuchi approximation on the complete set of regions of size $n-1$ given n attributes is identical to the Kirkwood superposition approximation. Most importantly, the Kikuchi approximation is exact when the initial regions are the maximal cliques of a Markov network. In general, Kikuchi approximation is exact when only the initial regions have their counting numbers greater than zero (Yedidia et al., 2001).

8.3 Learning Interaction Models

In the previous section we have described two ways of defining the joint probability distribution based on the set of submodels $\mathcal{M}$. The first is to discover the parameter vector $\boldsymbol{\beta}$ that defines the Boltzmann distribution consistent with the interactions. The second approach is to approximate the Boltzmann distribution with the region-based approximation. Regardless of which approximation we pick, a practical learning algorithm has to decide which submodels to pick and how to estimate them from the data. This will be the topic of the present section.

We will adopt the Bayesian framework, based on an explicit description of the model in terms of its parameters $\phi = (\mathcal{M}, \boldsymbol{\Theta}, \vartheta)$. It is desirable that the structure $\mathcal{M}$ is independent of the submodel prior ϑ and the submodel parameters $\boldsymbol{\Theta}$. It would be paradoxical that some submodel $P(X_1, Y)$ changed because some new attribute X_2 was introduced into the model. Submodels and marginals of the joint should remain invariant given the overall model structure. Our goal is achieved by the following factorization of the prior $P(\phi) = P(\mathcal{M})P(\vartheta)P(\boldsymbol{\Theta}|\vartheta) = P(\mathcal{M})P(\vartheta)\prod_i P(\boldsymbol{\theta}_i|\vartheta)$. We will now address two additional constraints: first, the submodels must be a posteriori consistent in spite of the conditional independence of their parameters; second, the models should be parsimonious: as simple as possible but not simpler.

For a certain model structure $\mathcal{M}$, we compute the class predictive distribution $\hat{P}(y|\mathbf{x}, \phi)$ by applying Algorithm 8.1 to the set of submodels. The resulting region graph $\mathcal{G}_R$ is used as the basis for the joint approximation:

$$\hat{P}'(\mathbf{v}|\mathcal{M}) \propto \prod_{\langle \mathcal{R}, c_\mathcal{R} \rangle \in \mathcal{G}_R} P(\mathbf{v}_\mathcal{R})^{c_\mathcal{R}} \tag{8.14}$$

We do not employ the joint approximation directly, we will not even attempt to normalize it. Instead we will compute the conditional probability distribution of the label y in the context of the attribute vector $\mathbf{x}$, and at the same time we will also perform the normalization solely in the context of $\mathbf{x}$:

$$\hat{P}(y|\mathbf{x}, \mathcal{M}) \triangleq \frac{\hat{P}'(y, \mathbf{x}|\mathcal{M})}{\sum_{y' \in \Re_Y} \hat{P}'(y', \mathbf{x}|\mathcal{M})} \tag{8.15}$$

For prediction we thus integrate the model structure out ('Bayesian model averaging'), because this technique has shown its virtues a number of times (Buntine, 1991, Hoeting et al., 1999, Cerquides and López de Màntaras, 2003). The final result of our inference based on data $\mathcal{D}$ will thus be the following class predictive distribution, obtained by

integrating over several choices of the structure $\mathcal{M}$ and possibly other components of the prior:

$$\hat{P}(y|\mathbf{x}) \propto \int P(\phi|\mathcal{D})\hat{P}(y|\mathbf{x},\phi)d\phi \tag{8.16}$$

We can see that an individual structure is weighted accordingly to its likelihood and its prior. Because of that, we implement the Bayesian maxim that all models are equally justifiable if they have the same likelihood and the same prior. In the following sections we will discuss the details of our priors and our greedy algorithm for efficiently integrating in the space of ϕ. Still, the value of ϕ with the maximum a posteriori probability is interesting as the best individual model in the ensemble.

It should be easy to see that our algorithms are based on some method of fusion used to come up with $\hat{P}$. Although the Kikuchi approximation is used, effectively nothing changes if Bethe approximation of the actual Boltzmann distribution is used instead. We do employ (8.15) to obtain the class-predictive distribution for a specific structure $\mathcal{M}$, conditioning the joint probability model. However, there is no reason why an inherently conditional model is used instead, for example maximizing the conditional entropy, or maximizing the conditional likelihood. An example of such conditional models are the conditional random fields (CRF) (Lafferty et al., 2001). The main difference is that the purpose of CRF here would be to fuse the interactions, and not to model the data directly.

8.3.1 Estimating Consistent Submodels

The submodels have no specific ordering, and should be estimated independently from data $\mathcal{D}$. After the estimation, we work with posterior predictive distributions $P(\mathbf{v}_R|\mathcal{D})$, without referring back to their parameters. It is important, however, to assure that the predictive submodels are in fact consistent. Consistency means that there does exist some joint probability model $P(\mathbf{V}|\mathcal{D})$ for data $\mathcal{D}$ so that each submodel $P(\mathbf{V}_\mathcal{S}) \in \mathcal{M}$ and $\mathcal{W} = \mathcal{V} \setminus \mathcal{S}$: $P(\mathbf{v}_\mathcal{S}) = \int_{\Re_\mathcal{W}} P(\mathbf{v}_\mathcal{S}, \mathbf{v}_\mathcal{W}|\boldsymbol{\theta})d\mathbf{v}_\mathcal{W}, \forall \mathbf{v}_\mathcal{S} \in \Re_\mathcal{S}$. For example, an indication of inconsistency is the existence of a subset of attributes $\mathcal{W} \subseteq \mathcal{V}$ that appear in two submodels $\mathcal{S}_1$ and $\mathcal{S}_2$: $\mathcal{W} \subseteq \mathcal{S}_1 \cap \mathcal{S}_2$, but where the marginalizations do not match:

$$\exists \mathbf{v}_\mathcal{W} \in \Re_\mathcal{W} : \int_{\Re_{\mathcal{S}_1 \setminus \mathcal{W}}} P(\mathbf{v}|\boldsymbol{\theta})d\mathbf{v}_{\mathcal{S}_1} \neq \int_{\Re_{\mathcal{S}_2 \setminus \mathcal{W}}} P(\mathbf{v}|\boldsymbol{\theta})d\mathbf{v}_{\mathcal{S}_2} \tag{8.17}$$

While maximum likelihood estimation would result in consistent submodels, Bayesian modelling requires some forethought. Namely, each submodel is modelled based on the same prior, but independently of other submodels, including those that overlap with it. Some popular choices of parameter priors, such as the Laplacean prior, would result in inconsistent submodel posteriors. Imagine estimating two entangled coins using the Laplacean prior. If a single coin c_1 is estimated independently, we will obtain the posterior predictive probability of $p_H = (1 + \#_{c_1=H})/(2 + \#)$. If we estimate two co-tossed coins simultaneously, and marginalize c_2 out, we obtain a non-matching

$$p_H = \frac{2 + \#(c_1 = H, c_2 = H) + \#(c_1 = H, c_2 = T)}{4 + \#}.$$

Let us now consider a submodel on attributes $\mathbf{X}_s = \{X_1, X_2, \ldots, X_k\}$. All the attributes are assumed to be nominal, and the multinomial submodel would be appropriate.

The multinomial submodel is parameterized by the vector $\boldsymbol{\theta}_{\mathcal{S}}$ whose dimensionality corresponds to the cardinality of $\prod_{i=1}^{k} |\mathcal{X}_i|$. A coordinate $\boldsymbol{\theta}_{\mathcal{S}:x_1,\ldots,x_k}$ can be interpreted as the probability of occurrence of $(x_1, \ldots x_k)$. What we need is a prior $P(\boldsymbol{\theta}_{\mathcal{S}})$ that assures that the posterior predictive distribution $P(\mathbf{x}_{\mathcal{S}}|\mathcal{D}) = \int P(\boldsymbol{\theta}_{\mathcal{S}}|\mathcal{D})P(\mathbf{x}_{\mathcal{S}}|\boldsymbol{\theta}_{\mathcal{S}})d\boldsymbol{\theta}_{\mathcal{S}}$ will be consistent with all submodels that share attributes with $\mathbf{X}_{\mathcal{S}}$.

It is quite easy to see that the following choice of the symmetric Dirichlet prior fulfills the demand of predictive consistency, if the same value of ϑ is used for all the submodels:

$$P(\boldsymbol{\theta}_{\mathcal{S}}|\vartheta) = \text{Dirichlet}(\alpha, \ldots, \alpha), \quad \alpha = \frac{\vartheta}{\prod_{i=1}^{k} |\Re_{X_i}|} \tag{8.18}$$

This prior is best understood as the expected number of outliers: to any data set, we add ϑ uniformly distributed instances. There is also an implied assumption of no structural zeros: not making such an assumption may result in zero likelihood of the test data.

In some cases, it is desirable not to integrate the parameters $\boldsymbol{\theta}_{\mathcal{S}}$ out. Instead, we would want a posterior distribution over $\boldsymbol{\theta}_{\mathcal{S}}$. The main problem here is to assure the consistency in each posterior sample: this can be achieved by sampling in the space of consistent parameter values, by ensuring that the prior probability of inconsistent parameters is zero. For each such sample, we perform the maximum entropy fusion. A further discussion of the relation between Bayesian and MaxEnt inference appears in (Cheeseman and Stutz, 2004). Again, we will not follow this approach and will only work on the submodel posterior means.

8.3.2 Parameterizing the Structure

The structure in the context of Kikuchi-Bayes is simply a selection of the submodels. $P(\mathcal{M})$ models our prior expectations about the structure of the model. Parsimony means that we should not select all the submodels, and the motivation for this is not just the subjective desire for simplicity but also the frequentist problem of objective identifiability and the decision-theoretic desire to minimize the expected loss. We will now provide a parsimonious prior that asserts a higher prior probability to simpler selections of submodels.

The primary question is how to quantify the complexity of the set of submodels. Neither the number of submodels nor the total number of parameters across the submodels in $\mathcal{M}$ would be sensible choices: some submodels describe attributes with a greater number of values, and some submodels may be partly contained within other submodels. An interesting quantification of complexity that solves this dilemma is given by Krippendorff (1986) in the context of loglinear models without structural zeros. Let us assume a set of overlapping submodels of the attribute vector $\mathbf{V}$, and the resulting region graph $\mathcal{R}$ obtained using the CVM. The number of *degrees of freedom* of the joint model $\mathcal{M}$ with a corresponding region graph $\mathcal{R}$ is:

$$df_{\mathcal{M}} \triangleq \sum_{\langle \mathcal{S}, c \rangle \in \mathcal{R}} c\Big(-1 + \prod_{X \in \mathcal{S}} |\Re_X| \Big) \tag{8.19}$$

The overlap between submodels is hence handled in an analogous way both for fusion in (8.15) and for the assessment of degrees of freedom.

The following prior corresponds to the assumption of exponentially decreasing prior probability of a structure with an increasing number of degrees of freedom (or effective parameters):

$$P(\mathcal{M}) \stackrel{\triangle}{\propto} \exp\left\{-\frac{m\ df_{\mathcal{M}}}{m - df_{\mathcal{M}} - 1}\right\} \tag{8.20}$$

We discourage the degrees of freedom from exceeding the number of training instances m. This choice of the prior has a frequentist justification: it corresponds to the Akaike information criterion (AIC) with small-sample correction (Burnham and Anderson, 2002). Akaike information criterion and the underlying objective of minimizing the expected loss on an independent sample from the model itself has been justified in philosophy of science (Forster and Sober, 1994). Performing MAP inference of the structure parameter $\mathcal{M}$ with such a prior would correspond to maximizing the AIC. Thus, our prior corresponds to the subjective choice of the frequentist paradigm along with a particular loss function. A Bayesian will make sure that the prior is properly normalized, of course, and will attempt to account for multiple models.

A submodel can be included into $\mathcal{M}$ without increasing the degrees of freedom. Assume that Kikuchi approximation has considerable approximation error fusing a set of interactions $\mathcal{M}_c \subset \mathcal{M}$. We can perform the perfect MaxEnt fusion solely on the attributes involved in $\mathcal{M}_c$, form a submodel and include that submodel. Some care must be consequently taken to assess the degrees of freedom correctly. Furthermore, we should note that ordinary MaxEnt is not necessarily appropriate for classification, as we have seen in Sect. 5.3.2. But in the simple joint case, it is easy to see that Kikuchi approximation would be exact had we performed the above procedure on all maximal cliques of the Markov network that corresponds to $\mathcal{M}$.

8.3.3 The Prior and the Likelihood Function for Classification

Our objective is predictive class probability estimation with the Kikuchi approximation (8.15). We need to define the prior on the structure variable $\mathcal{M}$ and the likelihood of $\mathcal{M}$ given the data. If $\mathcal{M}$ is going to be used for prediction, the effective degrees of freedom are fewer ("Conditional density estimation is easier than joint density estimation."). Assuming a single attribute X_i, the degrees of freedom of the conditional model $P(Y|X_i)$ correspond to the difference between the cardinality of the range of both Y and X_1 at once less the cardinality of the range of X_1 alone: $df_{Y|X_i} = |\Re_{X_i} \times \Re_Y| - |\Re_{X_1}|$. In general, if we condition upon a subset of attributes $\mathcal{Y} \subseteq \mathcal{X}$, the degrees of freedom of the resulting conditional model will be defined as:

$$df_{\mathcal{M}_\mathcal{Y}} \triangleq \sum_{\langle \mathcal{S}, c\rangle \in \mathcal{R}} c\left(\prod_{X\in\mathcal{S}} |\Re_X| - \prod_{\substack{X\in\mathcal{S}\\ X\notin\mathcal{Y}}} |\Re_X|\right) \tag{8.21}$$

The prior $P(\mathcal{M}_\mathcal{Y})$ is obtained by plugging (8.21) into (8.20).

The growth of structures should be guided by whether the addition of a submodel is of benefit in predicting the label. The following conditional likelihood function takes this into account:

$$\hat{P}(\mathbf{v}^{(1)\ldots(m)}|\mathcal{M}_\mathcal{Y}) \triangleq \prod_{i=1}^{m} \hat{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \mathcal{M}_\mathcal{Y}) \tag{8.22}$$

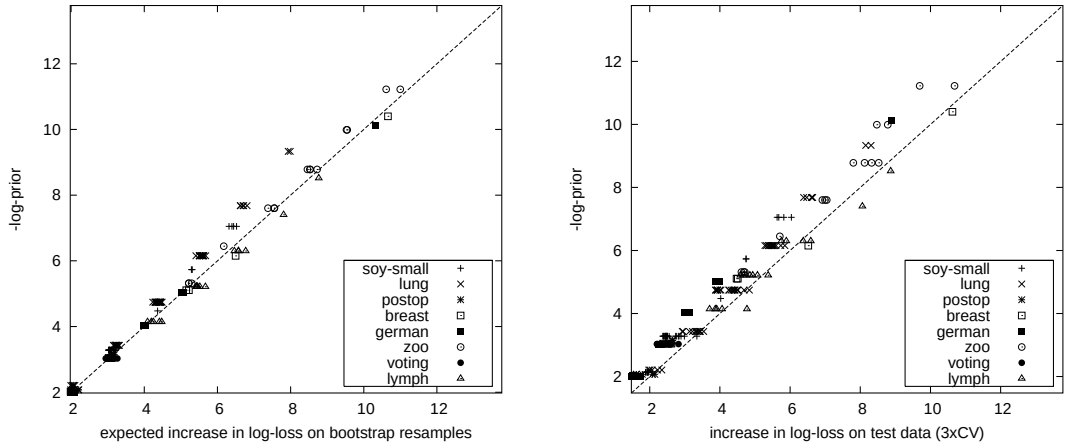


Figure 8.1: The negative logarithm of the parsimonious prior is well-aligned with the expected log-loss across bootstrap resamples in conditional density estimation (left). It is also reasonably well-calibrated with 3-fold cross-validation (right).

Because $\mathcal{M}$ was assumed to be independent of ϑ and $\mathbf{\Theta}$, we prepare $\mathbf{\Theta}$ in advance, before assessing $\mathcal{M}$. The $\hat{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \mathcal{M})$ is obtained by applying the Bayes rule on a Kikuchi approximation. A different approach to submodel fusion is conceivable, e.g., based on replacing the Kikuchi approximation with a maximum entropy one.

We can examine the resulting conditional prior empirically on several smaller benchmark domains with the number of instances in the order of magnitude of 100 and discretized attributes: 'soybean-small', 'lung', 'post-op', 'lymphography', 'german credit' and 'breast-cancer'. We have compared the posterior log-likelihood of a particular model $\mathcal{M}$, $P(\mathbf{v}^{(1)\ldots(m)}|\mathcal{M}_{\mathcal{Y}}, \vartheta = 0)P(\mathcal{M}_{\mathcal{Y}})$ with the expected total log-loss of the maximum likelihood estimates on nonparametric bootstrap resamples $\mathcal{D}^*$, across many resamples: $\mathbb{E}_{\mathcal{D}^*}\{-\sum_{\mathbf{v}^{(i)}\in\mathcal{D}^*} \log P(y^{(i)}|\mathbf{x}^{(i)})\}$. The sampling zeros were assumed to be structural zeros, i.e., if a particular attribute-class combination did not appear in the training data, it was assumed to be impossible and did not count towards the *df* (Jakulin and Bratko, 2004b). The result is shown in Fig. 8.1, and it can be seen clearly that the prior captures the increase in the variance of the loss on an independent sample.

8.3.4 Structure Search and Path Averaging

It is important to note that only those interactions that include the class label can affect the predictions. Assume that there is a certain interaction $P(\mathcal{X}')$ that does not contain any labelled attributes. If the attribute values are given $\mathbf{X} = \mathbf{x}$, the contribution of $\mathcal{X}'$ will be a constant $\psi_{\mathcal{X}'}(\mathbf{x}') = p_{X'}$ for all the values of $\mathcal{Y}$. We obtain the following expression:

$$\hat{P}(y|\mathbf{x}) = \frac{p_{X'} \prod_i \psi_i(y, \mathbf{x}_i)}{\sum_{y'} p_{X'} \prod_i \psi_i(y', \mathbf{x}_i)} = \frac{\prod_i \psi_i(y, \mathbf{x}_i)}{\sum_{y' \in \Re_Y} \prod_i \psi_i(y', \mathbf{x}_i)}$$

Therefore, for a practical classification problem with a single labelled attribute and n unlabelled attributes, there is only a single relevant submodel of order 1, $P(Y)$, n relevant submodels of order 2, $P(X_i, Y)$, and in general $\frac{n!}{k!(n-k)!} = \binom{n}{k}$ submodels of order $k+1$.

This assumption is not valid, however, in the context of semi-supervised learning and in severe cases of missing attribute values. In such situations, we can infer the values of a certain attribute X' based on other attributes, and employ this inferred information in predicting the label.

Although integrating the structure out using (8.16) is theoretically simple, we need to sample in the space of $\mathcal{M}$. It is very expensive to exhaustively survey the whole lattice of possible structures. Even if we did that, we would not be able to explain what our model is. We will adopt a simpler approach: hill-climbing. We will greedily ascend to the local maximum a posteriori structure by including the best individual region at each step, one that maximizes the posterior structure probability. Even once we get there, we keep descending for a while, as long as the structure's likelihood keeps increasing. On the termination of ascent, we *integrate out the stage of the path.* In other words, we perform Bayesian model averaging with respect to the *length* of the greedy path to the top and beyond.

This way, we obtain a compact depiction of the optimal hilltop (maximum a posteriori structure), the continuing of the path towards the dangerous peak (maximum likelihood structure). Integrating out the stage of the path prevents overconfidence in a particular structure and overfitting on the test data. Furthermore, the model average is essentially transparent and interpretable, as we can easily present the ordering of the regions as they were included into the model.

During the climb, we are guided by one-level lookahead (Buntine, 1991). This can be done efficiently with Kikuchi-Bayes using the tricks of Caruana et al. (2004): including a new region corresponds to just multiplying the approximate joint PMF with another term and renormalizing for each instance. With the considerable increase in performance that ensues, we can afford to find the best region at every step of the forward selection.

In addition to the look-ahead, we use the step-wise forward selection algorithms, normally used for loglinear models (Jobson, 1992). We first ascend by adding regions of size k attributes. When this becomes impossible, we continue ascending through addition of regions of size $k+1$ attributes. The purpose of the step-wise approach is both to increase the performance by decreasing the fanout in the search tree and to smooth the path. For example, we prevent immediately adding the initial region ABY if adding AY and BY is just as good. Still, we grow models faster than we would by only attempting unitary increases in their degrees of freedom: we skip forward by adding whole regions. In all, other search algorithms could be used, especially stochastic ones, but we should be careful as counting the same model structure multiple times would interfere with the prior.

The algorithm for greedy model search is shown as Algorithm 8.2. The end result is a sequence of ℓ pairs $\langle p_i, \mathcal{S}_i \rangle \in \mathcal{M}$. We compute the Bayesian model average (Hoeting et al., 1999) for the sequence as:

$$\hat{P}(y|\mathbf{x}) = \frac{\sum_{i=1}^{\ell} p_i \hat{P}(y|\mathbf{x}, CVA(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i))}{\sum_{i=1}^{\ell} p_i} \tag{8.23}$$

The Occam window C is used to stop the search once the new models would have an overwhelmingly small impact on the outcome. This improves the efficiency and restricts the number of models we have to average over. Otherwise, our approach to Bayesian model averaging along a greedy ascent was inspired by and has much in common with the ideas from computational learning theory, such as boosting (Freund et al., 2004), and regularization paths (Hastie et al., 2004).

$\mathcal{M} \leftarrow \emptyset$
$\mathcal{S}' \leftarrow \{Y\}$, $\phi' \leftarrow CVA(\mathcal{S}')$ {Initial model.}
$k \leftarrow 1$, $j \leftarrow 1$, $p \leftarrow 0$, $p' \leftarrow 0$
repeat
  if $\hat{P}(\phi'|\mathcal{D}) > p \wedge CP(\phi')\hat{P}(\phi'|\mathcal{D}) \geq p'$ **then** {better likelihood, a relevant posterior}
    $p \leftarrow \hat{P}(\phi'|\mathcal{D})$ {new likelihood}
    if $pP(\phi') > p'$ **then** {improvement in the posterior}
      $p' \leftarrow pP(\phi')$ {new maximum a posteriori model}
    end if
    $\mathcal{M}_j \leftarrow \langle pP(\phi'), \mathcal{S}' \rangle$ {include the best interaction of the previous iteration}
    $j \leftarrow j + 1$
    for all $\mathcal{S} \subseteq \mathcal{X}, |\mathcal{S}| = k$ **do** {examine all interactions of the present order}
      $\phi_{\mathcal{S}} \leftarrow CVA(\mathcal{M} \cup \{\mathcal{S} \cup \{Y\}\})$ {parameters obtained by including an interaction}
    end for
    $\langle \mathcal{S}', \phi' \rangle \leftarrow \arg\max_{\langle \mathcal{S}, \phi \rangle} P(\phi)\hat{P}(\phi|\mathcal{D})$ {pick the best a posteriori interaction}
  else
    $k \leftarrow k + 1$ {try interactions of higher order}
  end if
until $k > K$
return $\mathcal{M}$

Algorithm 8.2: This algorithm yields an initial set of regions with their posterior probabilities $\mathcal{M} = [\langle p_1, \mathcal{S}_1 \rangle, \ldots, \langle p_\ell, \mathcal{S}_\ell \rangle\}$ based on the data $\mathcal{D}$. The data is represented with the set of attributes $\mathcal{X}$, and the objective is to predict the label Y. The algorithm examines all interactions of the order $K + 1$ or less for possible inclusion in the predictive model. The window $C = 100$ is used to limit how far away from the best a posteriori model it is worth venturing.

The actual implementation is highly efficient, and avoids recomputing the full region-based approximation. Instead, it is possible to include a region to an existing region graph in constant time. Furthermore, we employ various caching techniques, and operate on log-probabilities. Finally, to avoid a combinatorial explosion, a certain limit is placed on the maximum number of 2-way and 3-way interactions that will be examined (in our case, 1500). The choice of which ones will be considered is based on an estimate obtained for the lower-order interactions with a calculation that resembles interaction information.

8.3.5 Examples

If our intention is to visualize the learned structure, a reasonable choice is to pick the structure with the maximum a posteriori probability. An example is shown in Fig. 8.2. Interactions of size 4 are not merely a theoretical curiosity, but identify certain intuitive symmetries and features of the tic-tac-toe board.

Although we perform a Bayesian model average, we can show the whole chain by enumerating individual interactions, and labelling them with the posterior probability of the structure. The enumeration provides a measure of importance, and the posterior structure probability an indication of what complexity is most appropriate. The posterior probability is not shown if it is lower than C^{-1}. Two examples of resulting models are shown in Fig. 8.3.

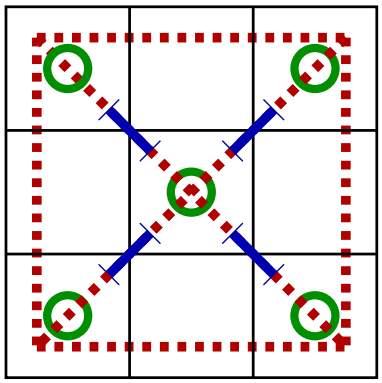


Figure 8.2: The tic-tac-toe game board comprises 9 squares, each described with a 3-valued attribute with the range $\{\times, \circ, \_\}$. The goal is to develop a predictive model that will indicate if a board position is winning for $\times$ or not: this is the 2-valued class attribute. The illustration shows the interactions of the MAP model identified by our algorithm: 2-way interactions (5 green circles), 3-way interactions (4 blue serif lines), and 4-way interactions (6 red dashed lines). Each interaction includes the label.

8.3.6 Experiments

We have taken 46 UCI data sets. The data sets were discretized with the Fayyad-Irani method and the missing values were interpreted as special values. For each data set, we performed 5 replications of 5-fold cross-validation.

The comparison included the Kikuchi-Bayes algorithm, both with model averaging (kBMA) and with the maximum a posteriori structure (kMAP). It was compared with the naïve Bayes (NB) and tree-augmented naïve Bayes (TAN), using the same Dirichlet prior for estimating the conditional probabilities. We examined multinomial logistic regression with the baseline (Agresti, 2002), as implemented in the Orange toolkit (Jakulin, 2002), and with the well-known C4.5 (Quinlan, 1993, Demšar and Zupan, 2004) as a typical representative of classification trees.

All recommendations regarding optimal performance both for SVM and logistic regression were used. For example, the most frequent class was used as the baseline, and the most frequent value of a discretized attribute was used to count towards the baseline in logistic regression. Furthermore, care was taken to prevent singularities, and a weak prior was used to prevent infinite logarithmic loss. In all, we have put a considerable effort to assure that the comparison will be fair to alternative methods. We list the results along with the commentary in Tables 8.1 through 8.6.

Simple Probabilistic Models and Calibration of the Prior

Table 8.1 shows that Kikuchi-Bayes with model averaging manages to outperform several of the today's most frequently used probabilistic classifiers: multinomial logistic regression with the baseline, tree-augmented naïve Bayes and the naïve Bayesian classifier. At the same time, Kikuchi-Bayes is efficient in spite of the fact that it follows a fully Bayesian approach by treating the structure as a nuisance variable and that it uses exhaustive lookahead in exploring the structure space: most data sets were processed in a fraction of a second, although a large number of attributes reduces the performance ('yeast class'

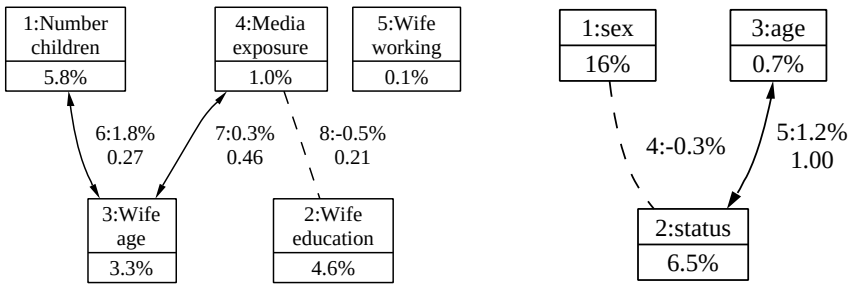


Figure 8.3: The Kikuchi-Bayes model can be visualized effectively using interaction graphs of Chapter 6. Two elements are included: the rank of an interaction, and the weight of the model in the ensemble. For example, in 'Titanic' the *sex* attribute ranks first - it is the most important attribute of survival chances, followed in that order by *status* and *age*. Further interactions involved a positive and a negative interaction with *status*. Practically all the predictive weight in the ensemble is assigned to the final model. On the other hand, in 'CMC' we interpolate between three models: 6, 7 and 8.

with 79 attributes, 'audiology' with 69 attributes). Logistic regression had the worst error rate, and TAN the worst log-loss. The log-loss performance does not correlate well with the error rate. Nevertheless, the same Kikuchi-Bayes models are ranked as the best.

Still, there are a few data sets where kBMA is inferior. A relatively large proportion of them are marked with '*' meaning that the number of degrees of freedom of the naïve Bayesian classifier exceeded the number of instances. In frequentist terms, this would indicate severe overfitting. Nevertheless, such a complex model performs quite well according to the cross-validated score. When the number of attributes exceeds 50, the performance of Kikuchi-Bayes drops. In such cases, the order of interactions should be decreased for tractability, or the attributes could be handled with more efficient methods, such as with those of Sect. 6.5.

The prior we are employing to discount the complexity is not well-calibrated with cross-validation performance, as shown in Fig. 8.4. This does not matter, though: the correction would be a constant factor. The influence of this constant factor would be eliminated in normalization. We are not using absolute posterior probabilities anywhere, we merely use them to assign weights to individual structures. Nevertheless, the final loss estimate does depend on the number of cross-validation folds.

Nevertheless, we can conclude that a combination of Kikuchi-Bayes and logistic regression almost always outperforms the naïve Bayesian classifier and the tree-augmented naïve Bayes. But, there are several domains where the Kikuchi-Bayes does not perform well. Interestingly, the nature of these domains conflicts with the assumption independently and identically distributed instances (IID): the instances are not a random sample. For example, the domain 'car' is a function that was sampled exhaustively. 'Zoo' is a list of representative animals. 'Tic-tac-toe' is a list of all possible board positions. However, cross-validation or bootstrap as evaluation protocols already enforce certain assumptions, and the learning algorithm has to match them in order to obtain good performance. So the learning algorithm does not have the freedom of choosing the assumptions: it has to match the assumptions of the evaluation protocol. The choice of the evaluation protocol

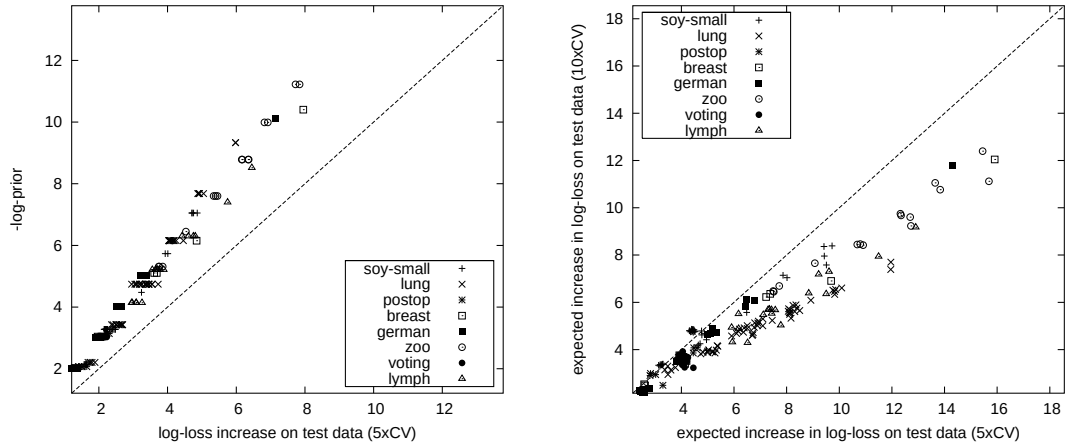


Figure 8.4: The negative logarithm of the parsimonious prior is conservatively biased in comparison to 5-fold cross-validation (left). But on the other hand, 10-fold cross-validation loss is also a biased estimate of 5-fold cross-validation loss (right). The consequence of this misalignment is that Kikuchi-Bayes is systematically excessively conservative about the complexity of its models.

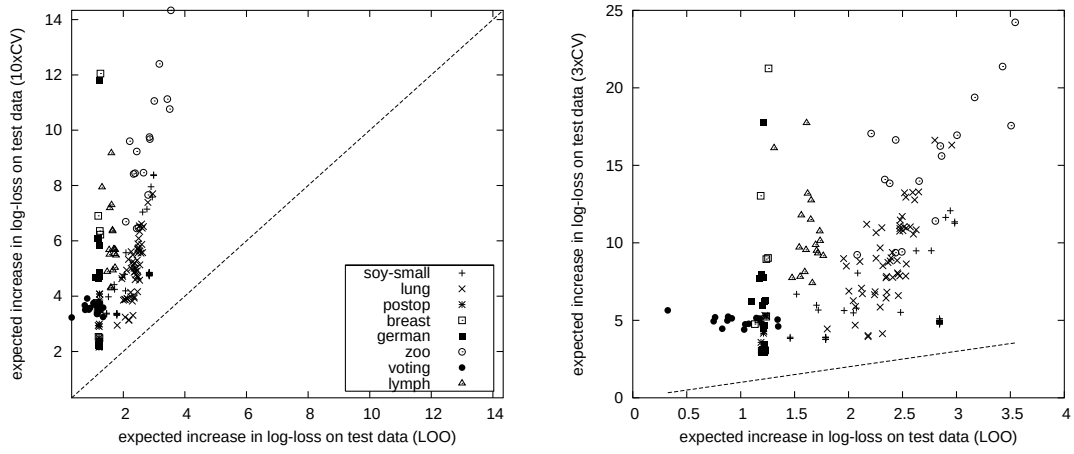


Figure 8.5: There does not seem to be much correlation between leave-one-out (LOO) estimates of expected loss and those of cross-validation with a smaller number of folds. Therefore, the efficiency of LOO estimates is of little benefit when cross-validation estimates are used in the end.

can be likened to the choice of the prior.

Does Kikuchi-Bayes Overfit?

How important are interactions of higher order? We approach this question by applying Algorithm 8.2 with different settings of K, the maximum order of interaction. $K2$ denotes the MAP Kikuchi-Bayes with $K = 2$. $K3_T$ is Kikuchi-Bayes with $K = 3$ prevented from forming patterns of interactions that cannot be factorized in an exact fashion. $K3$ and $K4$ are unrestricted with respect to the structure of interactions. The prefix b indicates that Bayesian model averaging is used instead of the MAP model in $K*$. Additionally, we include KX, the Kikuchi-Bayes model with all possible 3-way interactions involving the label.

The results are summarized in Table 8.2 for the log-loss, and in Table 8.3 for error rate. Although high-order interactions help with few data sets, they rarely hurt the performance. This implies that our method for penalizing the complexity of a model does seem to be effective. Bayesian model averaging almost consistently helps. The ability to factorize the interaction model does not seem particularly important, although it might prove to be a useful heuristic for reducing the search complexity.

The terrible performance of KX in Tables 8.2 and 8.3 indicates the importance of selecting interactions. Adding an interaction in the context of Kikuchi-Bayes may sometimes reduce the model's likelihood, not just its posterior probability. We refer to this as *approximation error* due to the use of suboptimal Kikuchi approximation. There would be no decrease in likelihood had we used MaxEnt instead, or if we only used the maximal cliques of the Markov network as initial regions.

Even if the joint approximation error did not increase when an interaction was added, the class-predictive approximation error $D(P(Y|\mathbf{X})\|\hat{P}(Y|\mathbf{X}))$ can well increase. This is noticeable already in the example of the naïve Bayesian classifier, as we have seen in Sect. 7.3.4 and in the examples of Sect. 5.3.2. The nature of the difference between the joint and the class-predictive approximation error probably lies in the problem of double-counting of evidence (Jakulin and Bratko, 2003). It has been shown that double-counting of evidence does not hurt the error rate as much as it hurts the log-loss (Domingos and Pazzani, 1997, Rish et al., 2001). As a confirmation of this observation, we can notice that NB and TAN are not as disadvantaged when error rate is used as the loss function, but this does not help them gain considerably in rank.

The Dangers of Myopic Interaction Selection

We could imagine a simple algorithm for constructing models that decides on what interactions should be included using mutual information $I(X;Y)$, or the importance of a 3-way interaction, as assessed using interaction information $I(X_1;X_2;Y)$. We could decide on a particular significance testing threshold γ, and include all interactions whose P-value is lower than γ.

Unfortunately, such an algorithm does not work well for a number of reasons:

- Interaction information is based on a non-normalized probability distribution.
- Interaction information does not directly estimate the class-predictive accuracy.

				log-loss / instance					error rate				
domain	t_K	n	df	NB	TAN	LR	kMAP	kBMA	NB	TAN	LR	kMAP	kBMA
horse-colic	1.89	369	228	1.67	·5.97	1.81	√0.83	**0.83**	**25.7**	·67.3	√35.0	√30.1	√30.6
hepatitis	0.47	155	48	0.78	·1.31	√0.77	√0.48	**0.43**	√15.6	√17.5	·19.1	**14.5**	√15.0
ionosphere	3.71	351	129	√0.64	·0.74	0.69	√0.39	**0.33**	**7.4**	√8.2	·13.6	√9.6	√9.6
vehicle	0.42	846	205	·1.78	1.14	0.93	√0.69	**0.66**	·39.6	**29.7**	√33.4	√31.4	√31.3
voting	0.23	435	48	·0.60	0.53	0.37	√0.21	**0.15**	·9.3	√7.9	√6.7	√4.6	**4.6**
monk2	0.01	601	17	0.65	0.63	·0.65	√0.45	**0.45**	38.2	36.2	·39.6	√27.6	**26.8**
p-tumor*	0.39	339	552	√3.17	·4.76	√2.76	2.65	**2.61**	**54.7**	√61.5	√63.6	·71.3	·71.3
heart	0.15	920	167	1.25	·1.53	1.24	√1.11	**1.10**	**42.8**	√44.1	·46.2	√44.8	√44.8
post-op	0.01	88	19	√0.93	·1.78	√0.81	√0.79	**0.67**	√33.4	√32.7	·34.5	**28.4**	√28.6
wdbc	0.57	569	61	0.26	0.29	·0.42	√0.15	**0.13**	√4.2	√4.4	·7.8	**4.0**	√4.1
promoters*	37.5	106	227	√0.60	·3.14	√0.70	√0.59	**0.54**	√13.4	30.4	·57.4	**10.4**	√10.6
lymph	0.39	148	94	√1.10	·1.25	√0.91	√0.98	**0.86**	√20.1	**16.1**	√23.1	·26.5	√25.7
cmc	0.04	1473	55	1.00	·1.03	0.97	√0.93	**0.92**	47.8	√45.8	·49.7	√43.6	**43.4**
adult	1.11	32561	134	·0.42	0.33	0.35	0.30	**0.30**	·16.4	14.3	**13.6**	√13.9	√13.9
crx	0.19	690	58	√0.49	·0.93	√0.39	√0.37	**0.36**	√14.1	·17.1	√14.1	**13.3**	√13.8
krkp	6.52	3196	69	·0.29	0.19	0.08	√0.06	**0.05**	·12.4	7.8	√2.5	**1.6**	√1.7
glass	0.03	214	90	√1.25	·1.76	√1.07	1.12	**1.05**	**28.3**	√29.2	√32.0	·32.1	√31.4
australian	0.16	690	49	√0.46	·0.94	√0.39	√0.41	**0.38**	√14.3	·17.6	√15.4	√14.3	**14.3**
titanic	0.01	2201	8	·0.52	√0.48	0.50	√0.48	**0.48**	·22.3	**21.1**	√22.2	**21.1**	√21.1
segment	0.74	2310	617	0.38	·1.06	0.45	**0.17**	**0.17**	√6.5	·14.2	√7.7	**5.4**	**5.4**
lenses	0.00	24	14	√2.44	·2.99	√0.89	**0.34**	0.39	√28.3	·35.8	√26.7	**12.5**	√15.0
monk1	0.01	556	16	0.50	0.09	·0.50	**0.01**	√0.02	25.4	**0.0**	·25.5	**0.0**	**0.0**
breast-LJ	0.03	286	24	√0.62	·0.89	**0.58**	√0.67	√0.58	**27.8**	√28.4	√28.3	·29.0	√28.7
monk3	0.01	554	17	·0.20	√0.11	**0.10**	√0.11	√0.11	·3.6	√1.6	√1.7	√1.1	**1.1**
bupa	0.01	345	12	·0.62	√0.60	**0.60**	√0.62	√0.61	√33.9	√32.8	·34.5	√33.2	**32.8**
tic-tac-toe	0.03	958	27	·0.55	0.49	**0.06**	√0.08	√0.07	·29.8	23.8	**2.0**	√3.1	√2.9
pima	0.02	768	19	√0.50	√0.49	**0.46**	·0.51	√0.48	√22.1	√22.1	**21.8**	·22.4	√22.0
iris	0.00	150	15	√0.27	·0.32	**0.21**	√0.27	√0.23	·6.3	√6.0	√5.6	**5.2**	**5.2**
spam	39.9	4601	156	·0.53	0.32	**0.16**	0.19	√0.19	·9.7	√6.9	**5.9**	√6.2	√6.2
breast-wisc	0.03	683	28	√0.21	·0.23	**0.13**	√0.21	√0.18	**2.6**	√3.4	√3.9	√3.9	·4.0
german	0.64	1000	68	√0.54	·1.04	**0.52**	0.65	√0.59	√24.5	·27.3	**24.4**	√26.3	√26.3
anneal	6.16	898	204	√0.07	·0.17	**0.02**	0.11	0.11	√1.3	·2.9	**0.3**	2.4	2.5
ecoli	0.01	336	92	√0.89	·0.94	**0.68**	√0.85	√0.83	**15.3**	√15.4	·16.8	√16.4	√16.2
hayes-roth	0.00	160	24	0.46	·1.18	**0.26**	0.45	0.45	√14.9	·29.9	√17.0	**13.5**	**13.5**
balance-scale	0.00	625	40	0.51	·1.13	**0.28**	0.51	0.51	√9.3	·15.0	**8.5**	√9.3	√9.3
soy-large*	5.95	683	822	√0.57	√0.47	**0.37**	·0.68	0.68	√9.0	√8.4	**7.7**	·27.0	27.0
o-ring	0.00	23	7	√0.83	√0.76	**0.66**	·1.41	√1.00	**13.0**	·22.6	√17.4	√22.6	√19.1
lung-cancer*	35.0	32	233	5.41	·6.92	**1.24**	√2.37	√1.62	**51.9**	√63.8	·70.6	√60.6	√61.9
audiology*	81.2	226	1783	3.55	·5.56	**1.40**	2.24	2.23	√40.8	62.7	**26.0**	·68.6	·68.6
soy-small*	5.29	47	115	√0.00	**0.00**	·0.15	0.00	0.00	**0.0**	**0.0**	·2.1	**0.0**	**0.0**
mushroom	1.33	8124	72	·0.01	**0.00**	0.00	0.00	0.00	·0.4	**0.0**	**0.0**	√0.0	√0.0
shuttle	0.01	253	15	·0.16	**0.06**	√0.10	√0.07	√0.07	·6.7	√2.8	**2.5**	√3.6	√2.9
car	0.02	1728	48	0.32	**0.18**	·0.33	0.19	0.19	14.6	**5.9**	·16.7	√6.5	√6.5
zoo*	0.23	101	124	**0.38**	·0.46	√0.38	√0.40	√0.40	**3.6**	√6.3	√7.5	·12.9	√12.1
wine	0.10	178	50	**0.06**	·0.29	√0.09	√0.19	√0.14	**0.9**	√3.1	√2.2	·4.3	√3.6
yeast-class*	138	186	376	**0.01**	√0.03	·0.90	0.25	0.23	**0.1**	√0.3	·34.9	√2.9	√2.9
avg rank				3.68	·3.99	√2.54	2.84	**1.95**	2.98	3.20	·3.34	√2.87	**2.62**

Table 8.1: A comparison of Kikuchi-Bayes with the maximum a posteriori structure (kMAP) and with Bayesian model averaging (kBMA), logistic regression with the baseline (LR), naïve Bayesian classifier (NB), and the tree-augmented naïve Bayes (TAN). The best result is typeset in bold, and the results of those methods that matched or outperformed the best method in at least 2 of the 25 experiments are tagged with √. df are the degrees of freedom of the ordinary naïve Bayesian classifier, and n is the number of instances. The sparse data sets with fewer instances than degrees of freedom are tagged with '*'. t_K marks the time in seconds spent by the Kikuchi-Bayes algorithm for learning the model structure from k-way interactions, $k \leq 4$, on a contemporary notebook computer. The domains are sorted with respect to what method performed best.

	log-loss / instance										
domain	NB	TAN	KX	$K2$	$K3_T$	$K3$	$K4$	$bK2$	$bK3_T$	$bK3$	$bK4$
adult	0.42	0.33	*·0.67*	0.31	0.30	0.30	0.30	0.31	√0.30	√0.30	**0.30**
krkp	*·0.29*	0.19	0.25	0.26	0.11	0.08	√0.06	0.26	0.11	0.08	**0.05**
monk2	0.65	0.63	0.51	*·0.65*	0.60	0.54	√0.45	0.65	0.60	0.53	**0.45**
spam	0.53	0.32	*·3.74*	√0.21	√0.19	√0.19	√0.19	√0.21	√0.19	√0.19	**0.19**
tic-tac-toe	*·0.55*	0.49	0.41	0.53	0.53	0.42	√0.08	0.53	0.52	0.42	**0.07**
titanic	0.52	√0.48	√0.48	*·0.52*	√0.48	√0.48	√0.48	0.52	√0.48	√0.48	**0.48**
glass	√1.25	√1.76	*·4.59*	1.12	1.12	1.12	1.12	√1.05	√1.05	**1.05**	**1.05**
heart	1.25	1.53	*·2.77*	√1.10	√1.11	√1.11	√1.11	√1.10	√1.10	**1.10**	**1.10**
horse-colic	1.67	*·5.97*	5.67	√0.83	√0.83	√0.83	√0.83	√0.83	√0.83	**0.83**	**0.83**
iris	√0.27	√0.32	*·2.87*	√0.27	√0.27	√0.27	√0.27	√0.23	√0.23	**0.23**	**0.23**
lymph	√1.10	√1.25	√1.98	√0.98	√0.98	√0.98	√0.98	√0.86	√0.86	**0.86**	**0.86**
p-tumor*	√3.17	*·4.76*	2.99	2.65	2.65	2.65	2.65	√2.61	√2.61	**2.61**	**2.61**
promoters*	√0.60	3.14	*·3.15*	√0.59	√0.59	√0.59	√0.59	√0.54	√0.54	**0.54**	**0.54**
vehicle	1.78	1.14	*·4.98*	√0.82	√0.69	√0.69	√0.69	√0.81	√0.66	**0.66**	**0.66**
audiology*	√3.55	*·5.56*	3.11	√2.24	√2.24	√2.24	√2.24	**2.23**	**2.23**	**2.23**	**2.23**
segment	0.38	1.06	*·1.79*	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**
cmc	1.00	1.03	*·1.24*	√0.93	√0.93	√0.93	√0.93	√0.93	**0.92**	√0.92	√0.92
ionosphere	√0.64	0.74	*·3.93*	√0.38	√0.39	√0.39	√0.39	√0.34	**0.33**	√0.33	√0.33
wdbc	0.26	0.29	*·1.16*	√0.14	√0.15	√0.15	√0.15	√0.13	**0.13**	√0.13	√0.13
ecoli	√0.89	√0.94	*·2.71*	√0.85	√0.85	√0.85	√0.85	**0.83**	**0.83**	√0.83	√0.83
lung-cancer*	5.41	6.92	*·8.26*	√2.37	√2.37	√2.37	√2.37	**1.62**	**1.62**	√1.62	√1.62
australian	√0.46	0.94	*·1.45*	√0.37	√0.39	√0.39	√0.41	**0.36**	√0.38	√0.38	√0.38
breast-LJ	√0.62	√0.89	*·1.06*	√0.57	√0.67	√0.67	√0.67	**0.56**	√0.58	√0.58	√0.58
breast-wisc	√0.21	√0.23	*·1.86*	√0.17	√0.21	√0.21	√0.21	**0.17**	√0.18	√0.18	√0.18
crx	√0.49	0.93	*·1.22*	√0.36	√0.37	√0.37	√0.37	**0.35**	√0.36	√0.36	√0.36
german	√0.54	1.04	*·1.54*	√0.53	0.64	0.64	0.65	**0.53**	√0.59	√0.59	√0.59
hepatitis	√0.78	1.31	*·2.03*	√0.48	√0.48	√0.48	√0.48	**0.43**	√0.43	√0.43	√0.43
post-op	√0.93	√1.78	*·2.70*	√0.79	√0.79	√0.79	√0.79	**0.67**	√0.67	√0.67	√0.67
voting	0.60	0.53	*·3.23*	√0.16	√0.21	√0.21	√0.21	**0.15**	√0.15	√0.15	√0.15
balance-scale	**0.51**	1.13	*·1.27*	√0.51	√0.51	**0.51**	**0.51**	**0.51**	√0.51	√0.51	√0.51
monk1	*·0.50*	0.09	0.10	0.49	√0.08	√0.08	**0.01**	0.49	0.08	0.08	√0.02
hayes-roth	√0.46	1.18	*·1.71*	**0.45**	**0.45**	**0.45**	**0.45**	√0.45	√0.45	√0.45	√0.45
lenses	√2.44	2.99	*·3.87*	**0.34**	**0.34**	**0.34**	**0.34**	0.39	0.39	0.39	0.39
pima	√0.50	√0.49	√0.56	**0.48**	√0.49	√0.49	√0.51	√0.48	√0.48	√0.48	√0.48
monk3	√0.20	0.11	**0.10**	√0.20	√0.11	√0.11	√0.11	√0.20	√0.11	√0.11	√0.11
shuttle	0.16	√0.06	**0.04**	0.17	√0.07	√0.07	√0.07	*·0.17*	√0.07	√0.07	√0.07
bupa	√0.62	**0.60**	√0.62	√0.62	√0.61	√0.62	√0.62	√0.62	√0.61	√0.61	√0.61
car	0.32	**0.18**	*·2.62*	0.32	0.19	0.19	0.19	0.32	0.19	0.19	0.19
mushroom	0.01	**0.00**	*·0.03*	0.01	0.00	0.00	0.00	0.01	0.00	0.00	0.00
o-ring	√0.83	**0.76**	√0.81	√1.41	√1.41	√1.41	√1.41	√0.99	√0.99	√1.00	√1.00
soy-large*	√0.57	**0.47**	*·2.88*	√0.68	√0.68	√0.68	√0.68	√0.68	√0.68	√0.68	√0.68
soy-small*	√0.00	**0.00**	*·1.10*	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
anneal	**0.07**	√0.17	*·1.39*	0.12	√0.11	√0.11	√0.11	0.12	√0.11	√0.11	√0.11
wine	**0.06**	√0.29	*·7.30*	√0.19	√0.19	√0.19	√0.19	√0.14	√0.14	√0.14	√0.14
yeast-class*	**0.01**	√0.03	*·1.27*	0.25	0.25	0.25	0.25	0.23	0.23	0.23	0.23
zoo*	**0.38**	√0.46	*·1.79*	√0.40	√0.40	√0.40	√0.40	√0.40	√0.40	√0.40	√0.40
avg rank	7.79	7.88	*·9.65*	6.57	6.26	6.03	5.85	5.09	4.05	√3.61	**3.22**

Table 8.2: We have noticed no deterioration by increasing the maximum initial region size, so the prior effectively prevents overfitting: the rank of classifiers with an increasing size of interactions is decreasing monotonically. However, attempting the inclusion of large regions is sometimes futile: the initial regions of size 3 or even just 2 were perfectly sufficient in many natural data sets. Although these higher-order interactions are relatively rare in real-life data (they only appear in the first 6 data sets and in 'monk1'), we should have the capacity to handle them.

	error rate										
domain	NB	TAN	KX	$K2$	$K3_T$	$K3$	$K4$	$bK2$	$bK3_T$	$bK3$	$bK4$
adult	16.4	√14.3	*·20.9*	14.6	√13.9	√13.9	√13.9	14.6	√13.9	√13.9	**13.9**
monk2	*·38.2*	36.2	√27.6	34.3	35.3	√30.2	√27.6	34.3	35.3	√29.7	**26.8**
balance-scale	**9.3**	15.0	*·17.5*	**9.3**	**9.3**	**9.3**	**9.3**	**9.3**	**9.3**	**9.3**	**9.3**
cmc	47.8	√45.8	*·50.0*	√45.1	√43.6	√43.6	√43.6	√45.3	**43.4**	**43.4**	**43.4**
hayes-roth	√14.9	29.9	*·33.6*	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**	**13.5**
iris	√6.3	√6.0	*·56.7*	**5.2**	**5.2**	**5.2**	**5.2**	**5.2**	**5.2**	**5.2**	**5.2**
monk1	*·25.4*	**0.0**	**0.0**	25.4	**0.0**	**0.0**	**0.0**	25.4	**0.0**	**0.0**	**0.0**
monk3	*·3.6*	√1.6	√1.2	*·3.6*	√1.1	√1.1	√1.1	*·3.6*	√1.1	**1.1**	**1.1**
segment	√6.5	14.2	*·85.0*	**5.4**	**5.4**	**5.4**	**5.4**	**5.4**	**5.4**	**5.4**	**5.4**
soy-small*	**0.0**	**0.0**	*·63.8*	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**
tic-tac-toe	*·29.8*	23.8	19.4	27.8	26.6	20.8	√3.1	27.8	26.7	20.4	**2.9**
australian	√14.3	√17.6	*·30.5*	√14.6	√14.1	√14.1	√14.3	√14.6	√14.1	**14.1**	√14.3
spam	9.7	√6.9	*·38.2*	√6.9	√6.3	√6.2	√6.2	√6.9	**6.2**	√6.2	√6.2
pima	√22.1	√22.1	√25.5	**21.7**	√22.2	√22.2	√22.4	**21.7**	√22.0	√22.1	√22.0
voting	9.3	√7.9	*·30.3*	**4.4**	√4.6	√4.6	√4.6	**4.4**	√4.6	√4.6	√4.6
wdbc	√4.2	√4.4	*·10.9*	√4.0	√4.0	√4.0	√4.0	**3.9**	√4.0	√4.1	√4.1
crx	√14.1	√17.1	*·23.4*	√14.4	√13.7	**13.3**	**13.3**	√14.3	√13.9	√13.7	√13.8
hepatitis	√15.6	√17.5	√20.6	**14.5**	**14.5**	**14.5**	**14.5**	√15.4	√15.0	√15.0	√15.0
krkp	*·12.4*	7.8	6.5	7.6	3.4	√2.4	**1.6**	7.7	3.4	√2.3	√1.7
lenses	√28.3	√35.8	√60.0	**12.5**	**12.5**	**12.5**	**12.5**	√15.0	√15.0	√15.0	√15.0
post-op	√33.4	√32.7	√39.5	**28.4**	**28.4**	**28.4**	**28.4**	√28.6	√28.6	√28.6	√28.6
promoters*	√13.4	*·30.4*	√27.4	**10.4**	**10.4**	**10.4**	**10.4**	√10.6	√10.6	√10.6	√10.6
bupa	√33.9	√32.8	**31.4**	√34.0	√33.5	√33.2	√33.2	√34.6	√32.6	√32.4	√32.8
shuttle	√6.7	√2.8	**2.0**	√6.7	√3.6	√3.6	√3.6	√6.7	√2.9	√2.9	√2.9
titanic	√22.3	√21.1	**21.0**	√22.3	√21.1	√21.1	√21.1	√22.3	√21.1	√21.1	√21.1
car	14.6	**5.9**	*·63.8*	14.9	√6.5	√6.5	√6.5	14.9	√6.5	√6.5	√6.5
lymph	√20.1	**16.1**	*·49.3*	√26.5	√26.5	√26.5	√26.5	√25.7	√25.7	√25.7	√25.7
mushroom	*·0.4*	**0.0**	0.3	√0.0	√0.0	√0.0	√0.0	√0.0	√0.0	√0.0	√0.0
soy-large*	√9.0	**8.4**	*·89.3*	27.0	27.0	27.0	27.0	27.0	27.0	27.0	27.0
vehicle	39.6	**29.7**	*·50.5*	36.3	√31.4	√31.4	√31.4	36.2	√31.3	√31.3	√31.3
anneal	**1.3**	√2.9	*·24.0*	√2.8	√2.4	√2.4	√2.4	√2.8	√2.4	√2.5	√2.5
audiology*	**40.8**	62.7	*·75.8*	68.6	68.6	68.6	68.6	68.6	68.6	68.6	68.6
breast-LJ	**27.8**	√28.4	√33.6	√29.7	√29.0	√29.0	√29.0	√29.5	√28.7	√28.7	√28.7
breast-wisc	**2.6**	√3.4	*·19.8*	√3.9	√3.9	√3.9	√3.9	√3.9	√4.0	√4.0	√4.0
ecoli	**15.3**	√15.4	*·82.3*	√16.4	√16.4	√16.4	√16.4	√16.2	√16.2	√16.2	√16.2
german	**24.5**	√27.3	*·29.2*	√25.4	√26.4	√26.3	√26.3	√25.3	√26.2	√26.4	√26.3
glass	**28.3**	√29.2	*·68.8*	√32.1	√32.1	√32.1	√32.1	√31.4	√31.4	√31.4	√31.4
heart	**42.8**	√44.1	*·51.0*	√44.7	√44.8	√44.8	√44.8	√44.7	√44.8	√44.8	√44.8
horse-colic	**25.7**	*·67.3*	56.3	√30.1	√30.1	√30.1	√30.1	√30.6	√30.6	√30.6	√30.6
ionosphere	**7.4**	√8.2	*·34.9*	√9.6	√9.6	√9.6	√9.6	√9.9	√9.6	√9.6	√9.6
lung-cancer*	**51.9**	√63.8	√71.9	√60.6	√60.6	√60.6	√60.6	√61.9	√61.9	√61.9	√61.9
o-ring	**13.0**	√22.6	√21.7	√22.6	√22.6	√22.6	√22.6	√19.1	√19.1	√19.1	√19.1
p-tumor*	**54.7**	√61.5	*·75.2*	71.3	71.3	71.3	71.3	71.3	71.3	71.3	71.3
wine	**0.9**	√3.1	*·67.9*	√4.3	√4.3	√4.3	√4.3	√3.6	√3.6	√3.6	√3.6
yeast-class*	**0.1**	√0.3	*·11.1*	√2.9	√2.9	√2.9	√2.9	√2.9	√2.9	√2.9	√2.9
zoo*	**3.6**	√6.3	*·59.4*	√12.9	√12.9	√12.9	√12.9	√12.1	√12.1	√12.1	√12.1
avg rank	6.00	6.14	*·9.51*	6.78	5.74	5.49	5.30	6.51	√4.86	4.88	**4.78**

Table 8.3: The naïve Bayes and tree augmented naïve Bayes are not as disadvantaged when the quality of a classifier is evaluated using the error rate. Nevertheless, Kikuchi-Bayes outranks them with a margin. We can notice that error rate is a relatively 'noisy' loss function: even when $bK4$ was not the best, it nevertheless outperformed the winning classifier in at least some of the experiments: $bK4$ is not √-tagged only in two domains ('p-tumor' and 'audiology').

- Significance testing is not directly concerned with maximizing the expected predictive accuracy of the classifier. Significance only indicates the reliability of a probability estimate, not the reduction in approximation error.
- Myopic interaction selection disregards the approximation error due to overlapping interactions.
- Selection of interactions based on the whole model can indirectly manage the approximation error.

It is interesting to examine the reasons for failure of interaction selection based on significance testing on a few domains where the differences between methods are most accentuated. We will disregard the generalization error and focus on the approximation error assessed on the training set itself. Furthermore, we will employ the class-predictive loss in assessing the interaction. For example, the following definition of a class-predictive P-value will be used:

$$\gamma = \Pr\left\{ D\left(P^*(Y|\mathbf{X}_\mathcal{S}) \| P(Y_\mathcal{S}|\mathbf{X}_\mathcal{S})\right) \geq D(P(Y|\mathbf{X}_\mathcal{S}) \| \hat{P}(Y_\mathcal{S}|\mathbf{X}_\mathcal{S})) \right\}$$

Here, P^* means an independently drawn resample, P the 'true' model, and $\hat{P}$ the no-interaction model. We can estimate the P-value using bootstrap, as we have seen in Sect. 4.2.1. The KL-divergence in our case is such:

$$D(P(Y|\mathbf{X}_\mathcal{S}) \| \hat{P}(Y|\mathbf{X}_\mathcal{S})) = \sum_{\mathbf{x}_\mathcal{S}} \sum_{y} P(\mathbf{x}_\mathcal{S}, y) \log_2 \frac{P(\mathbf{x}_\mathcal{S}, y) \sum_{y'} \hat{P}(\mathbf{x}_\mathcal{S}, y')}{P(\mathbf{x}_\mathcal{S}) \hat{P}(\mathbf{x}_\mathcal{S}, y)}$$

We plotted the performance of $K2$, $K3$ and $K3_T$ at various levels of the significance testing parameter γ. Figure 8.6 illustrates 'tic-tac-toe', where KX obtains solid performance: the restrictions of $K1$ and $K3_T$ prevent the utilization of these pairs.

Highly significant interactions deteriorate the performance in 'voting': after the initial gains (based on two 2-way interactions and one 3-way interaction), additional interactions result in large approximation errors. The factorizable model of $K3_T$ is better than $K3$ at any setting of γ, but adding interactions into the tree may also increase the approximation error.

On the other hand, 'segment' shows the opposite situation: 3-way interactions cause major approximation errors in spite of highly significant clusters. This domain has 2310 instances, and many interactions are highly significant. Unfortunately, significance only indicates the reliability of the probability estimate, not the reduction in estimation error, and the local learning neglects the additional approximation error. We can see that the curve of the $K3$ is not monotonically decreasing either, and the last leaf of the tree causes a deterioration.

'Pima' is a particularly well-behaved domain. The structure of interactions is rather acyclic and $K3_T$ and $K3$ largely overlap. In the end, $K3$ has the potential of gain, but, if unrestrained, also slight deterioration. In comparison, the deterioration in 'voting' is much more distinct. As in 'segment', the tree-based model is better than $K3$ at any setting of γ. It is incorrect to view γ as a domain-dependent tuning parameter: γ does affect how many clusters will get included, but it has meaning of its own that should remain unrelated to the issue of approximation error. Modelling the approximation error in Kikuchi-Bayes remains an open issue.

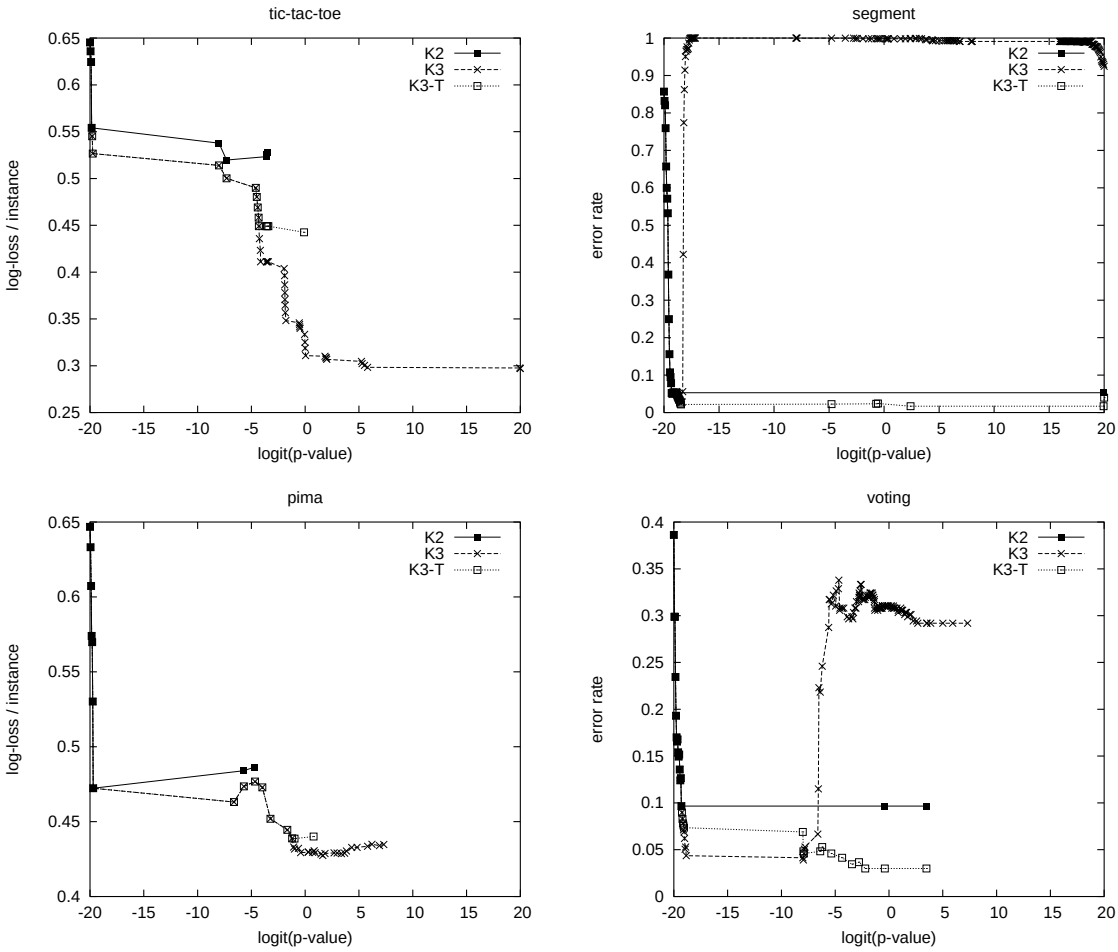


Figure 8.6: The classification performance depends on the significance testing threshold γ. The horizontal scale indicates the logit-transformed value of the P-value threshold $\log(\gamma/(1-\gamma))$ used as a parameter for Kikuchi-Bayes learning.

Comparisons with Support Vector Machines

Support vector machines (Vapnik, 1999, Schölkopf and Smola, 2002) are often acknowledged as the state-of-the-art in machine learning. Although there are disadvantages to them, such as inefficiency, abundance of parameters, black-box models, dependence on the data representation, they achieve excellent performance overall. There are several theories for their success, but probably the key advantage is that SVM learning algorithms actually attempt to maximize the classification performance and not an indirect proxy, such as maximum likelihood. The noisiness of the error rate is tackled by an appropriate formulation of the optimization problem.

Our comparison included two state-of-the-art SVM implementations: $SVM^{multiclass}$ V1.01 (Tsochantaridis et al., 2004, Crammer and Singer, 2001, Joachims, 1999), and LIBSVM V2.8 (Fan et al., 2005, Chang and Lin, 2005). A k-valued discrete attribute was represented with a k-dimensional vector for the SVM classifier, so that the active attribute value was placed at $+1$, and the inactive attribute value at -1: this method yielded the best results in our experiments. We have not, however, attempted to execute internal cross-validation to optimize the parameter values. Instead, we have used the default parameters as hardwired in the implementations.

Our first experiment in Table 8.4 indicates that $SVM^{multiclass}$ with the dot product kernel unsurprisingly outperforms the Kikuchi-Bayes algorithm in error rate. But surprisingly, the versatile radial basis function (RBF) kernel does not outperform Kikuchi-Bayes, not even the MAP version. What is interesting is that the dot product kernel, a method that assumes linear separability, manages to be the best ranked method. Furthermore, it is surprising that Kikcuhi-Bayes with its simplicity and disregard for error rate nevertheless outperforms C4.5, an algorithm that was tuned for the UCI collection of data sets.

In later experiments (Tables 8.5 and 8.6), we included the latest version of LIBSVM, which yields excellent results, but also involves a decrease in performance as compared to $SVM^{multiclass}$. LIBSVM also supports probability estimation, so we can assess its log-loss as well. Interestingly, LIBSVM's performance is better using the RBF kernel than using the linear kernel. $SVM^{multiclass}$ has a hard time competing against LIBSVM.

It has to be stressed that the naïve Bayesian classifier, logistic regression and SVM with the dot product kernel have the same representation if the label is binary. They only differ with respect to the criterion used for fitting that representation. Especially for log-loss, it seems that our own implementation of logistic regression could benefit from better regularization: LIBSVM is distinctly more mature.

A comparison of the timings for Kikuchi-Bayes and LIBSVM with the dot product kernel shows that the two methods's performance follows a different characteristic: SVM running time is primarily tied to the number of instances, whereas Kikuchi-Bayes running time is primarily correlated with the number of attributes. A hybrid method could take the best of both worlds: it would combine the interpretability of Kikuchi-Bayes models with the utility of optimization methods for fusing submodels without approximation loss. The path towards integration has already been set by the formulations of Altun et al. (2004) and Lafferty et al. (2004). Kikuchi-Bayes can be used as efficient means to learning the structure of kernels, but the fusion can be then performed using the convex optimization. Furthermore, classification trees could be used instead of the multinomial models we use for submodels.

			error rate					
domain	t_K	t_{SVM}	SVM_{dot}	SVM_{rbf}	C4.5	LR	kMAP	kBMA
iris	0.00	0.09	**5.2**	√6.4	√5.3	√5.6	**5.2**	**5.2**
soy-small*	5.29	0.11	**0.0**	**0.0**	√0.9	√2.1	**0.0**	**0.0**
monk3	0.01	0.19	*·3.6*	√2.6	**1.1**	√1.7	√1.1	**1.1**
bupa	0.01	0.20	√37.2	√33.7	√33.8	√34.5	√33.2	**32.8**
car	0.02	1.68	16.3	9.4	√8.9	*·16.7*	√6.5	**6.5**
cmc	0.04	3.21	√45.4	48.6	√45.6	*·49.7*	√43.6	**43.4**
hayes-roth	0.00	0.15	22.4	*·30.4*	24.9	√17.0	**13.5**	**13.5**
monk1	0.01	0.29	25.4	√2.2	√2.8	*·25.5*	**0.0**	**0.0**
voting	0.23	0.24	√4.6	√7.7	√5.0	√6.7	√4.6	**4.6**
crx	0.19	0.69	√14.6	√16.6	**13.3**	√14.1	**13.3**	√13.8
lenses	0.00	0.05	√19.2	√21.7	√16.7	√26.7	**12.5**	√15.0
titanic	0.01	0.80	√22.4	√21.8	√21.3	√22.2	**21.1**	√21.1
anneal	6.16	1.96	2.0	√1.0	√1.4	**0.3**	2.4	*·2.5*
balance-scale	0.00	0.50	√8.5	13.1	*·34.1*	**8.5**	√9.3	√9.3
german	0.64	1.43	√24.5	√27.4	√27.7	**24.4**	√26.3	√26.3
mushroom	1.33	7.57	√0.1	**0.0**	**0.0**	**0.0**	√0.0	√0.0
pima	0.02	0.40	√23.2	√24.3	√23.1	**21.8**	√22.4	√22.0
shuttle	0.01	0.11	√6.7	√4.3	√3.2	**2.5**	√3.6	√2.9
tic-tac-toe	0.03	0.73	13.5	√5.8	*·14.8*	**2.0**	√3.1	√2.9
audiology*	81.2	17.0	√25.8	√24.2	**23.4**	√26.0	*·68.6*	*·68.6*
australian	0.16	0.60	√14.6	√16.5	**13.4**	√15.4	√14.3	√14.3
breast-LJ	0.03	0.20	√28.0	√26.9	**26.8**	√28.3	√29.0	√28.7
krkp	6.52	4.02	*·6.2*	3.4	**0.7**	2.5	1.6	1.7
ecoli	0.01	0.34	√17.8	**15.6**	√15.8	√16.8	√16.4	√16.2
glass	0.03	0.28	√31.0	**26.4**	√30.6	√32.0	√32.1	√31.4
lung-cancer*	35.0	0.22	√57.5	**56.3**	√59.4	√70.6	√60.6	√61.9
monk2	0.01	0.35	34.3	**23.3**	36.6	*·39.6*	27.6	√26.8
vehicle	0.42	2.61	√29.4	**28.9**	√29.0	√33.4	√31.4	√31.3
zoo*	0.23	0.21	√9.1	**3.2**	√7.7	√7.5	√12.9	√12.1
yeast-class*	138	0.72	**0.0**	**0.0**	√3.8	*·34.9*	√2.9	√2.9
adult	1.11	436	**13.3**	*·20.0*	13.8	√13.6	13.9	13.9
breast-wisc	0.03	0.23	**2.4**	√2.8	√4.6	√3.9	√3.9	√4.0
heart	0.15	3.46	**42.2**	√44.5	√45.9	√46.2	√44.8	√44.8
hepatitis	0.47	0.15	**13.7**	√17.9	√20.9	√19.1	√14.5	√15.0
horse-colic	1.89	1.68	**28.3**	√30.5	√36.8	√35.0	√30.1	√30.6
ionosphere	3.71	0.57	**7.5**	√7.9	√11.2	√13.6	√9.6	√9.6
lymph	0.39	0.21	**16.8**	√18.5	√23.5	√23.1	√26.5	√25.7
o-ring	0.00	0.05	**13.9**	√17.4	√20.0	√17.4	√22.6	√19.1
p-tumor*	0.39	6.07	**54.8**	√60.6	√57.9	63.6	*·71.3*	*·71.3*
post-op	0.01	0.08	**28.0**	√33.2	√29.5	√34.5	√28.4	√28.6
promoters*	37.5	0.47	**9.6**	√18.3	√23.2	*·57.4*	√10.4	√10.6
segment	0.74	26.1	**5.0**	√6.0	√6.2	*·7.7*	√5.4	√5.4
soy-large*	5.95	11.0	**6.6**	√7.4	√7.0	√7.7	*·27.0*	27.0
spam	39.9	13.2	**5.8**	*·29.9*	√7.1	√5.9	√6.2	√6.2
wdbc	0.57	0.33	**2.3**	√3.1	√4.0	*·7.8*	√4.0	√4.1
wine	0.10	0.15	**1.6**	√3.1	√6.2	√2.2	√4.3	√3.6
avg rank			**2.97**	3.49	3.70	*·4.16*	3.46	√3.23

Table 8.4: Support vector machines, using the $SVM^{multiclass}$ implementation, outperform Kikuchi-Bayes when an appropriate kernel is used. There is an interesting trade-off involving the time consumption of SVM versus Kikuchi-Bayes: Kikuchi-Bayes is extremely fast even when there are many instances, but is slow when there are many attributes ('audiology'). On the other hand, SVM copes with a large number of attributes, but becomes highly inefficient with a large number of instances ('adult').

error rate	LIBSVM		$SVM^{multiclass}$					
domain	SVM^{p}_{dot}	SVM^{p}_{rbf}	SVM_{dot}	SVM_{rbf}	C4.5	LR	NB	kBMA
cmc	√46.0	√44.6	√45.4	48.6	√45.6	·49.7	47.8	**43.4**
hayes-roth	√14.5	√13.8	22.4	·30.4	24.9	√17.0	√14.9	**13.5**
lenses	√29.2	√40.0	√19.2	√21.7	√16.7	√26.7	√28.3	**15.0**
monk1	25.4	**0.0**	25.4	√2.2	√2.8	·25.5	25.4	**0.0**
monk3	**1.1**	·3.6	·3.6	√2.6	**1.1**	√1.7	·3.6	**1.1**
soy-small*	**0.0**	**0.0**	**0.0**	**0.0**	√0.9	√2.1	**0.0**	**0.0**
titanic	√22.4	√21.2	√22.4	√21.8	√21.3	√22.2	√22.3	**21.1**
voting	√4.7	√4.6	√4.6	√7.7	√5.0	√6.7	·9.3	**4.6**
horse-colic	√30.4	√27.3	√28.3	√30.5	·36.8	√35.0	**25.7**	√30.6
ionosphere	√9.7	√8.3	√7.5	√7.9	√11.2	√13.6	**7.4**	√9.6
lung-cancer*	√56.9	√70.0	√57.5	√56.3	√59.4	√70.6	**51.9**	√61.9
o-ring	√22.6	√13.0	√13.9	√17.4	√20.0	√17.4	**13.0**	√19.1
wine	√1.5	√1.3	√1.6	√3.1	√6.2	√2.2	**0.9**	√3.6
anneal	√0.5	·3.0	2.0	√1.0	√1.4	**0.3**	√1.3	2.5
mushroom	**0.0**	√0.1	√0.1	**0.0**	**0.0**	**0.0**	·0.4	√0.0
pima	√22.7	√22.5	√23.2	√24.3	√23.1	**21.8**	√22.1	√22.0
shuttle	√2.7	√6.6	√6.7	√4.3	√3.2	**2.5**	√6.7	√2.9
audiology*	√24.9	√30.5	√25.8	√24.2	**23.4**	√26.0	√40.8	·68.6
australian	√15.6	√14.5	√14.6	√16.5	**13.4**	√15.4	√14.3	√14.3
breast-LJ	√29.2	√26.9	√28.0	√26.9	**26.8**	√28.3	√27.8	√28.7
crx	√13.4	√14.5	√14.6	√16.6	**13.3**	√14.1	√14.1	√13.8
krkp	3.3	6.1	6.2	3.4	**0.7**	2.5	·12.4	1.7
monk2	34.3	33.2	34.3	**23.3**	36.6	·39.6	38.2	√26.8
yeast-class*	**0.0**	√0.4	**0.0**	**0.0**	√3.8	·34.9	√0.1	√2.9
zoo*	√6.3	√9.9	√9.1	**3.2**	√7.7	√7.5	√3.6	√12.1
breast-wisc	√3.5	√2.5	**2.4**	√2.8	√4.6	√3.9	√2.6	√4.0
hepatitis	√15.9	√14.8	**13.7**	√17.9	√20.9	√19.1	√15.6	√15.0
lymph	√21.8	√18.0	**16.8**	√18.5	√23.5	√23.1	√20.1	√25.7
wdbc	√3.0	√2.8	**2.3**	√3.1	√4.0	·7.8	√4.2	√4.1
bupa	√34.2	**32.1**	√37.2	√33.7	√33.8	√34.5	√33.9	√32.8
car	7.2	**4.9**	16.3	9.4	8.9	·16.7	14.6	√6.5
ecoli	√14.7	**13.2**	√17.8	√15.6	√15.8	√16.8	√15.3	√16.2
german	√24.9	**24.1**	√24.5	√27.4	√27.7	√24.4	√24.5	√26.3
glass	√25.6	**24.3**	√31.0	√26.4	√30.6	√32.0	√28.3	√31.4
heart	√42.2	**41.3**	√42.2	√44.5	45.9	·46.2	√42.8	44.8
iris	√5.6	**4.9**	√5.2	√6.4	√5.3	√5.6	√6.3	√5.2
p-tumor*	√54.5	**53.9**	√54.8	60.6	√57.9	63.6	√54.7	·71.3
post-op	**27.3**	**27.3**	√28.0	√33.2	√29.5	√34.5	√33.4	√28.6
promoters*	√9.2	**7.2**	√9.6	√18.3	23.2	·57.4	√13.4	√10.6
spam	√5.9	**5.8**	√5.8	·29.9	7.1	√5.9	9.7	√6.2
adult	**13.1**	13.9	√13.3	·20.0	13.8	13.6	16.4	13.9
balance-scale	**5.6**	√8.2	√8.5	13.1	·34.1	√8.5	√9.3	√9.3
segment	**4.1**	6.6	√5.0	6.0	6.2	·7.7	6.5	√5.4
soy-large*	**5.8**	√6.7	√6.6	√7.4	√7.0	√7.7	√9.0	·27.0
tic-tac-toe	**1.7**	10.5	13.5	5.8	14.8	√2.0	·29.8	√2.9
vehicle	**27.3**	√32.2	√29.4	√28.9	√29.0	√33.4	·39.6	√31.3
avg rank	√3.60	**3.42**	4.23	4.74	4.97	·5.66	4.95	4.43

Table 8.5: LIBSVM outperforms all other methods in classification accuracy.

log-loss / instance				LIBSVM		$SVM^{multiclass}$					
domain	t_K	t_{SVM}	t^p_{SVM}	SVM^p_{dot}	SVM^p_{rbf}	SVM_{dot}	SVM_{rbf}	C4.5	LR	NB	kBMA
cmc	0.04	3.21	1.92	0.96	√0.94	3.20	*·3.44*	1.07	0.97	1.00	**0.92**
lenses	0.00	0.05	0.07	√0.72	0.88	√0.66	√0.73	√0.51	√0.89	√2.44	**0.39**
monk2	0.01	0.35	0.29	0.64	0.54	*·2.12*	1.44	0.73	0.65	0.65	**0.45**
shuttle	0.01	0.11	0.11	√0.10	0.16	*·0.36*	√0.23	√0.07	√0.10	0.16	**0.07**
titanic	0.01	0.80	0.67	0.53	0.50	*·1.67*	1.63	√0.49	0.50	0.52	**0.48**
soy-small*	5.29	0.11	0.12	*·0.31*	0.27	0.08	0.08	0.11	0.15	**0.00**	0.00
wine	0.10	0.15	0.16	√0.09	√0.08	√0.09	√0.17	√0.27	√0.09	**0.06**	√0.14
yeast-class*	138	0.72	0.19	0.06	0.06	0.01	0.01	0.13	*·0.90*	**0.01**	0.23
anneal	6.16	1.96	0.52	√0.05	0.11	*·0.13*	√0.07	0.08	**0.02**	√0.07	0.11
bupa	0.01	0.20	0.13	√0.60	√0.61	*·2.09*	1.89	0.68	**0.60**	√0.62	√0.61
hayes-roth	0.00	0.15	0.09	√0.35	√0.28	1.10	*·1.48*	0.67	**0.26**	0.46	0.45
mushroom	1.33	7.57	10.4	0.00	0.01	0.01	0.00	0.00	**0.00**	*·0.01*	0.00
pima	0.02	0.40	0.36	√0.48	√0.48	1.49	*·1.56*	0.53	**0.46**	√0.50	√0.48
tic-tac-toe	0.03	0.73	0.48	√0.08	0.23	*·0.90*	0.39	0.46	**0.06**	0.55	√0.07
audiology*	81.2	17.0	0.51	1.31	1.49	1.44	√1.36	**1.04**	1.40	*·3.55*	2.23
krkp	6.52	4.02	5.42	0.10	0.17	*·0.48*	0.27	**0.03**	0.08	0.29	√0.05
monk3	0.01	0.19	0.15	0.11	0.14	*·0.22*	√0.16	**0.07**	0.10	0.20	0.11
soy-large*	5.95	11.0	0.98	0.46	0.47	√0.44	0.50	**0.27**	√0.37	0.57	*·0.68*
zoo*	0.23	0.21	0.19	0.47	*·0.51*	√0.47	**0.21**	√0.34	√0.38	√0.38	√0.40
o-ring	0.00	0.05	0.11	√0.80	√0.54	**0.50**	√0.60	√0.56	0.66	√0.83	√1.00
australian	0.16	0.60	0.50	√0.36	**0.36**	0.92	*·1.04*	√0.37	√0.39	√0.46	√0.38
breast-LJ	0.03	0.20	0.18	0.59	**0.55**	*·1.53*	1.47	0.62	√0.58	√0.62	√0.58
breast-wisc	0.03	0.23	0.17	0.12	**0.09**	√0.15	0.18	0.20	√0.13	√0.21	√0.18
car	0.02	1.68	0.66	0.18	**0.14**	*·1.18*	0.68	0.30	0.33	0.32	0.19
ecoli	0.01	0.34	0.20	√0.55	**0.50**	*·1.02*	0.89	√0.66	0.68	√0.89	√0.83
german	0.64	1.43	2.34	√0.51	**0.50**	1.64	*·1.83*	0.82	√0.52	√0.54	0.59
glass	0.03	0.28	0.13	√0.78	**0.75**	*·1.62*	1.38	1.10	1.07	1.25	√1.05
heart	0.15	3.46	1.60	1.03	**1.00**	2.79	*·2.94*	1.70	1.24	1.25	√1.10
hepatitis	0.47	0.15	0.14	√0.39	**0.33**	√0.67	*·0.87*	0.66	√0.77	√0.78	√0.43
horse-colic	1.89	1.68	0.60	√0.71	**0.69**	1.62	1.74	1.33	*·1.81*	1.67	√0.83
ionosphere	3.71	0.57	0.20	√0.26	**0.20**	0.42	0.45	0.36	*·0.69*	0.64	√0.33
iris	0.00	0.09	0.09	0.24	**0.17**	√0.26	√0.32	√0.20	√0.21	√0.27	√0.23
lymph	0.39	0.21	0.16	√0.56	**0.48**	√0.82	√0.90	0.86	0.91	*·1.10*	√0.86
monk1	0.01	0.29	0.19	0.49	**0.01**	*·1.55*	√0.13	√0.05	0.50	0.50	√0.02
p-tumor*	0.39	6.07	0.39	√1.93	**1.92**	3.10	*·3.42*	2.76	2.76	3.17	2.61
spam	39.9	13.2	35.9	√0.16	**0.16**	0.48	*·2.45*	0.29	√0.16	0.53	√0.19
voting	0.23	0.24	0.16	√0.13	**0.12**	0.27	0.45	√0.18	0.37	*·0.60*	√0.15
wdbc	0.57	0.33	0.18	√0.10	**0.09**	√0.14	√0.19	0.19	*·0.42*	0.26	√0.13
adult	1.11	436	2678	**0.29**	0.31	1.32	*·1.98*	0.35	0.35	0.42	0.30
balance-scale	0.00	0.50	0.19	**0.17**	√0.22	0.53	0.81	*·1.11*	0.28	0.51	0.51
crx	0.19	0.69	0.54	**0.34**	√0.35	0.92	*·1.05*	√0.37	√0.39	√0.49	√0.36
lung-cancer*	35.0	0.22	0.13	**1.02**	√1.17	1.90	1.86	√1.54	√1.24	*·5.41*	√1.62
post-op	0.01	0.08	0.09	**0.61**	√0.62	√1.20	*·1.42*	√0.64	√0.81	√0.93	√0.67
promoters*	37.5	0.47	0.16	**0.23**	√0.24	√0.44	*·0.82*	0.74	0.70	√0.60	√0.54
segment	0.74	26.1	4.59	**0.14**	0.18	0.38	0.45	0.27	*·0.45*	0.38	√0.17
vehicle	0.42	2.61	1.10	**0.56**	0.60	*·1.91*	1.88	0.90	0.93	1.78	0.66
avg rank				√2.83	**2.65**	6.13	*·6.36*	4.21	4.35	5.76	3.72

Table 8.6: LIBSVM yields excellent performance with respect to log-loss as well. It is somewhat surprising that C4.5 achieves lower log-loss than logistic regression. To obtain the probabilities from classifiers, we have admixed the uniform distribution with a small weight $(1/(|\Re_Y| + |\mathcal{D}|))$: without this, the log-loss could be infinite. This admixing was performed for $SVM^{multiclass}$, LR and C4.5.

CHAPTER 9

Discussion

We hope that the work has demonstrated the utility of interactions as a descriptive metaphor of the data. We have provided a definition of an interaction, and many examples of interactions identified in real-life data. We have provided ways of inducing the patterns of interactions from the data, for the goals of exploratory data analysis (visualization), confirmatory data analysis (significance testing), and for predictive modelling (classification). Our methods are practical and effective, and we have already applied them to numerous cases ranging from medicine, data mining, economics, to political science.

A more detailed list of contributions of this dissertation appeared already in Ch. 1, so we will conclude with a summary of our subjective impressions. We will list implications for the field of machine learning. Finally, we will mention some possibilities for further work.

Although there should be little doubt that interactions are a useful metaphor, it is clear that it is not the only metaphor that works. Other ideas and concepts in machine learning are also important. A good way of appreciating the differences is to examine the experimental results of Ch. 8. We can see that interactions offer a very good explanation for some situations, such as the 'CMC' and the 'Titanic' data sets. For some situations, such as 'KRKP' and 'Monk3' the classification trees are distinctly more appropriate: the interaction graphs are sprawling and ineffective as compared to a well-structured depiction of a single set of rules. Furthermore, the extreme success of SVM with dot product kernels and of logistic regression reminds us that many practical problems do not involve sophisticated interactions: instead the problems of weighting and fusion are the ones that require a proper solution.

It is easy to think that an ensemble-based approach will solve the above conundrum. We believe otherwise: these methods are complementary and need to be integrated, not just combined. The following summary in the notation of Salthe captures the idea best:

[[[rules, support vectors and constructs] interactions] voting, weighting and fusion]

This means that rules are to be found within an interaction, and that fusion methods should operate on top of interactions.

Although Kikuchi-Bayes does not win against the cutting-edge SVM classifiers, it is fair to mention that we are using relatively crude Cartesian products within an interaction,

and relatively crude Kikuchi approximations for the fusion. Clearly, these methods should be improved. We have already suggested MaxEnt, hypertree factorization and convex optimization as means of addressing the problem of fusion in our chapter on classification. We also suggest methods such as (Buntine and Jakulin, 2004, Clinton et al., 2004) and (Della Pietra et al., 1997) for handling the inference of structure within an interaction in our chapter on visualization.

The problem of a large number of attributes is important for text mining, and our approach is relatively basic. There are various ideas that can be tackled, including those we mention in the chapter on attribute selection. Although we have provided probabilistic models for continuous attributes, the results in that area are either simplistic (multivariate Gaussian model), or potentially non-robust (EM). More work is clearly needed in that area.

There are some ideas in addition to the notion of interactions that proved their worth. First, the Bayesian philosophy is a solid and cohesive theory of learning. Bayesian model averaging showed its unambiguous effectiveness in prediction, and Bayesian tests of significance are the ones that proved to be least sensitive to problems. One should not be afraid of assuming priors. Even if we assume cross-validation as an evaluation method, we can select a prior appropriate for the task at hand. Although others have noted this earlier, it was surprising how well the notion of a parsimonious prior describes the performance of a learning algorithm on independent data.

Another idea that 'works' is the notion of trajectory integration. Our application of this idea has been inspired by the work on boosting and regularization paths. However, we have transplanted these heuristics into a Bayesian context; their role is to balance the breadth and depth in the consideration of models. Our computation of degrees of freedom according to Krippendorff (1986) should be seen as a heuristic, and a more thorough formal study is required. In addition to that, it is not clear what is the right way of assessing degrees of freedom for other models. Still, our experimental framework should be well-suited to guide and check further work.

We can also informally criticize certain ideas. First, the notion of conditional probability adds unnecessary complexity to the problem of building predictive models. We find it easier to understand conditional probability as a computationally convenient special case of Kikuchi approximation. The models specified in terms of conditional probabilities are often unnecessarily complex, and offer confusing and unfounded indications of causality. Still, we appreciate the work of Pearl (2000) on formalizing causality, which shows nicely that conditional independence is a way of accommodating causality, but not a way of inferring causality.

The second idea we criticize is the blind trust in cross-validation and the blind pursuit for classification accuracy. We hope that the analysis in Chapters 2 and 8 shows how important is the training/test proportion used in cross-validation. Furthermore, several authors have pointed out the leave-one-out is not always a sensible evaluation method in the context of uncertainty, and our results seem to confirm this. Finally, classification accuracy is a rather unreliable measure of classifier performance in the context of uncertainty, and either proper p-loss functions, ROC measures, or margin width (SVM) are to be used instead. In the tables of Ch. 8 it can be seen how much more noise there is in classification accuracy results.

Still, one can criticize that the work in Ch. 8 is done solely in the context of logarithmic loss. Although log-loss is a well-behaved generic loss function that assures the resulting

probabilities to be sensible in the frequentist sense, and although most utility functions are expressed for a decision specified in terms of the expected loss, this criticism is well-deserved. We point out the work of Rubin (1987), Rissanen (2001), Grünwald (1998): the prior and the likelihood function can be adjusted for a specific loss function at hand, and the Bayesian modelling remains the same from that point onwards. Furthermore, Grünwald and Dawid (2004) show how MaxEnt inference can be adjusted for other loss functions. The likelihood function can also be adjusted appropriately when the IID assumption is unfounded; IID merely means that the instances are independent given the model.

We have also observed that relatively complex models, whose degrees of freedom exceeded the number of instances, nevertheless managed to achieve very good performance as assessed through cross-validation. This phenomenon should be explained properly: it might either be due to hidden dependencies, a flaw in our assumptions, a flaw of our degrees of freedom assessment, or a flaw of the cross-validation procedure of the type we have already encountered in Sect. 4.5.5 and in Fig. 2.6.

Interaction analysis also seems to be very well-aligned with human intuitions about the data. Indeed, positive and negative interactions seem intuitive, and they help understand many 'anomalies' such as the mismatch between correlation coefficients and the regression coefficients, the XOR problem, the redundancies. Even the convenience of conditional independence is merely a special case of using interaction information and region-based approximations.

A particular conundrum not handled very well in the present work is the peculiar and deep influence of the loss functions. In some chapters we use joint loss functions where the outcome involves all attributes. On other chapters we use conditional loss functions that only involve the labelled attributes. But one should be highly conscious of what loss function should be used. For example, the Kikuchi approximation is suited to fusing interactions under the maximum *joint* entropy criterion. However, a different notion of maximum conditional entropy criterion is more suitable for learning class-predictive models (as assessed through the loss functions on the outcomes). The same applies for the chain rule. In all, we should also speak of conditional and joint fusion methods.

DODATEK A

Povzetek v slovenskem jeziku

Strojno učenje na osnovi interakcij med atributi

Aleks Jakulin

Povzetek

V pričujočem delu definiramo koncept interakcije, ki označuje medsebojno povezanost atributov. Stopnjo interakcije definiramo kot povečanje koristnosti modela s tem, da mu dovolimo hkrati upoštevati vrednost več atributov. Izkaže se, da so korelacijski koeficienti, informacijski prispevek ter interakcijski prispevek vsi posebni primeri tega splošnejšega koncepta.

Če se učimo modela na podlagi omejene količine podatkov in danega prostora hipotez, moramo vedno upoštevati negotovost glede pravilnosti posamičnega modela. Pristop k učenju, ki upošteva to načelo, je Bayesova statistika. Poleg tega na proces učenja vpliva tudi funkcija koristnosti kot končni vzrok in algoritem kot gonilo procesa sprehajanja po prostoru hipotez. V splošnem lahko proces opišemo na podlagi Aristotelove vzročnosti, ki razloži neskladja med pristopi k strojnemu učenju: nekateri pristopi se razlikujejo pri izbiri vzrokov, ali pa nekaterim vzrokom posvečajo več pozornosti kot drugim.

Interakcija v tem okviru je element prostora hipotez, teorija o informacijah pa določi tudi uporabo verjetnosti kot osnovnega gradnika hipotez, logaritmično napako kot funkcijo koristnosti, ob tam pa nudi učinkovit izrazni jezik za ocenjevanje različnih odnosov med atributi in modeli. Ker pa je bistvo teorije o informacijah analiza verjetnostnih modelov, uporabimo Bayesovo statistiko, da se teh naučimo iz podatkov. S tem tudi odpremo možnost negotove obravnave količin kot so medsebojna informacija, pa tudi preskusov značilnosti, ki temeljijo na Kullback-Leiblerjevi divergenci. Vseeno pa moramo ločiti med razgradnjo entropije (kjer se lahko pojavljajo tako soodvisnosti oziroma negativne interakcije kot tudi sodejavnosti oziroma pozitivne interakcije) ter primerjave med modeli (kjer se pojavljajo samo pozitivne razdalje med koristnostmi).

Da bi človeku lahko učinkovito prikazali vzorce interakcij v podatkih, predstavimo več tipov grafičnih predstavitev. Informacijski graf služi prikazu informacijsko-teoretičnih količin, kjer ploščino merimo v bitih. Interakcijska matrika prikaže stopnjo in tip interakcije med vsakim parom atributov. Interakcijski graf izloči najmočnejše posamične interakcije, ki jih označimo s pomembnostjo, tipom, stopnjo značilnosti. Obstaja tudi več prikazov strukture znotraj posamične interakcije, kot recimo pravil, taksonomij in podobno.

Interakcije lahko v praktičnem strojnem učenju uporabimo za izbiro atributov: negativne interakcije znižajo kvaliteto posamičnega atributa. Druga možnost pa je, da dejanski model, ki je rezultat učenja, predstavimo kar z množico interakcij. S pomočjo metode maksimalne entropije ali s hitrim in učinkovitim Kikučijevim približkom lahko ustvarimo napovedni model z združevanjem tudi prekrivajočih se interakcj. Napovedi s tem Kikuči-Bayesovim modelom posplošujejo Bayesove mreže in se dobro izkažejo v primerjavi z drugimi postopki strojnega učenja.

Ključne besede

- strojno učenje
- klasifikacija, razpoznavanje vzorcev, uvrščanje
- interakcija, soodvisnost, sodejavnost, odvisnost, neodvisnost
- teorija o informacijah, entropija, medsebojna informacija
- Bayesova statistika, preizkus značilnosti

A.1 Uvod

A.1.1 Učenje

Ko se učimo modelov, to vedno počnemo v nekem izraznem jeziku. Imamo jezike pravil, jezike enačb, jezike iz besed. Ko opisujemo pojave, ljudje velikokorat rečemo, da je nekaj povezano z nečim drugim. Ampak kako bi formalno definirali pomen te 'povezanosti'? To je osnovni cilj te disertacije. Povrh tega pa je naš cilj tudi pokazati, kako je lahko ta pojem koristen v praktičnih uporabah, tako za subjektivne koristi razumljivega prikazovanja podatkov človeku, kot tudi za objektivno merjene koristi večje točnosti pri napovedovanju. Namreč, če je naš pojem povezanosti, recimo mu *interakcija*, res koristen, se mora to poznati tudi pri klasifikacijski točnosti.

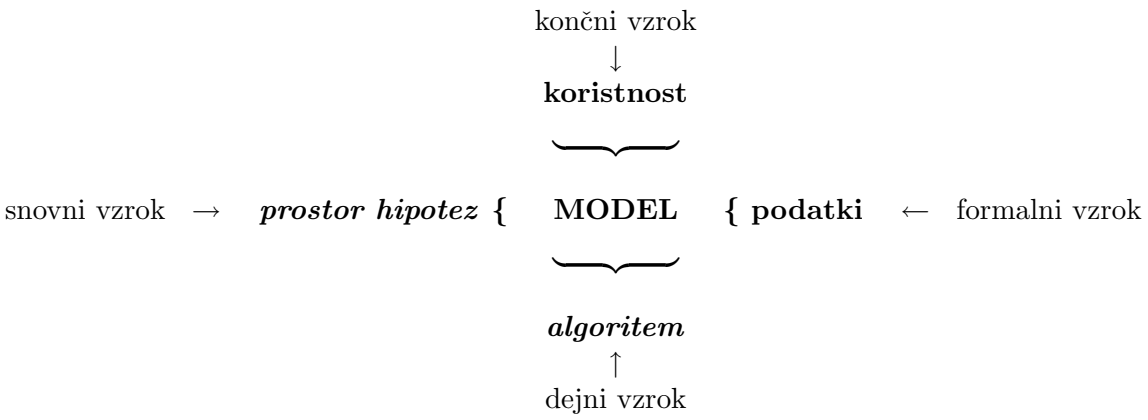
Preden se lotimo modeliranja, si moramo definirati nekaj pojmov, ki so zanj potrebni. Pojmi so naslednji:

- **Prostor hipotez:** Ta prostor določi, kaj je smiselna izjava. Prostor linearnih modelov za dva atributa je $\mathbb{R}^2$, kjer prva koordinata določi naklonjenost premice, druga pa odmik premice od izhodišča. Nekateri prostori imajo potencialno neskončno dimenzionalnost, recimo prostor odločitvenih dreves, prostor izračunljivih funkcij, ali prostori, s katerimi delajo postopki podpornih vektorjev (SVM).

- **Predstavitev podatkov:** Ponavadi so podatki predstavljeni kot seznam opažanj, primerov. Vsako opažanje je opisano z množico vrednosti atributov. Pri tem velikokrat predpostavimo, da imajo atributi nek stalen pomen, ki se ne spreminja od primera do primera.

- **Algoritem učenja:** Da bi seznamu opažanj pridali pomen v prostoru hipotez, potrebujemo nek postopek, ki se sprehaja po prostoru hipotez in sestavlja hipotezo, ki je skladna s podatki. To je naloga algoritma učenja. V preteklosti so se uporabljali že v naprej pripravljeni modeli, ki niso zahtevali preiskovanja, algoritmi učenja so bili le enostavni izračuni. Danes pa se področje strojnega učenja večinoma ukvarja z neskončnimi prostori hipotez. Algoritem je nosilec izkušenj iz preteklih problemov učenja, tako da zna učinkovito preiskovati prostor hipotez.

- **Funkcija koristnosti:** Enostavna funkcija koristnosti je dvojiška: je hipoteza resnična ali ne? Žal tega ne vemo. Podatki si lahko nasprotujejo, lahko jih je premalo. Četudi se hipoteza popolnoma ujema s podatki, se mogoče ne bo več ujemala na novih podatkih. V vsej tej zmešnjavi je naloga funkcije koristnosti, da tehta, kako koristna je neka hipoteza: po kriterijih skladnosti s podatki, njene enostavnosti in mogoče skladnosti s preteklimi podobnimi problemi.



Slika A.1: Štirje Aristotelovi vzroki učenja. Simbol $a\{b$ pomeni, da je a splošnejši od b. Vzroki omejujejo model. Podatki so formalni vzrok, saj v procesu učenja ne vemo, kaj je resnični model. Negotovost se mora zato pojaviti v prostoru hipotez.

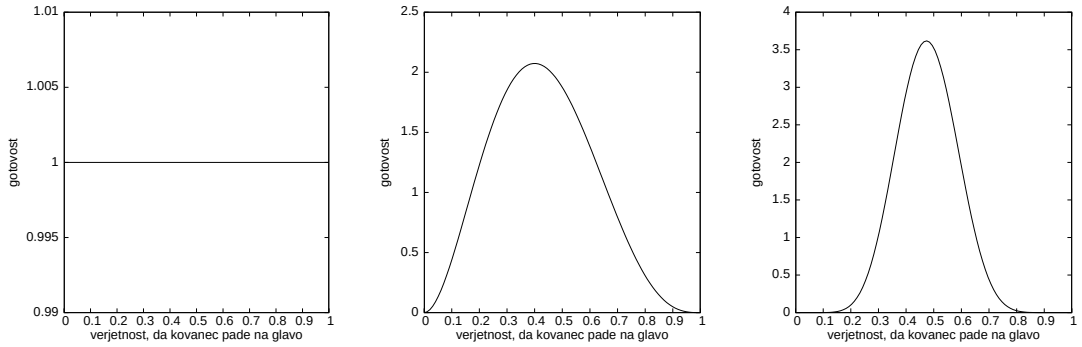
Je hipoteza tudi model? Načeloma je lahko več hipotez enako koristnih. Ker je koristnost osnovni kriterij (to predpostavimo, to je definicija koristnosti), so zato vse te hipoteze enakovredne in ni razloga, da bi lahko izbirali. Epikurovo načelo neopredeljenosti (Kirchherr et al., 1997) pravi: *Obdrži vse hipoteze, ki so skladne z dejstvi.* Torej izbirati načeloma niti ne smemo. Model je lahko skupek (ensemble) več hipotez.

Slika A.1 prikazuje Aristotelov model učenja (Jakulin, 2004) opisan kot prepletanje dveh specifikacijskih hierarhij (Salthe, 1993). Aristotelova vzročnost, ne da bi se tega zavedali, v veliki meri opisuje strukturo našega razmišljanja. Nekateri se opredelijo na določene vzroke in poskušajo odstraniti vse ostale. Nekatere veje znanosti, recimo, odstranijo vse vzroke razen dejnega in naravo predstavijo kot mehanski algoritem ali velikanski računalnik. Vendar pa je ta model vzročnosti vedno z nami, ne zato, ker bi bil resničen ali nujen, ampak ker na ta način ljudje razmišljamo.

A.1.2 Negotovost

Bi lahko napovedali izid meta kovanca? Mogoče v principu že, v praksi pa brez popolnega nadzora nad eksperimentom ne. Vseeno pa lahko nekaj le rečemo o kovancu. Lahko pa povemo, da ponavadi pade približno enako cifer kolikor glav, če kovanec ni zvit. Prostor hipotez v našem primeru predstavlja binomska porazdelitev z enim samim parametrom, verjetnostjo p. Recimo, da imamo nekaj metov kovanca. Kako sklepamo o verjetnosti p? Odgovor na to vprašanje ponuja Bayesova statistika (Bernardo and Smith, 2000, Gelman et al., 2004a). V osnovi lahko neko gotovost (belief) za vsako verjetnost (probability). Za primer kovanca prikažemo porazdelitve gotovosti na sliki A.2.

Podatki so predstavljeni kot množica $\mathcal{D}$. Vektor $\mathbf{X}$ predstavlja neoznačene atribute, ki so ponavadi dani. Vektor $\mathbf{Y}$ včasih označuje atribute, ki so označeni, in se jih trudimo napovedati. Posamični primer je potem $\mathbf{v}^{(i)} = \langle \mathbf{x}^{(i)}, \mathbf{y}^{(i)} \rangle$, kjer so $\mathbf{X} = \mathbf{x}$ in $\mathbf{Y} = \mathbf{y}$ vrednosti atributov za dani primer i. Zalogo vrednosti posamičnega atributa X (ki je del vektorja $\mathbf{X}$) označimo z $\Re_X = \{x_1, x_2, \ldots, x_k\}$. Formalno lahko rečemo, da je parameter p vsebovan v opisu hipoteze, ki ga izrazimo kot vektor parametrov $\mathbf{\Theta}$.



Slika A.2: Bayesov skupek hipotez. Vsaka verjetnost, da bo kovanec padel kot glava je posamična vrednost parametra p. Vse možne vrednosti tega parametra tvorijo skupek, vendar niso vse enako gotove. Zato vsaki verjetnosti pripišemo aposteriorno gotovost pri danih podatkih. Na začetku so gotovosti vseh verjetnosti p enake, vendar se z večjim in večjim številom podatkov tudi naš skupek hipotez bolj in bolj natančno opredeljuje glede dejanske verjetnosti (od leve proti desni). Vseeno pa smo vedno vsaj malo negotovi glede verjetnosti.

Izraz P uporabljamo tako za verjetnosti kot tudi za gotovosti: matematično se ne razlikujeta, le pomensko. Bistvo Bayesove statistike temelji na eksplicitni predpostavki apriorne gotovosti glede parametrov (prior) $P(\boldsymbol{\Theta}|\mathcal{H})$. Včasih poznamo razvidnost (evidence) ali verjetnost, da bomo opazili dane podatke $P(\mathcal{D}|\mathcal{H})$. Funkcija zanesljivosti (likelihood) nam ovrednoti gotovost v vrednost parametrov $\boldsymbol{\Theta} = \boldsymbol{\theta}$ pri danih podatkih $P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H})$. Končni cilj je ponavadi aposteriorna gotovost (posterior) ali zaupanje v parametre pri danih podatkih $P(\boldsymbol{\Theta}|\mathcal{D}, \mathcal{H})$. Vse to opišemo z enačbo (MacKay, 2003):

$$P(\boldsymbol{\Theta}|\mathcal{D}, \mathcal{H}) = \frac{P(\boldsymbol{\Theta}|\mathcal{H})P(\mathcal{D}|\boldsymbol{\Theta}, \mathcal{H})}{P(\mathcal{D}|\mathcal{H})}. \tag{A.1}$$

Vidimo, da je vse v kontekstu $\mathcal{H}$: to je prostor hipotez, ki pa ga ponavadi ne omenjamo eksplicitno. Moramo pa se zavedati, da je to osnovna predpostavka. Funkcija zanesljivosti ponavadi tudi predpostavi neodvisnost učnih primerov:

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) = \prod_{\mathbf{x} \in \mathcal{D}} P(\mathbf{x}|\boldsymbol{\theta}). \tag{A.2}$$

Ker ponavadi razvidnosti ne poznamo, jo tudi lahko odstranimo, tako da predpostavimo, da je aposteriorno gotovost normalizirana pri danih podatkih:

$$P(\boldsymbol{\Theta}|\mathcal{D}, \mathcal{H}) = \frac{P(\boldsymbol{\Theta}|\mathcal{H})P(\mathcal{D}|\boldsymbol{\Theta}, \mathcal{H})}{\int_{\Re_{\boldsymbol{\Theta}}} P(\boldsymbol{\theta}|\mathcal{H})P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H})d\boldsymbol{\theta}} \tag{A.3}$$

Pri praktičnem napovedovanju se lahko zavedamo negotovosti glede izbire modela, ali pa tudi ne. Če se te negotovosti nočemo zavedati, jo marginaliziramo (integrate out), tako da povprečimo napovedi vseh smiselnih modelov:

$$P(\mathbf{X}|\mathcal{D}) = \int_{\Re_{\boldsymbol{\Theta}}} P(\mathbf{X}|\boldsymbol{\Theta})P(\boldsymbol{\Theta}|\mathcal{D})d\boldsymbol{\Theta} \tag{A.4}$$

Temu pravimo Bayesovsko povprečenje modelov, ali BMA (Bayesian model averaging). Marginalizacija pri zgornjem primeru kovanca, pri čemer smo uporabili enakomerno apriorno porazdelitev, nam da Laplacovo oceno verjetnosti:

$$p^H_{BMA} = \frac{n_H + 1}{n_H + n_T + 2}$$

Seveda pa je Laplacova ocena enaka, če smo videli 10000 metov kovanca ali dva meta kovanca z enakim razmerjem med cifro in glavo: negotovost glede verjetnosti je izgubljena. Zato ne modela ne smemo vedno obravnavati le kot povprečje.

Včasih nas zanima napovedovati le vrednost označenih atributov $\mathbf{Y}$. Za tako vrsto modeliranja, primer katere je regresija, uporabimo pogojno funkcijo zanesljivosti. Če napovedujemo $\mathbf{y}$ iz $\mathbf{x}$, je pogojna funkcija zanesljivosti ob predpostavki neodvisnosti med primeri definirana kot

$$P(\mathcal{D}|\boldsymbol{\theta}, \mathcal{H}) = \prod_{\langle \mathbf{x},\mathbf{y}\rangle \in \mathcal{D}} P(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}) \tag{A.5}$$

Aposteriorno gotovost glede modela $\hat{P}(\boldsymbol{\Theta}|\mathcal{D})$ lahko interpretiramo kot nekakšno analogijo funkcije koristnosti. S takim pogledom izberemo aposteriorno najbolj gotov model (maximum a posteriori, MAP) $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \hat{P}(\boldsymbol{\theta}|\mathcal{D}) = \arg\max_{\boldsymbol{\theta}} \hat{P}(\boldsymbol{\theta})\hat{P}(\mathcal{D}|\boldsymbol{\theta}) \tag{A.6}$$

Precej znanih postopkov v strojnem učenju, kot recimo princip najkrajšega opisa (MDL), je le posebnih primerov MAP.

A.2 Teorija informacije

A.2.1 Osnovne količine

Pri teoriji informacije vnaprej predpostavimo nek določen verjetnostni model P. O tem modelu lahko marsikaj povemo z uporabo količin iz teorije informacij. Osnovna količina je Shannonova entropija (Shannon, 1948):

$$H(A) \triangleq -\sum_{a \in \Re_A} P(a) \log_2 P(a) \tag{A.7}$$

Po definiciji, $0 \log_2 0 = 0$. Tu smo jo izračunali za atribut A. Načeloma pa jo lahko izračunamo za poljubno podmnožico atributov $\mathbf{Y}$ pri pogoju $\mathbf{X} = \mathbf{x}$:

$$H(\mathbf{Y}|\mathbf{x}) \triangleq -\sum_{\mathbf{y} \in \Re_{\mathbf{Y}}} P(\mathbf{x}, \mathbf{y}) \log_2 P(\mathbf{y}|\mathbf{x}) \tag{A.8}$$

Važno pa je, da so atributi diskretni: večina lastnosti količin iz teorije informacij izhaja iz te predpostavke.

Druga pomembna količina je Kullback-Leiblerjeva divergenca (Kullback and Leibler, 1951), ki meri razdaljo med dvema verjetnostnima porazdelitvama, ki sta lahko tudi pogojni:

$$D(P(Y|X) \| Q(Y|X)) \triangleq \sum_{x \in \Re_X, y \in \Re_Y} P(y, x) \log_2 \frac{P(y|x)}{Q(y|x)} \tag{A.9}$$

Tu je porazdelitev P referenčna, Q pa alternativna: KL-divergenca namreč ni simetrična.

Za entropijo in KL-divergenco se skriva funkcija logaritmične napake $L(\mathbf{x}, Q) = -\log_2 Q(\mathbf{x})$. Logaritmična napaka je korektna (proper), saj bo v kontekstu nedeterminističnih dogodkov minimum dosegla ravno pri pravilni verjetnosti; klasifikacijska točnost ali informacijska vsebina odgovora nimata te lastnosti. Torej, če model Q opisuje naše napovedi, je $L(\mathbf{x}, Q)$ napaka našega modela ob dogodku $\mathbf{x}$. Entropija je minimalna možna stopnja napake, če je resnica P:

$$H(\mathbf{X}|P) = \inf_Q \mathbb{E}_{\mathbf{x}\sim P}\{L(\mathbf{x}, Q)\}$$

Eksplicitno smo napisali, da entropija temelji na modelu P: to je vedno res, vendar tega ponavadi ne navajamo. Kullback-Leiblerjeva divergenca pa je obžalovanje (regret) modela Q pri resnici P. Obžalovanje označuje presežek napake preko stopnje, v katero smo itak prisiljeni:

$$D(P(\mathbf{X})\|Q(\mathbf{X})) = \mathbb{E}_{\mathbf{x}\sim P}\{L(\mathbf{x}, Q) - L(\mathbf{x}, P)\}$$

Vidimo, da pri nekaterih napovednih problemih napovedujemo $\mathbf{Y}$ pri danem $\mathbf{X} = \mathbf{x}$, za kar moramo uporabiti pogojne funkcije napake, pogojno entropijo ter pogojno KL-divergenco. Povrh tega je včasih bolj primerna kaka druga funkcija napake in ne ravno logaritmična. Takrat KL-diverenco in entropijo ustrezno popravimo, a previdno, saj se nekatere lastnosti ne ohranijo (Grünwald and Dawid, 2004).

Zelo pomembna količina je medsebojna informacija ali informacijski prispevek, ki jo sicer uporabljamo za oceno pomembnosti atributov pri napovedovanju razreda:

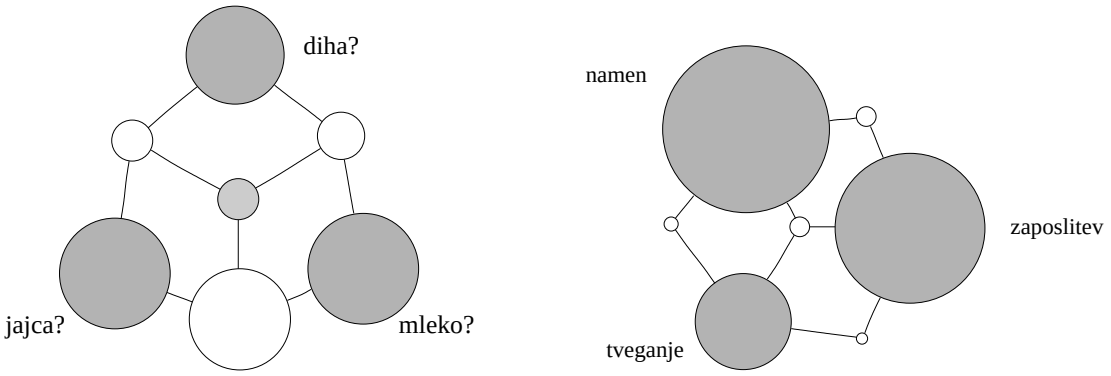
$$\begin{aligned} I(A;B) &\triangleq \sum_{a\in\Re_A, b\in\Re_B} P(a,b)\log_2\frac{P(a,b)}{P(a)P(b)} \\ &= D(P(A,B)\|P(A)P(B)) = D(P(A|B)\|P(A)) = D(P(B|A)\|P(B)) \\ &= H(A) + H(B) - H(A,B) = H(A) - H(A|B) = I(B;A) = H(B) - H(B|A) \end{aligned} \tag{A.10}$$

Vidimo, da medsebojno informacijo lahko interpretiramo na dva načina: kot KL-divergenco med dvema verjetnostnima modeloma, ali pa kot vsoto in razliko entropij podskupin atributov. Poznamo tudi pogojno medsebojno informacijo s podobnimi lastnostmi:

$$\begin{aligned} I(A;B|C) &\triangleq \sum_{a,b,c} P(a,b,c)\log_2\frac{P(a,b|c)}{P(a|c)P(b|c)} = H(A|C) + H(B|C) - H(AB|C) \\ &= H(A|C) - H(A|B,C) = H(AC) + H(BC) - H(C) - H(ABC). \end{aligned} \tag{A.11}$$

Tule se pojavlja več simbolov. Včasih pišemo vejico $(,)$, ki veže najmočneje in pomeni 'in', povezuje dva atributa med seboj, podobno kot $P(A,B)$ ali kot unija napovedi. Podpičje $(;)$ veže malce šibkeje in ločuje dve skupini atributov med seboj, podobno kot $P(A)P(B)$, ali kot presek napovedi. Pogoj $(|)$ veže še šibkeje in loči pogoj na desni strani od izraza med atributi na levi strani, podobno kot $P(A|B)$, ali kot odštevanje napovedi. Končno, dvočrta $(\|)$ loči dva modela.

Ko dovolimo več skupin atributov, pridemo do interakcijskega prispevka (McGill, 1954,



Slika A.3: Levo: Primer soodvisnosti ali negativne 3-interakcije med atributi 'leže jajca?', 'diha?' in 'daje mleko?' za različne živali. Če vemo, da leže jajca, skoraj zagotovo vemo, da nam žival ne daje mleka; zato nam informacija, da diha (torej, da ni podvodna žival) ne pove nič novega, četudi je malo podvodnih živali z mlekom. **Desno:** Primer sodejavnosti ali pozitivne 3-interakcije med atributi 'namen' 'tveganje' in 'zaposlitev' nekega kreditojemalca: če ima brezposeln človek velike namene pri najemu kredita, se nam zdi to veliko bolj tvegano kot veliki nameni sami po sebi ali brezposelnost sama po sebi.

McGill and Quastler, 1955, Demšar, 2002):

$$\begin{aligned} I(A;B;C) &\triangleq I(A;B|C) - I(A;B) = I(A,B;C) - I(A;C) - I(B;C) \\ &= H(AB) + H(BC) + H(AC) - H(A) - H(B) - H(C) - H(ABC) \\ &= I(B;C;A) = I(C;B;A) = I(A;C;B) = I(B;A;C) \end{aligned} \tag{A.12}$$

Ta služi kot nekakšna mera interakcije med tremi atributi. Najbolje ga razumemo kot razbitje entropije med tremi atributi $H(A,B,C)$ na vsoto posamičnih entropij $H(A), H(B), H(C)$, 2-informacij med njimi $I(A;B), I(B;C), I(A;C)$, ter 3-interakcije med njimi $I(A;B;C)$, tako da velja:

$$H(A,B,C) = H(A) + H(B) + H(C) + I(A;B) + I(B;C) + I(A;C) + I(A;B;C)$$

Interakcijski prispevek med dvema atributoma je kar medsebojna informacija.

Če iz podatkov ocenimo verjetnostni model, lahko s pomočjo informacijskih grafov prikažemo odnose med atributi. Vsakemu atributu pripišemo temen krog, katerega ploščina je proporcionalna entropiji tega atributa, torej naši negotovosti glede vrednosti tega atributa. Pozitivne interakcije ali sodejavnosti prikazujejo prekrivanje med informacijo dveh atributov, kar pomeni, da nam vrednost enega atributa nekaj pove o vrednosti drugega atributa. To v informacijskem grafu prikažemo kot bel krog, ki povezuje dva atributa in katerega ploščina ustreza interakcijskemu prispevku. Negativne interakcije ali soodvisnosti imajo pomen odvečnosti, kjer nam en atribut pove isto kot drugi atribut o tretjem atributu. To v informacijskem grafu prikažemo kot temen krog, ki povezuje prekrivajoče se pozitivne interakcije med seboj. Primer obeh tipov interakcij je na sliki A.3.

A.2.2 Posplošeni interakcijski prispevek

Tudi interakcijski prispevek lahko interpretiramo kot primerjavo med dvema modeloma: med dejansko verjetnostno porazdelitvijo $P(A, B, C)$ in približkom Kirkwooda $\hat{P}_K$ (Kirkwood and Boggs, 1942, Matsuda, 2000):

$$\hat{P}_K(a, b, c) \triangleq \frac{P(a, b)P(a, c)P(b, c)}{P(a)P(b)P(c)} = P(a|b)P(b|c)P(c|a) \tag{A.13}$$

Ob tem velja $I(A; B; C) = D(P(A, B, C)\|\hat{P}_K(A, B, C))$. Mogoče se zdi čudno, da je interakcijski prispevek lahko negativen, vendar to razložimo enostavno: približek Kirkwooda namreč ni normaliziran, saj se verjetnosti ne seštejejo v 1.

Vseeno pa lahko definiramo interakcijski prispevek za poljubno število atributov, tako da posplošimo definicijo v (McGill, 1954). Torej, interakcijski prispevek za skupino $\mathcal{S}$ atributov je:

$$I(\mathcal{S}) \triangleq -\sum_{\mathcal{T}\subseteq\mathcal{S}} (-1)^{|\mathcal{S}\setminus\mathcal{T}|} H(\mathcal{T}) = I(\mathcal{S} \setminus X|X) - I(\mathcal{S} \setminus X),\ X \in \mathcal{S} \tag{A.14}$$

A.3 Interakcije in Modeli

A.3.1 Brez-interakcijski približki

Videli smo, da je interakcijski prispevek na nek način primerjava med dvema modeloma, vendar pa en od teh dveh modelov ni pravilen. Da bi prišli do koristnosti interakcije moramo med seboj primerjati dva modela: eden mora dopuščati obstoj interakcije, drugi pa je ne sme dovoliti. Interakcijo lahko potem tu obravnavamo kot omejitev.

Recimo, da imamo nek referenčni model $P(\mathbf{X}|\boldsymbol{\theta})$ na k atributih $\mathbf{X} = [X_1, X_2, \ldots, X_k]^T$. Ta model nima nobenih posebnih omejitev, zato lahko vsebuje poljubne interakcije. Zdaj pa temu modelu preprečimo, da bi vseboval interakcije. To naredimo tako, da najdemo najslabši alternativni model $\hat{P}(\mathbf{X}|\hat{\boldsymbol{\theta}})$, ki pa se še ujema z referenčnim v vseh projekcijah atributov, ki so manjše od k. To zapišemo takole:

$$\hat{\boldsymbol{\theta}} = \arg\max_{\substack{\boldsymbol{\theta}'\in\Re_{\boldsymbol{\Theta}} \\ \forall X_i: \int \hat{P}(\mathbf{X}|\boldsymbol{\theta}')dx_i = \int P(\mathbf{X}|\boldsymbol{\theta})dx_i}} H(\mathbf{X}|\hat{P}(\mathbf{X}|\boldsymbol{\theta}')) \tag{A.15}$$

Uporabili smo princip maksimalne entropije (MaxEnt) (Jaynes, 2003), da bi dobili brez-interakcijski približek (part-to-whole approximation), ki pa se še vedno ujema z referenčnim modelom na vsaki podmnožici $k - 1$ ali manj atributov. V praksi za ta namen uporabimo algoritem *generalized iterative scaling* (GIS) (Darroch and Ratcliff, 1972, Cziszár, 1998), ki ga inicializiramo z enakomerno porazdelitvijo.[1]

Maksimizacija entropije je težak optimizacijski problem z omejitvijo. Dualen mu je problem minimizacije napake modela, ki mu je onemogočeno dopuščati interakcije. Izkaže se, da ima to lastnost Boltzmannova porazdelitev (Darroch et al., 1980, Jaynes, 2003):

$$\hat{P}(\mathbf{x}|\boldsymbol{\beta}) \triangleq \frac{1}{Z}\prod_{\mathcal{S}\subset\mathcal{X}} \psi_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}) = \exp\left\{\sum_{\mathcal{S}\subset\mathbf{X}} \log\psi_{\mathcal{S}}(\mathbf{x}_{\mathcal{S}}) - \log Z\right\} \tag{A.16}$$

[1] Dejansko ta postopek ne poskuša maksimizirati entropije, ampak zadostiti omejitvam, ob čemer pa se poskuša čimmanj oddaljiti od enakomerne porazdelitve.

Namesto omejitev imamo zdaj funkcijo, ki avtomatsko izpolnjuje dane omejitve. $\mathcal{S}$ je prava podmnožica množice atributov $\mathcal{X} = \{X_1, X_2, \ldots, X_k\}$, tako, da $|\mathcal{S}| < k$. Sicer zadošča, da upoštevamo le take potenciale, za katere velja $|\mathcal{S}| = k - 1$, saj so manjši vsebovani v njih. Nenegativni funkciji $\psi(\mathbf{x}_\mathcal{S})$ pravimo potencial na $\mathcal{S}$. Z je particijska funkcija, ki poskrbi, da je Boltzmannova porazdelitev normalizirana:

$$Z \triangleq \int_{\Re_\mathbf{X}} \prod_{\mathcal{S}\subset\mathcal{X}} \psi_\mathcal{S}(\mathbf{x}_\mathcal{S}) d\mathbf{x} \tag{A.17}$$

Parametrizacija $\boldsymbol{\beta}$ določi potenciale. Ker je neobstoj interakcij reda k in več že zagotovljen, moramo za brez-interakcijski približek le izbrati parametre tako, da bo $\hat{P}$ čimbližje referenčni porazdelitvi:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}'} D(P(\mathbf{X}|\boldsymbol{\theta}) \| \hat{P}(\mathbf{X}|\boldsymbol{\beta}')) \tag{A.18}$$

Ta problem ponavadi rešimo z metodo Lagrangovih multiplikatorjev.

Ker pa sta oba zgornja postopka računsko zahtevna, lahko uporabljamo normaliziran Kirkwoodov približek:

$$\hat{P}_K(\mathbf{x}) = \frac{1}{Z} \prod_{\mathcal{S}\subset\mathcal{X}} P(\mathbf{x}_\mathcal{S})^{(-1)^{1+|\mathcal{X}\setminus\mathcal{S}|}} \tag{A.19}$$

Kakršenkoli brez-interakcijski približek $\hat{P}$ že uporabimo, količino interakcije ovrednotimo kot $D(P\|\hat{P})$. Če je vrednotenje količine interakcije v kontekstu napovedovanja $\mathbf{Y}$ iz $\mathbf{X}$, pazimo, da tudi v (A.15) in (A.18) uporabimo ustrezen kriterij, $H(\mathbf{Y}|\mathbf{X}, \hat{P})$ ter $D(P(\mathbf{Y}|\mathbf{X})\|\hat{P}(\mathbf{Y}|\mathbf{X}))$. Torej, zamislimo si, da je referenčnemu modelu dovoljeno upoštevati interakcijo pri napovedovanju C iz A in B, alternativnemu pa ne. Potem je informacijski prispevek zmanjšanje v koristnosti ob prehodu iz referenčnega na alternativni model. Konkretno je tu alternativni model nenormalizirana napoved, ki spominja na naivni Bayesov klasifikator:

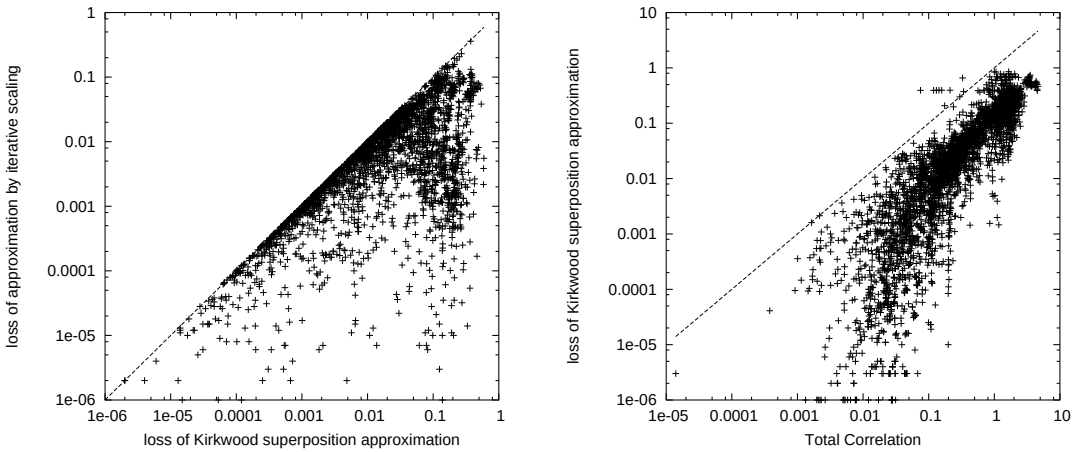
$$I(A;B;C) = D\left(P(C|A,B)\middle\|P(C)\frac{P(A|C)P(B|C)}{P(A)P(B)}\right) = D\left(P(C|A,B)\middle\|\frac{P(C|A)P(C|B)}{P(C)}\right)$$

Na sliki A.4 vidimo, da je približek Kirkwooda slabši od tistega z metodo maksimalne entropije. Vseeno pa je boljši od popolne faktorizacije in boljši od približka, ki predpostavi pogojno neodvisnost.

A.3.2 Preskus značilnosti interakcije

V prejšnjem razdelku smo vedeli, da je naš referenčni model P. Vendar v praksi P ni znan: imamo le predpostavke in podatke. Zato smo glede P negotovi. To predstavlja podlago preskusov značilnosti (significance tests) ter intervalov zaupanja (confidence intervals). Pri preskusih značilnosti nas zanima verjetnost, da bo napaka referenčnega modela večja od napake alternativnega modela in to ob predpostavki, da je referenčni model pravilen. Pri intervalih zanimanja pa se posvetimo verjetnostni porazdelitvi napake med obema modeloma, spet ob predpostavki, da je referenčni model pravilen.

Lahko se odločimo, da bo referenčni model tisti, ki dovoljuje interakcijo. Ob tem je alternativni model njegov brez-interakcijski približek. Tej izbiri bomo tu sledili, čeprav bi lahko izbor tudi obrnili ter vzeli brez-interakcijski približek kot referenčni model. V nadaljevanju bomo tudi predpostavili, da so vsi naši atributi diskretni.



Slika A.4: Normalizirani približek Kirkwooda ima skoraj vedno večjo napako od brez-interakcijskega približka glede na referenčni model (levo). Je pa približek Kirkwooda vseeno skoraj vedno boljši od modela, ki ne dopušča nobene interakcije $P(A)P(B)P(C)$ (desno).

Asimptotični pristop

Predpostavimo nek referenčni model $P(\mathbf{X})$, kjer je $\mathbf{X}$ vektor diskretnih atributov. Naš referenčni model ima popolno svobodo in lahko opisuje poljubne interakcije. Naključno vzorčimo n primerov iz P, in ocenimo $\hat{P}$ z metodo relativne frekvence iz vzorca. KL-divergenca med resnico P in oceno parametrov v pravilnem prostoru hipotez P' ponavadi ne bo nič, četudi je naš prostor hipotez pravilen in četudi se naša ocena parametrov trudi minimizirati napako. Napako med obema modeloma $D(P'\|P)$ lahko opišemo z verjetnostno porazdelitvijo:

$$\frac{2n}{\log_2 e} D(P'\|P) \underset{n\to\infty}{\sim} \chi^2_{|\Re_{\mathbf{X}}|-1} \tag{A.20}$$

Ob tem smo uporabili dejstvo, da je KL-divergenca pomnožena z $2n/\log_2 e$ enaka Wilksovi statistiki G^2. Pri velikih n ima G^2 porazdelitev χ^2_{df}, kjer je df število prostostnih stopenj. Ta asimptotični približek ni dober ko $n/df < 5$. Če bi hoteli vrednotiti značilnost 3-interakcije med tremi 3-vrednostnimi atributi, bi potrebovali vsaj 135 primerov.

Ob zgornji interpretaciji, stopnjo značilnosti (P-value) alternativnega modela $\hat{P}$ ocenimo kot $\phi \triangleq \Pr\left\{\chi^2_{df}(x) \geq \frac{2n}{\log_2 e} D(P\|\hat{P})\right\}$. Če je $\hat{P}$ brez-interakcijski približek, govorimo o stopnji značilnosti interakcije.

Število prostostnih stopenj je odvisno od lastnosti referenčnega modela. Če bi, recimo, referenčni model predpostavil neodvisnost med atributi, bi število prostostnih stopenj bilo enako $\sum_i |\Re_{X_i} - 1|$. Paziti moramo še na to, da če ocenimo zgornji referenčni model iz podatkov, kjer se nekatere vrednosti ne pojavijo, število teh vrednosti odštejemo od števila prostostnih stopenj. Če to naredimo, se zgornja enačba zelo dobro ujema s postopkom prevzorčenja z vračanjem (bootstrap). Ob tem naj opozorimo, da smo predpostavili, da je ocena $D(P\|\hat{P})$ na podlagi omejenega učnega vzorca zanesljiva. To ni nujno utemeljeno in rešitve bomo obravnavali v nadaljevanju.

Prečno preverjanje

Pri asimptotičnem računanju stopnje značilnosti predpostavimo, da je mogoče $D(P\|\hat{P})$ oceniti popolnoma natančno. To ni vedno utemeljeno, zato bomo v nadaljevanju razbili množico danih učnih primerov na učno in testno. Na učni množici ocenimo referenčni model P' ter alternativni $\hat{P}'$, na testni pa se še enkrat naučimo referenčnega, kar označimo z $\dot{P}$. Rezultat tega preizkusa preko velikega števila prečnih preverjanj je CV-vrednost:

$$\nu \triangleq \Pr\{D(\dot{P}\|\hat{P}') \geq D(\dot{P}\|P')\} \tag{A.21}$$

Torej ν pomeni verjetnost, da bo napaka alternativnega modela na testni množici večja od napake referenčnega modela. Če bi referenčni model bil ponavadi boljši, bi se alternativni model premalo prilegal podatkom. Da bi odstranili odvisnost od posamičnega razbitja, moramo prečno preverjanje izvesti velikokrat. To je še posebej pomembno, ko učnih primerov ni veliko.

S pomočjo prečnega preverjanja lahko ocenimo tudi interval zaupanja glede napake posamičnega modela, pa tudi razlike med referenčnim in alternativnim modelom. Odločiti se moramo, ali je osnova za izračun posamično prečno preverjanje, ali posamično razbitje na učno in testno množic primerov. Kakorkoli že, če izvedemo veliko število prečnih preverjanj ali razbitij na istih podatkih in pri istih postopkih učenja, lahko dobimo tudi več različnih rezultatov glede napake in razlik v napakah. Res je pričakovana napaka povprečje teh napak, vendar pa je dobro upoštevati tudi porazdelitev napak.

Ker KL-divergenca ni simetrična, je bolje uporabiti intervale zaupanja, ki temeljijo na percentilih. Simetrični 99% interval zaupanja glede napake alternativnega modela glede na referenčni model temelji na dveh številkah $w_<$ in $w_>$, kjer

$$\Pr\{D(\dot{P}\|\hat{P}) - D(\dot{P}\|P') \leq w_<\} = (100\% - 99\%)/2 \quad \wedge \tag{A.22}$$

$$\Pr\{D(\dot{P}\|\hat{P}) - D(\dot{P}\|P') \geq w_>\} = (100\% - 99\%)/2 \tag{A.23}$$

Verjetnost računamo preko velikega števila preizkusov. Interval potem zapišemo kot $[w_<, w_>]$, kar pomeni, da bo razlika med modeloma v 99% preizkusov s prečnim preverjanjem znotraj tega okvira, izpustili pa bomo zelo velike in zelo majhne vrednosti.

Četudi smo v tem razdelku govorili o prečnem preverjanju, lahko načeloma uporabimo poljubno metodologijo, ki temelji na prevzorčenju (resampling). Za vsako metodologijo potem dobimo svoj nabor značilnosti in intervalov zaupanja.

Bayesovski pristop k preizkusom značilnosti

Načelno mnenje v Bayesovi statistiki je, da preizkusi značilnosti niso primerni. Vsakemu modelu pripišejo svoj prostor hipotez $\mathcal{H}$, kvaliteta modela pa izhaja iz razvidnosti podatkov v tistem kontekstu. Težava tega pristopa je, da ni neposredno skladen z uporabo funkcije koristnosti. Tudi če to potem uvedemo, ponavadi izražajo vse ocene razlik med modeli popolno gotovost, četudi je aposteriorna gotovost na široko razpršena. Nekateri novejši pristopi, kot recimo DIC (Speigelhalter et al., 2003), upoštevajo aposteriorno razpršenost. Mi pa bomo to naredili v kontekstu preskusov značilnosti.

Verjetnost, da je neka možna vrednost parametrov $\hat{\boldsymbol{\theta}}$ z aposteriorno gotovostjo $P(\hat{\boldsymbol{\theta}}|\mathcal{D})$ v referenčnem prostoru hipotez $\mathcal{H}_1$ slabši približek drugi možni vrednosti $\boldsymbol{\theta}'$ z aposteriorno

gotovostjo $P(\boldsymbol{\theta}'|\mathcal{D})$, kot pa neka vrednost parametrov $\hat{\boldsymbol{\theta}}$ iz alternativne družine hipotez $\mathcal{H}_2$:

$$\beta \triangleq \iiint \mathbb{I}\{D_{\boldsymbol{\theta}|\boldsymbol{\theta}'}(P\|P') \geq D_{\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}}(P\|\hat{P})\}P(\boldsymbol{\theta}|\mathcal{D})P(\boldsymbol{\theta}'|\mathcal{D})P(\hat{\boldsymbol{\theta}}|\mathcal{D})d\boldsymbol{\theta}d\boldsymbol{\theta}'d\hat{\boldsymbol{\theta}} \tag{A.24}$$

B-vrednosti nimajo nekaterih lastnosti stopenj značilnosti.

A.3.3 Verjetnostni modeli za zvezne atribute

Interakcije lahko obravnavamo tudi za zvezne atribute. Kot vedno, pa je rezultat odvisen od tega, kakšen model si izberemo. Najprej si poglejmo koncept diferenčne entropije h, ki jo tudi merimo v bitih (Cover and Thomas, 1991):

$$h(\mathbf{X}|p) \triangleq -\int_{\Re_{\mathbf{X}}} p(\mathbf{x}) \log_2 p(\mathbf{x}) d\mathbf{x} = \mathbb{E}_p\{-\log_2 p(\mathbf{x})\} \tag{A.25}$$

Lastnosti diferenčne entropije se včasih razlikujejo od lastnosti navadne entropije (Shannon, 1948). Na primer, diferenčna entropija je lahko negativna ali nič:

$$\sigma \leq 1/\sqrt{2\pi e} : \; h(X|X \sim \text{Normal}(\mu, \sigma)) \leq 0 \tag{A.26}$$

Kullback-Leiblerjeva divergenca pa se pri zveznih atributih ne obnaša bistveno drugače:

$$D(p\|q) \triangleq \int_{\Re_{\mathbf{X}}} p(\mathbf{x}) \log_2 \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \tag{A.27}$$

Ker je analitični izračun entropije za marsikateri model problematičen, pride prav koncept empirične entropije (empirical entropy) $\hat{h}$ pri vzorcu $\mathcal{D}$:

$$\hat{h}(\mathbf{X}|p, \mathcal{D}) \triangleq -\frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \log_2 p(\mathbf{x}). \tag{A.28}$$

Oglejmo si naše informacijske količine na primeru večrazsežne normalne porazdelitve. Imejmo d-razsežni slučajni vektor $\mathbf{X} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$:

$$p(\mathbf{X} = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\} \tag{A.29}$$

$\boldsymbol{\mu}$ je vektor srednjih vrednosti, $\boldsymbol{\Sigma}$ pa kovariančna matrika. Če je $d = 1$, lahko zapišemo diferenčno entropijo v zaključeni obliki kot (Cover and Thomas, 1991):

$$h(X|\mu, \sigma) = \frac{1}{2} \log_2(2\pi e \sigma^2) \tag{A.30}$$

V splošnem pa (Billinger, 2004):

$$h(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{2} \log_2(|2\pi e \boldsymbol{\Sigma}|) \tag{A.31}$$

Tu $|\cdot|$ označuje determinanto.

d-razsežni naključni vektor $\mathbf{X} = [X_1, \ldots, X_d]$ lahko opišemo skupaj v obliki večrazsežne normalne porazdelitve (in dovoljujemo linearno interakcijo) v modelu p. Po drugi strani pa ga lahko opišemo kot produkt neodvisnih enorazsežnih normalnih porazdelitev q:

$$p: \qquad \mathbf{x} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{A.32}$$

$$q: \qquad \mathbf{x} \sim \prod_{i}^{d} \text{Normal}(\mu_i, \sigma_i) \tag{A.33}$$

V primeru $d = 2$ je q brez-interakcijski model. KL-divergenca v tem primeru in Pearsonov koeficient korelacije ρ sta lepo povezana v zaključeni obliki (Billinger, 2004):

$$D(p\|q) = I(X_1; X_2|p) = -\frac{1}{2}\log_2(1-\rho^2) = -\frac{1}{2}\log_2\left(\frac{|\boldsymbol{\Sigma}|}{\sigma_{X_1}\sigma_{X_2}}\right) \tag{A.34}$$

Seveda pa so izračuni z bolj naprednimi modeli, kot so recimo mešanice (mixture models), tudi bolj zapleteni.

A.4 Vizualizacija

A.4.1 Interakcijska analiza

Splošni postopek interakcijske analize na učnem problemu, ki ga upisujemo z atributi $\mathcal{X} = \{A_1, A_2, \ldots, A_m\}$ ima naslednjo obliko:

1. Ustvari eno-atributne projekcije $\mathcal{S}_1 = \{\{A_1\}, \{A_2\}, \ldots, \{A_m\}\}$, dvo-atributne projekcije $\mathcal{S}_2 = \{\{A_1, A_2\}, \{A_1, A_3\}, \ldots, \{A_{m-1}, A_m\}\}$ in tako naprej.

2. Če obstaja razredni atribut Y, ga dodaj vsaki projekciji $S \in \mathcal{S}_1 \cup \mathcal{S}_2 \cup \cdots$.

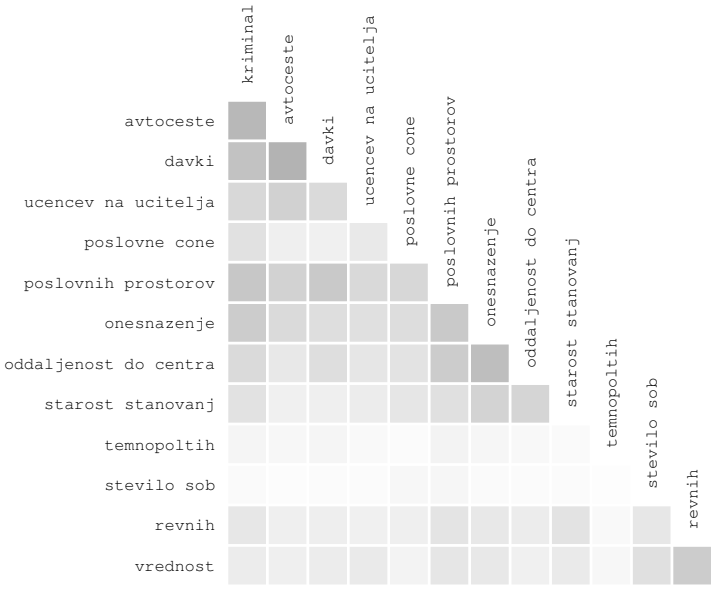
3. Nauči se verjetnostnega modela za vsak S.

4. Izračunaj interakcijski prispevek za vsak S.

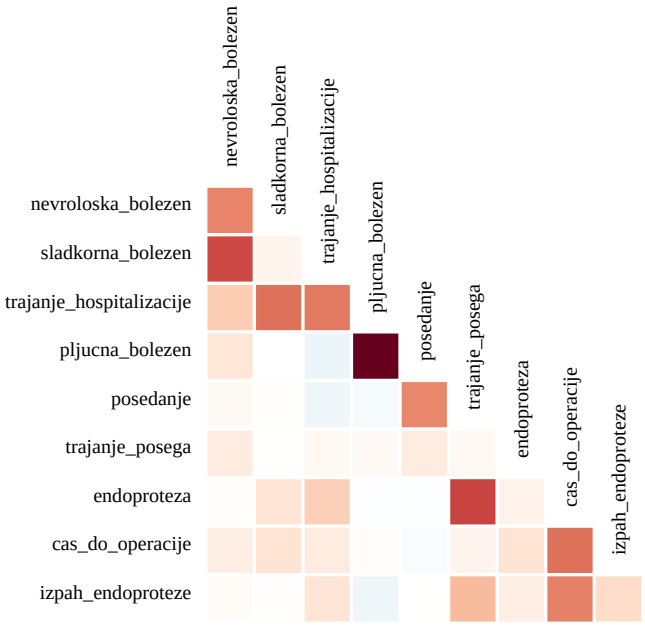
5. Prikaži rezultate:

   - Povzami vzorec interakcij v razumljivi obliki (interakcijska matrika, interakcijski dendrogram, večrazsežno lestvičenje).
   - Izloči najbolj očitne interakcije in jih prikaži v obliki grafa.
   - Osredotoči se na posamezno interakcijo in jo razloži z drugimi postopki, kot so recimo pravila ali skupine..

Zaradi učinkovitosti izvajamo interakcijsko analizo le do neke maksimalnega reda interakcij k. Ponavadi je $k = 2$. Kompleksnost v takem primeru je kvadratična, saj je $\binom{m}{k}$ načinov izbire k atributov izmed m. Označimo lahko več atributov, da bi določili kontekst za interakcijsko analizo, pri čemer je potem vseh preučevanih interakcij manj.



Slika A.5: Povezava med parom atributov pri domeni 'Boston housing' je močna, ko je ustrezni kvadratek med obema atributoma temen, in šibka, ko je svetel. Vidimo, da so močne povezave med avtocestami in davki ter med oddaljenostjo od centra ter onesnaženostjo. Razredni atribut *vrednost* spada v tretjo skupino.



Slika A.6: Na tej ilustraciji prikazujemo pomembnost posamičnih atributov ter pomembnost interakcij med njimi. Razred pri tej domeni je kvaliteta kolka po operaciji. Rdeča barva označuje pozitivne interakcije, modra pa negativne. Šibke interakcije so neobarvane, bele. Po diagonali so prikazane interakcije reda 2 med razredom in enim atributom, drugje pa reda 3 med dvema atributoma in razredom. Najpomembnejši atribut označuje, ali bolnik ima pljučno bolezen. Najpomembnejša interakcija je povezava med tipom endoproteze in trajanjem operacije.

A.4.2 Interakcije med atributi

Najlažje lahko interakcije med atributi in razredom prikažemo v matriki, kjer barva označuje tip in moč interakcije. Slika A.5 prikazuje 2-interakcijski prispevek med vsakim atributom, slika A.6 pa 3-interakcijski prispevek med vsakim parom atributov na domeni HHS.

Če bi radi tako interakcijsko matriko povzeli, se pojavi težava kompleksnosti atributov: pri atributih z več vrednostmi bo v povprečju obseg interakcije večji kot pri atributih z manj vrednostmi, kar spominja na nekatere težave pri ocenjevanju pomembnosti atributov (Kononenko, 1997, Robnik-Šikonja, 2001): Quinlan (1986) je predlagal razmerje informacijskega prispevka (gain ratio), López de Màntaras (1991) pa mero razdalje. Izkaže se, da sta obe rešitvi le posebna primera koeficientov Rajskega. Uporabljali bomo naslednjo definicijo razdalje med atributoma A in B, razdaljo Rajskega (Rajski, 1961), ki tudi izpolnjuje trikotniško neenakost:

$$\langle A, B\rangle_R \triangleq 1 - \frac{I(A;B)}{H(A,B)} \tag{A.35}$$

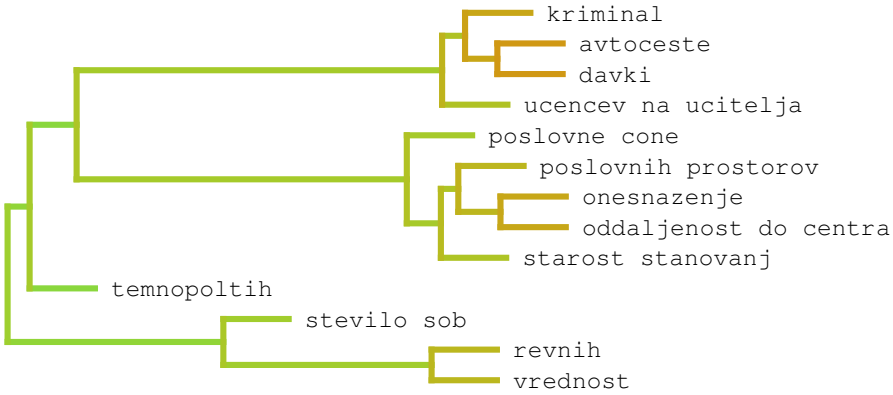
Razdalja Rajskega je vedno med 0 in 1. Ko je 1, sta atributa popolnoma neodvisna, zato daleč. Ko je 0, sta atributa popolnoma odvisna, zato zelo blizu. Ta definicija je uporabna tudi za oceno pomembnosti atributa, če je razredni atribut eden od A in B. Ko pa imamo eksplicitno naveden razred Y, lahko uporabimo kot razdaljo izraz, ki temelji na interakcijskem prispevku:

$$\langle A, B, Y\rangle_R \triangleq 1 - \frac{|I(A;B;Y)|}{H(A,B,Y)} \tag{A.36}$$

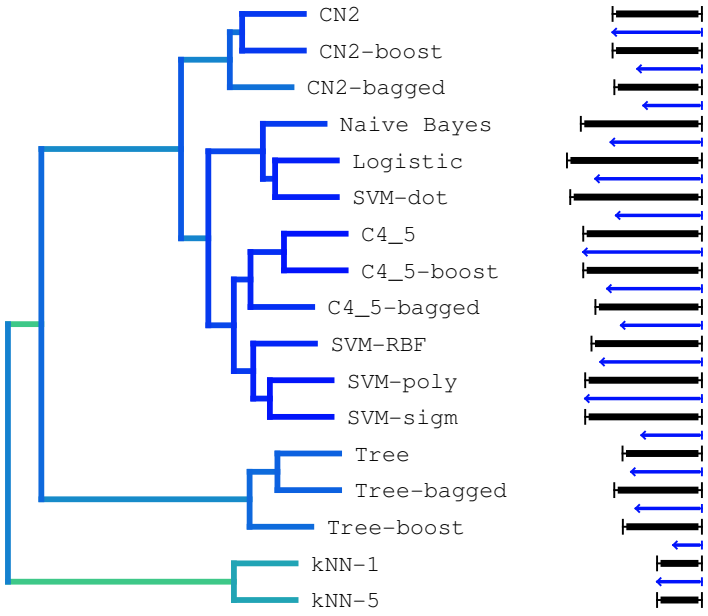
Čeprav ni formalne podlage za to razdaljo, v praksi daje smiselne rezultate. Bistveno je opažanje, da razdalja Rajskega temelji na konceptu funkcije koristnosti ter neke družine modelov. Torej, statistični model in funkcija koristnosti zadoščata, da ustvarimo metrični prostor modelov. S pomočjo razdalje Rajskega lahko iz matrike informacijskih ali interakcijskih prispevkov ustvarimo matriko različnosti ter jo povzamemo s postopki hierarhičnega razvrščanja (Kaufman and Rousseeuw, 1990), kar prikazujeta sliki A.7 in A.8.

V prejšnjem razdelku smo eksplicitno navajali posamične interakcije, ki so izstopale. Lahko pa jih tudi predstavimo v obliki grafa. Pri nadzorovanem učenju si lahko pomagamo z dejstvom, da je medsebojna informacija atributa kvečjemu enaka negotovosti glede razreda, $I(A;Y) \leq H(Y)$. To velja tudi za večje število atributov, na primer $I(A,B;Y) \leq H(Y)$. Zato lahko izrazimo medsebojno informacijo, pa tudi interakcijski prispevek kot delež negotovosti glede razreda $H(Y)$. Pri tem lahko upoštevamo še znano lastnost, $I(A,B;Y) = I(A;Y) + I(B;Y) + I(A;B;Y)$.

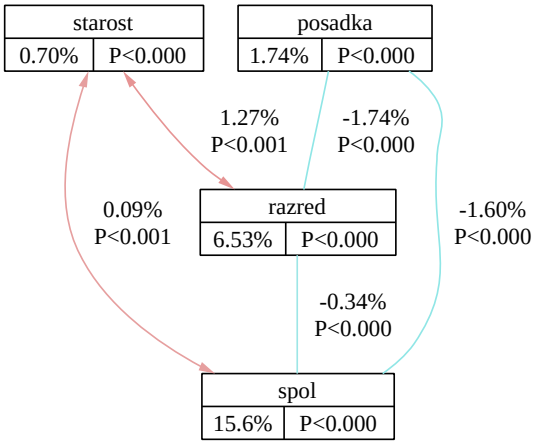
Na sliki A.9 je prikazan interakcijski graf domene 'Titanic'. Kot vidimo, obstaja kar nekaj statistično značilnih interakcij: očitno so konkretna pravila določala, kdo se je lahko rešil. Vidimo, da je bil najpomembnejši kriterij spol, ki je bil odgovoren za 15.6% negotovosti glede preživetja (ženske so imele prednost pri vstopu v rešilne čolne). Pomemben atribut je bil tudi potniški razred, saj so imeli plačniki dražjih vozovnic prednost. Je sicer negativna interakcija med spolom in razredom, vendar ta odpravi le majhen delež negotovosti; če upoštevamo znano lastnost $I(A;Y|B) = I(A;Y) + I(A;B;Y)$, je prispevek razreda k negotovosti potem, ko smo že kontrolirali za spol, še vedno velika: $6.53 - 0.34 = 6.19\%$. Po drugi strani pa vidimo, da že spol razloži velik del večjega števila



Slika A.7: Medsebojno informacijo med posamičnimi pari atributov lahko z razdaljo Rajskega pretvorimo v različnosti in jih povzamemo s hierarhičnih razvrščanjem (metoda *agnes*). Vidimo, da v tej domeni obstajajo tri skupine atributov. Rdeča barva označuje pozitivne iterakcije, zelena pa je nevtralna.



Slika A.8: Če napoved klasifikatorja interpretiramo kot atribut, lahko z interakcijskim dendrogramom prikažemo nekakšno taksonomijo algoritmov strojnega učenja. Za ta dendrogram smo uporabili 10-kratno prečno preverjanje na domeni 'CMC' ter okolje Orange (Demšar and Zupan, 2004). Črne črte označujejo pomembnost atributa, v tem primeru to, kako dobre so napovedi posamičnega klasifikatorja. Logistična regresija se je najboljše obnesla, vendar razlike niso velike. Modre črte označujejo obseg negativne interakcije med dvema sosednima klasifikatorjema.



Slika A.9: To je interakcijski graf domene 'Titanic'. Razred pri tej domeni je preživetje posamičnega potnika. Vsako vozlišče označuje posamičen atribut, in za vsak atribut vidimo, kolikšen delež entropije razreda odpravi, ter kakšna je njegova stopnja značilnosti. Pozitivne interakcije so označene z rdečimi dvosmernimi puščicami, negativne pa z modrimi črtami.

ponesrečencev med posadko, $1.74 - 1.60 = 0.14\%$. Očitna je negativna interakcija med razredom in posadko: atribut *razred* nam pove tudi to, ali gre za člana posadke ali za potnika, zato nam atribut *posadka* ne pove nič novega, ko že poznamo vrednost atributa *razred*.

Med pozitivnimi interakcijami je najmočnejša tista med starostjo in razredom. Razlaga je žalostna a enostavna: vsi otroci iz prvega in drugega razreda so preživeli, skoraj dve tretjini otrok iz tretjega razreda pa je umrlo. Druga pozitivna interakcija je podobno žalostna: med otroci tretjega razreda se je rešilo nadpovprečno veliko dečkov, manj pa deklic. Primerov teh pozitivnih interakcij v filmu Titanic nismo videli, kot tudi ne tega, da je umrlo zelo malo ženskih članov posadke. Razvidno je tudi, da so vse te interakcije statistično značilne.

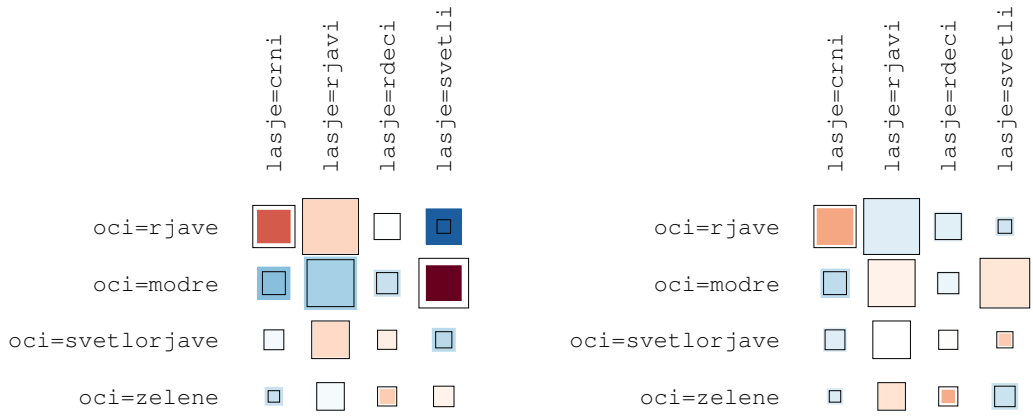
A.4.3 Notranjost interakcije

Pravila

Že pri prejšnjem primeru smo videli, da je mogoče razložiti vzroke interakcije pri posamičnih vrednostih atributov. Za take naloge ponavadi uporabimo t.i. mozaične diagrame. Lahko pa tudi ustvarimo matriko hkratnih vrednosti dveh atributov, z velikostjo označimo število primerov, z barvo pa napako brez-interakcijskega modela. Primer na sliki A.10 prikazuje interakcijo med atributoma *barva oči* in *barva las*. Da zmanjšali vpliv naključja, barvo napake določimo na podlagi Pearsonovih standardnih ostankov, ki so za n učnih primerov porazdeljeni po standardizirani normalni porazdelitvi Normal$(0, 1)$:

$$d(a, b) \triangleq \sqrt{n}\frac{P(a, b) - P(a)P(b)}{\sqrt{P(a)P(b)(1 - P(a))(1 - P(b))}} \tag{A.37}$$

Čeprav bi lahko model, ki dopušča interakcijo, opisali kar s celotnim kartezičnim produktom, obstaja krajša pot s pravili. Seveda moramo pravila obravnavati v kontekstu



(brez omejitev) (z omejitvijo)

Slika A.10: Prikazana je interakcija med barvo las in barvo oči. Velikost črnega kvadratka iz črt je proporcionalna verjetnosti tega izida. Velikost barvnega kvadrata označuje napovedano verjetnost, barva pa napako. Rdeča barva pomeni premajhno napoved, modra pa preveliko. Prvi model (levo) je brez interakcij, pri čemer sta najbolj izraziti napaki premajhne napovedi sopojavitve modrih oči in svetlih las ter previsoke napovedi sopojavitve rjavih oči in svetlih las. Če uvedemo pravilo R, KL-divergenca interakcije pade iz 0.178 na 0.045. Naslednji dve dobri pravili bi lahko povezovali rjave oči in črne lase ter zelene oči in rdeče lase.

verjetnostnih porazdelitev. Možna rešitev je tvorjenje konjunkcij med vrednostmi atributov (Kononenko, 1991). Tule pa bomo tvorili nov atribut, R, ki je definiran na podlagi atributov O (barva oči) in L (barva las) kot:

$$R(O,L) \triangleq \begin{cases} 1; & (O = \text{modre} \vee O = \text{zelene}) \wedge (L = \text{svetli}), \\ 0; & \text{sicer.} \end{cases} \tag{A.38}$$

Dopustimo le interakcijo med R in O ter med R in L. S tem atributom, ki ustreza pravilu "Modre in zelene oči so povezane s svetlimi lasmi." smo tako rekoč odpravili interakcijo. Če bi bilo prekrivajočih se pravil več, bi jih lahko obravnavali kot omejitve pri uporabi postopka GIS (Darroch and Ratcliff, 1972). Povrh tega se kompleksnost indukcije pravil zmanjša, saj je iskanje pravil potrebno le znotraj najdenih interakcij.

Taksonomije

Možno je definirati razdaljo med dvema vrednostma atributa. Osnovna ideja leži v vprašanju, ali bomo kaj izgubili pri sposobnosti napovedovanja razreda in drugih atributov, če ti dve vrednosti združimo. Recimo, da imamo atribut A, ki bi ga radi preučili. Najprej moramo ustvariti pomožni atribut $\dot{A}_{i,j}$ za vsak par vrednosti a_i in a_j v $\Re_A$:

$$\dot{a}_{i,j} \triangleq \begin{cases} 1 & ; \ A = a_i \ \vee \ A = a_j \\ 0 & ; \ \text{sicer.} \end{cases} \tag{A.39}$$

Vrednostna razdalja spominja na razdaljo Rajskega in je definirana v nekem kontekstu množice atributov $\mathcal{V}$:

$$\langle a_i, a_j \rangle_V \triangleq 1 - \frac{I(A; \mathcal{V} | \dot{A}_{i,j} = 1)}{H(A, \mathcal{V} | \dot{A}_{i,j} = 1)} \tag{A.40}$$

Če nas zanima le uporabnost razlikovanja med parom vrednosti pri napovedovanju razreda Y, lahko rečemo $\mathcal{V} = \{Y\}$. Ker je pri velikem $\mathcal{V}$ težavno oceniti model, na podlagi katerega bi potem izračunali medsebojno informacijo, lahko uporabimo Bethejev približek (Yedidia et al., 2004):

$$\langle a_i, a_j \rangle_V = 1 - \frac{\sum_{X \in \mathcal{V}} I(A; X | \dot{A}_{i,j} = 1)}{(1 - |\mathcal{V}|) H(A | \dot{A}_{i,j} = 1) + \sum_{X \in \mathcal{V}} H(A, X | \dot{A}_{i,j} = 1)} \tag{A.41}$$

S pomočjo tega približka in hierarhičnega razvrščanja smo izračunali smiselno taksonomijo držav rojstva prebivalcev ZDA na podlagi njihovih atributov, kot so opisani v domeni 'adult/census' (slika A.11). Tako taksonomijo je mogoče s pridom uporabiti tudi pri strojnem učenju.

A.5 Interakcije pri razvrščanju

Prejšnji razdelek je pokazal uporabnost interakcij pri prikazu širših odnosov med atributi, izločanja pomembnih interakcij ter tvorjenja novih atributov oziroma struktur znotraj atributov. Nemogoče bi bilo temeljito obravnavati vse možne uporabe, lahko pa se osredotočimo na dve: izbiro atributov v kontekstu naivnega Bayesovega klasifikatorja, ter posplošitev naivnega Bayesovega klasifikatorja, kjer je model definiran z množico interakcij na atributih.

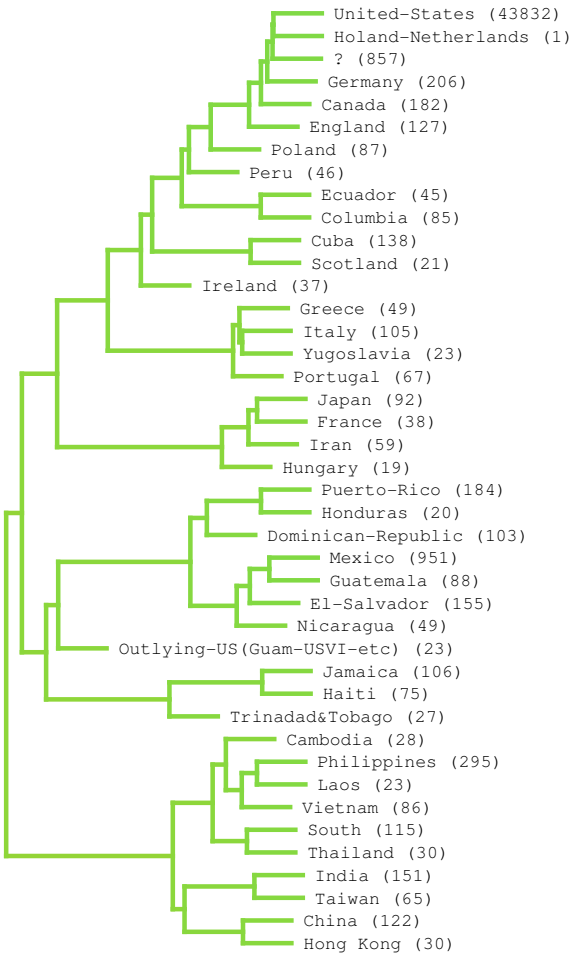
A.5.1 Izbira atributov z interakcijskim prispevkom

Pri naivnem Bayesovem klasifikatorju sta dva običajna postopka izbire atributov: uporaba internega prečnega preverjanja pri vodenju pohlepnega iskanja (Langley and Sage, 1994), ter uporaba ocen pomembnosti atributov pri rangiranju. V tem razdelku si bomo ogledali dva nova algoritma: prvi izvaja pohlepno iskanje kar na učni množici, saj se problemi povezani s kršenjem predpostavke o pogojni neodvisnosti pri razredu opazijo že tam; drugi algoritem uporablja interakcijsko informacijo kot hevristiko.

Recimo, da imamo na razpolago množico atributov $\mathcal{X}$ za napovedovanje razreda Y. Na neki točki iskanja smo nekaj atributov $\mathcal{A}$ že vključili v naš model, nekaj pa jih še lahko $\bar{\mathcal{A}}$. Pohlepni način izbire atributov se odloči za tisti atribut, ki bo napako naivnega Bayesovega klasifikatorja čimbolj zmanjšal. Če uporabimo funkcijo napake $D(P\|Q)$, kjer je P resnica, Q pa približek, lahko pohlepni algoritem surove sile opišemo z naslednjo izbiro:

$$\hat{X} = \arg\min_{X \in \bar{\mathcal{A}}} D\left(P(Y | \mathcal{A}, \bar{\mathcal{A}}) \middle\| \frac{1}{Z} P(Y) P(X|Y) \prod_{A \in \mathcal{A}} P(A|Y) \right) \tag{A.42}$$

V praksi seveda ne poznamo resnice, lahko pa računamo logaritemsko napako verjetnostnih napovedi, minimizacija katere je precej podobna zgornji enačbi. Algoritem se ustavi, ko vsak od atributov v $\bar{\mathcal{A}}$ kvečjemu poslabša napako na učni množici. Algoritem surove sile je lahko hiter, če imamo kvalitetno implementacijo naivnega Bayesovega klasifikatorja, je



Slika A.11: V tem dendrogramu lepo vidimo, da lahko z nekaj izjemami ločimo tri skupine izvora prebivalcev ZDA: Azija, Južna in Srednja Amerika ter Evropa in Severna Amerika. V oklepajih je navedeno število prebivalcev v posamični skupini.

pa vseeno približno kvadratičen s številom atributov. Vsako preverjanje seveda zahteva pregled vseh učnih primerov.

Malce hitrejši pristop temelji na neki oceni koristnosti atributa za napovedovanje razreda, $I(X;Y)$:

$$\hat{X} = \arg\max_{X \in \bar{\mathcal{A}}} I(X;Y) \tag{A.43}$$

Seveda pa je ta postopek kratkoviden in ne upošteva tega, da nam lahko en atribut pove isto kot drugi. Zato je Fleuret (2004) predlagal naslednji postopek:

$$\hat{X} = \arg\max_{X \in \bar{\mathcal{A}}} \left(\min_{A \in \mathcal{A}} I(X;Y|A) \right) \tag{A.44}$$

Izberemo torej najboljši atribut, ki v njemu najmanj ugodnem kontekstu atributa A vseeno doprinese čimveč. Ta postopek je še vedno kvadratičen s številom atributov, vendar pa je preverjanje takorekoč takojšnje. Slabost postopka je, da je nerobusten, saj upošteva le en atribut za oceno kvalitete. Druga slabost je, da ne loči med pozitivnimi in negativnimi interakcijami. Zato predlagamo naslednji postopek:

$$\hat{X} = \arg\max_{X \in \bar{\mathcal{A}}} \left(I(X;Y) + \sum_{A \in \mathcal{A}} \min\{0, I(X;Y;A)\} \right) \tag{A.45}$$

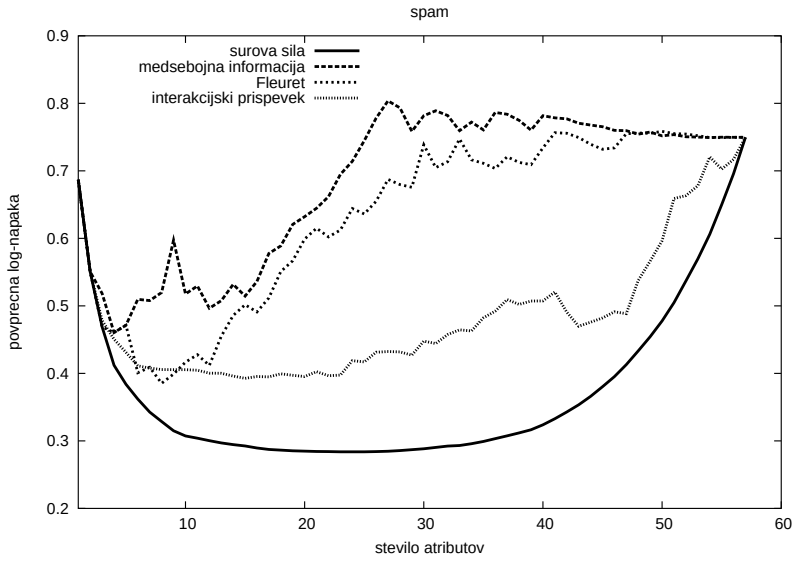
Tu seštejemo negativne interakcije vseh atributov, ki so že vključeni v model, poleg tega pa preprečujemo upoštevanje pozitivnih interakcij. Vsi trije hevristični postopki se ustavijo, ko ni več izboljšanja na učni množici.

Slika A.12 prikazuje delovanje teh treh hevristik na domeni 'Spam'. Uporabili smo logaritemsko funkcijo napake, kjer klasifikator kaznujemo z $-\log_2 p$. Tu je p verjetnost, ki jo je klasifikator pripisal pravilnemu razredu. Obširnejši preskus je prikazan v tabeli A.1. Očitno je najboljši postopek surove sile, sledi mu pa hevristika, ki temelji na interakcijskem prispevku. Sicer med Fleuretovo in našo hevristiko ni spektakularnih razlik, je pa naša precej bolj robustna. Najbolj očitno pa je izredno slabo delovanje kratkovidne ocene, ki je le malenkost boljša od popolne odsotnosti izbire atributov. Manjša učinkovitost informacijskega prispevka pri naivnem Bayesovem klasifikatorju se sklada tudi z ugotovitvami drugih (Mladenić, 1998).

A.5.2 Kikuči-Bayesov klasifikator

Že Demšar (2002) je opazil, da je interakcijski prispevek dobra hevristika za združevanje atributov, kar je sestavni del konstruktivne indukcije (Zupan, 1997). Že v prejšnjem razdelku je postalo jasno, da moramo za konkurenčne rezultate pri gradnji modela uporabljati postopke surove sile, kar pa je z uporabo novejših programskih tehnik lahko dokaj učinkovito (Caruana et al., 2004). Poleg tega bomo uporabili metode Bayesove statistike za sestavljanje modelov in preprečevanje pretiranega prilagajanja podatkom.

Najprej definirajmo pojem interakcijskega modela. Interakcijski model na atributih $\mathcal{V} = \{X_1, X_2, \ldots, X_m\}$ določa neka množica interakcij $\mathcal{M} = \{P(\mathcal{S}|\boldsymbol{\theta}_{\mathcal{S}});\ \mathcal{S} \subseteq \mathcal{V}\}$. Vsako interakcijo na atributih $\mathcal{S}$ opišemo s podmodelom $P(\mathbf{V}_{\mathcal{S}})$, s skupno verjetnostno porazdelitvijo atributov, ki so udeleženi v interakciji. Tu ne povemo, kako točno ta podmodel izgleda. Čeprav sami uporabljamo multinomski model na diskretnih atributih, bi



Slika A.12: Primerjava kontekstno-odvisne izbire atributov pri naivnem Bayesovem klasifikatorju pokaže, da sicer najboljše deluje Fleuretova hevristika. Po drugi strani pa je interakcijski prispevek, ki kaznuje atribute, udeležene v veliko število negativnih interakcij z drugimi atributi, veliko bolj gladko krivuljo, saj je bolj robustna.

lahko uporabili karkšnegakoli (drevesa, pravila, mešanice, ipd.). Ravno tako lahko enostavno predpostavimo tudi to, da nobena od podmnožic atributov $\mathcal{S}$ ni vsebovana v kaki drugi $\mathcal{S}' \supseteq \mathcal{S}$: v takem primeru bi bilo trivialno marginalizirati $P(\mathbf{V}'_{\mathcal{S}})$. Ravno tako lahko za probleme klasifikacije enostavno predpostavimo, da vsaka podmožica $\mathcal{S}$ vsebuje razredni atribut: sicer ta interakcija ne bi nič prispevala h kvaliteti napovedi. Končno, lahko si zaželimo konsistentnosti med podmodeli: načeloma naj bi obstajal nek skupni (joint) model $P(\mathbf{V})$, tako da je vsak posamičen podmodel marginalizacija le-tega.

Kikučijev približek

Osnovna problema, ki sta povezana z interakcijskimi modeli sta dva: kako na podlagi takih posamičnih interakcij napovedujemo razred in kako se iz podatkov naučimo strukture interakcij ter podmodelov. Prej smo že omenili Boltzmannovo porazdelitev, ki ima lepe lastnosti. Uporabili bi jo lahko tako, da bi vsaki interakciji pripisali potencial, potem pa poskušali najti take parametre potencialom, da bi se celotna porazdelitev ujemala s podmodeli. Razen v posebnih primerih bi to bilo precej počasno.

Po drugi strani smo na primeru približka Kirkwooda videli, da lahko pridemo tudi do enostavnejših struktur, ki ne delujejo slabo, moramo jih le normalizirati pred uporabo. Kikučijev približek (Kikuchi, 1951, Yedidia et al., 2004), ki je bil zamišljen kot približek pri računanju entropije, lahko obravnavamo kot posplošitev verižnega pravila (chain rule). Bistvo približka je v tem, da območje prekrivanja med dvema interakcijama odštejemo iz vpliva. Pri tem uporabimo postopek na sliki A.13, da dobimo graf regij. Na podlagi grafa regij $\mathcal{G}_R$ lahko izračunamo skupno verjetnostno porazdelitev:

$$\hat{P}'(\mathbf{v}|\mathcal{M}) \propto \prod_{\langle \mathcal{R}, c_{\mathcal{R}} \rangle \in \mathcal{G}_R} P(\mathbf{v}_{\mathcal{R}})^{c_{\mathcal{R}}} \tag{A.46}$$

	Brierjeva napaka				
	vsi atributi	surova sila	medsebojna inf.	Fleuret	interakcijski p.
lung	0.575 ± 0.085 √	**0.562 ± 0.079**	0.581 ± 0.085 √	0.567 ± 0.088 √	0.634 ± 0.069 √
soy-small	0.000 ± 0.000	**0.000 ± 0.000**	0.000 ± 0.000	0.000 ± 0.000	0.000 ± 0.000
zoo	0.125 ± 0.047 √	0.133 ± 0.046 √	**0.106 ± 0.038**	0.123 ± 0.046 √	0.125 ± 0.047 √
lymph	0.537 ± 0.056	**0.326 ± 0.033**	0.360 ± 0.027	0.368 ± 0.028	0.330 ± 0.018 √
wine	0.013 ± 0.008	0.008 ± 0.006 √	0.020 ± 0.011	0.017 ± 0.010	**0.006 ± 0.003**
glass	**0.618 ± 0.047**	0.630 ± 0.051 √	**0.618 ± 0.047**	**0.618 ± 0.047**	0.635 ± 0.053 √
breast	0.208 ± 0.016	0.198 ± 0.011 √	**0.192 ± 0.010**	0.194 ± 0.010 √	0.194 ± 0.010 √
ecoli	1.070 ± 0.056	**0.916 ± 0.042**	1.056 ± 0.054	1.070 ± 0.056	1.044 ± 0.049
horse-colic	1.013 ± 0.074	**0.423 ± 0.016**	0.426 ± 0.012 √	0.431 ± 0.012 √	0.431 ± 0.012 √
voting	0.091 ± 0.010	**0.037 ± 0.008**	0.044 ± 0.010 √	0.037 ± 0.009 √	0.040 ± 0.010 √
monk3	0.043 ± 0.004 √	**0.043 ± 0.004**	0.043 ± 0.004 √	0.043 ± 0.004 √	0.043 ± 0.004 √
monk1	0.174 ± 0.008 √	0.175 ± 0.008 √	**0.174 ± 0.008**	**0.174 ± 0.008**	**0.174 ± 0.008**
monk2	**0.229 ± 0.006**	0.229 ± 0.006 √	0.229 ± 0.006 √	0.229 ± 0.006 √	0.229 ± 0.006 √
soy-large	0.832 ± 0.095	**0.696 ± 0.088**	1.566 ± 0.103	1.039 ± 0.087	0.803 ± 0.082
wisc-cancer	**0.023 ± 0.004**	0.028 ± 0.005	0.031 ± 0.005	0.027 ± 0.005	0.026 ± 0.005 √
australian	0.112 ± 0.008 √	0.104 ± 0.008 √	0.111 ± 0.010 √	0.104 ± 0.008 √	**0.103 ± 0.010**
credit	0.111 ± 0.007	0.104 ± 0.007 √	0.110 ± 0.007	**0.100 ± 0.007**	0.102 ± 0.007 √
pima	0.160 ± 0.006	0.154 ± 0.005 √	0.154 ± 0.005 √	0.155 ± 0.005 √	**0.151 ± 0.005**
vehicle	0.589 ± 0.021	**0.446 ± 0.020**	0.584 ± 0.013	0.494 ± 0.019	0.487 ± 0.021
heart	0.713 ± 0.024	**0.664 ± 0.019**	0.696 ± 0.021	0.691 ± 0.021	0.670 ± 0.019 √
german	0.174 ± 0.007 √	**0.172 ± 0.005**	0.172 ± 0.007 √	0.174 ± 0.008 √	0.176 ± 0.008 √
cmc	0.445 ± 0.010	0.417 ± 0.006 √	**0.416 ± 0.006**	**0.416 ± 0.006**	**0.416 ± 0.006**
segment	0.262 ± 0.015	**0.150 ± 0.008**	0.287 ± 0.018	0.161 ± 0.009	0.200 ± 0.014
krkp	0.092 ± 0.004	**0.074 ± 0.002**	0.085 ± 0.003	0.078 ± 0.003	0.081 ± 0.002
mushroom	0.034 ± 0.003	**0.005 ± 0.001**	0.008 ± 0.000	0.008 ± 0.000	0.008 ± 0.001
adult	0.120 ± 0.002	**0.098 ± 0.001**	0.117 ± 0.002	0.101 ± 0.001	0.101 ± 0.001

domen	**NB**	**B**	**MI**	**F**	**I**
najboljša	3	15	5	4	5
dobra √	6	10	8	10	13
slaba	17	1	13	12	8

Tabela A.1: Najslabši postopek je odsotnost izbire atributov, najboljši pa uporaba surove sile. Med hevristikami je najslabša kratkovidna ocena z informacijskim prispevkom, najboljša pa uporaba interakcijskega prispevka, saj je ta največkrat med zmagovalci (√ pomeni, da je rezultat znotraj standardne napake najboljšega za domeno).

Približek Kirkwooda je posebni primer Kikučijevega približka za običajni izbor interakcij, ki definirajo brez-interakcijski model nekega višjega reda. $\hat{P}'$ v splošnem ni normaliziran in v splošnem je računanje particijske funkcije Z eden od najzoprnejših problemov pri delu z Boltzmannovo porazdelitvijo. Pri klasifikaciji pa to ni težava, saj normalizacijo izvedemo le pri dani vrednosti atributov $\mathbf{x}$ za razredni atribut Y:

$$\hat{P}(y|\mathbf{x},\mathcal{M}) \triangleq \frac{\hat{P}'(y,\mathbf{x}|\mathcal{M})}{\sum_{y'\in\Re_Y}\hat{P}'(y',\mathbf{x}|\mathcal{M})} \tag{A.47}$$

Učenje strukture interakcij

Naš algoritem učenja strukture izvaja iskanje po principu surove sile, pri čemer najprej preveri vse 2-interakcije. Ko s temi ne more več doseči izboljšanja, se loti 3-interakcij. Ko tu ne more več doseči izboljšanja, gre na 4-interakcije in tako naprej. Algoritem pa ima

$\mathcal{R}_0 \leftarrow \{\emptyset\}$ {Začetne regije.}
for all $\mathcal{S} \in \mathcal{M}$ **do** {za vsako začetno regijo}
  if $\forall \mathcal{S}' \in \mathcal{R}_0 : \mathcal{S} \not\subseteq \mathcal{S}'$ **then**
    $\mathcal{R}_0 \leftarrow \mathcal{R}_0 \cup \{\mathcal{S}\}$ {$\mathcal{S}$ ni odvečna}
  end if
end for
$\mathcal{R} \leftarrow \{\langle \mathcal{S}, 1\rangle;\ \mathcal{S} \in \mathcal{R}_0\}$
$k \leftarrow 1$
while $|\mathcal{R}_{k-1}| > 2$ **do** {možne podmnožice}
  $\mathcal{R}_k \leftarrow \{\emptyset\}$
  for all $\mathcal{I} = \mathcal{S}^\dagger \cap \mathcal{S}^\ddagger : \mathcal{S}^\dagger, \mathcal{S}^\ddagger \in \mathcal{R}_{k-1}, \mathcal{I} \notin \mathcal{R}_k$ **do** {možni preseki}
    $c \leftarrow 1$ {številnost preseka}
    for all $\langle \mathcal{S}', c'\rangle \in \mathcal{R}, \mathcal{I} \subseteq \mathcal{S}'$ **do**
      $c \leftarrow c - c'$ {upoštevaj številnost vseh regij, ki vsebujejo presek}
    end for
    if $c \neq 0$ **then**
      $\mathcal{R} \leftarrow \mathcal{R} \cup \{\langle \mathcal{I}, c\rangle\}$
    end if
    $\mathcal{R}_k \leftarrow \mathcal{R}_k \cup \{\mathcal{I}\}$
  end for
end while
return $\{\langle R, c\rangle \in \mathcal{R}; c \neq 0\}$ {Graf regij s številnostmi.}

Slika A.13: Na podlagi množice interakcij interakcijskega modela $\mathcal{M} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_\ell\}$ ta algoritem izdela graf regij Kikučijevega približka.

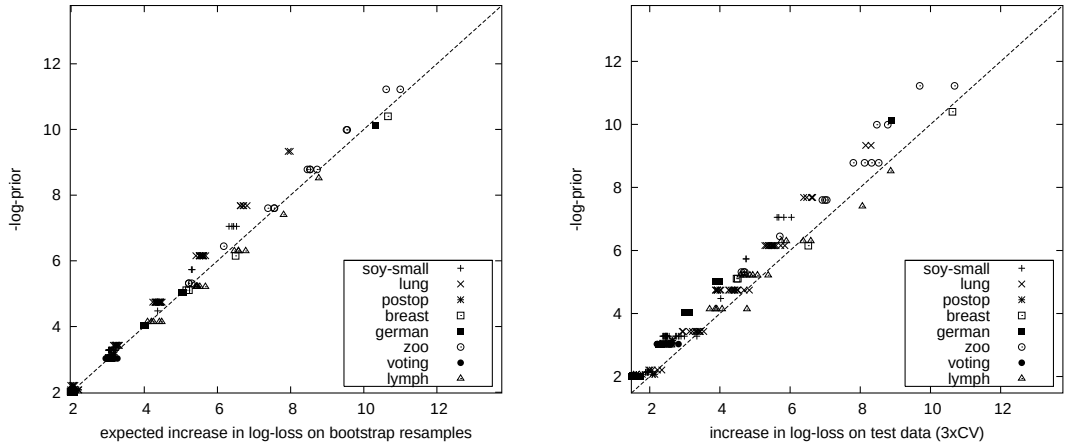
še dve lastnosti. Enostavnejše strukture imajo večjo apriorno gotovost, s tem omejimo globino iskanja ter preprečujemo preveliko prileganje podatkom. Druga lastnost pa je uporaba Bayesovskega povprečenja modelov: namesto, da bi uporabljali le eno strukturo za napovedovanje, združimo več struktur glede na njihovo aposteriorno gotovost.

Najprej moramo ovrednotiti kompleksnost modela. Uporabili smo naslednjo definicijo prostostnih stopenj, ki temelji na prilagoditvi kompleksnosti loglinearnih modelov (Krippendorff, 1986) za namene klasifikacije:

$$df_{\mathcal{M}_\mathcal{Y}} \triangleq \sum_{\langle \mathcal{S}, c\rangle \in \mathcal{R}} c\left(\prod_{X \in \mathcal{S}} |\Re_X| - \prod_{\substack{X \in \mathcal{S} \\ X \notin \mathcal{Y}}} |\Re_X| \right) \tag{A.48}$$

Kompleksnost torej izhaja iz produkta kardinalnosti zalog vrednosti atributov, udeleženih v vsako interakcijo, pri čemer pa upoštevamo, da je pri vsaki vrednosti samih atributov $\mathbf{x}$ napoved vrejetnosti razrednega atributa $P(Y|\mathbf{x})$ normalizirana, kar zmanjša število prostostnih stopenj za ena. To kompleksnost potem vstavimo v naslednji izraz za varčno apriorno gotovost (parsimonious prior), ki ustreza informacijskemu kriteriju Akaikeja (AIC) s korekcijo za majhne vzorce (Burnham and Anderson, 2002).

$$P(\mathcal{M}) \overset{\triangle}{\propto} \exp\left\{ -\frac{m\ df_\mathcal{M}}{m - df_\mathcal{M} - 1} \right\} \tag{A.49}$$



Slika A.14: Negativni logaritem varčne apriorne gotovost se dobro ujema s pričakovano logaritmično napako na neodvisnem vzorcu pri prevzorčenju (left) in z logaritmično napako na testni množici pri 3-kratnem prečnem preverjanju.

Slika A.14 dokazuje, da se ta varčna apriorna gotovost zelo dobro ujema z empiričnimi meritvami pričakovane napake. Seveda bi za drugačno funkcijo napake morali prilagoditi tudi izraz za apriorno gotovost.

Posamični podmodel ocenimo z naslednjim izrazom za apriorno gotovost, ki zagotavlja konsistentnost vseh podmodelov, če je le ϑ za vse enak:

$$P(\boldsymbol{\theta}_{\mathcal{S}}|\vartheta) = \text{Dirichlet}(\alpha, \ldots, \alpha), \;\; \alpha = \frac{\vartheta}{\prod_{i=1}^{k} |\mathcal{X}_i|} \tag{A.50}$$

V praksi to pomeni, da vsako verjetnost sopojavitve konjunkcije vrednosti atributov izračunamo kot $(\alpha + n)/(N + \vartheta)$, kjer je n število pojavitev konjunkcije, N pa število vseh primerov. Gre torej za varianto m-ocene (Cestnik, 1990), le da ocenjujemo skupne verjetnosti in ne pogojnih.

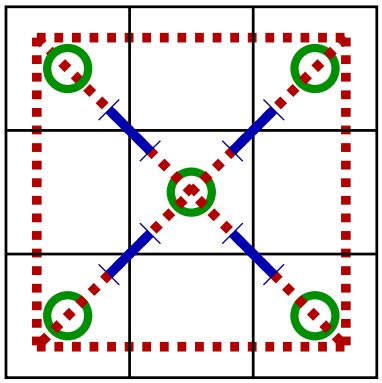
Potem, ko smo s postopkom preiskovanja prišli do točke, ko se je kvaliteta modela na učni množici začela slabšati, z naslednjim izrazom za vsakega od korakov pri iskanju izračunamo aposteriorno gotovost modela pri tistem koraku na učni množici $\mathcal{D} = \{\langle \mathbf{x}^{(1)}, y^{(1)} \rangle, \ldots, \langle \mathbf{x}^{(m)}, y^{(m)} \rangle\}$:

$$\hat{P}(\mathbf{v}^{(1)\ldots(m)}|\mathcal{M}_{\mathcal{Y}}) \triangleq \prod_{i=1}^{m} \hat{P}(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}, \mathcal{M}_{\mathcal{Y}}) \tag{A.51}$$

Če smo pri vsakem od skupno ℓ korakov dodali neko novo interakcijo $\mathcal{S}_i$ in dobili model z aposteriorno gotovostjo p_i, napovedi združimo z naslednjim izrazom za Bayesovsko povprečenje modelov (Hoeting et al., 1999):

$$\hat{P}(y|\mathbf{x}) = \frac{\sum_{i=1}^{\ell} p_i \hat{P}(y|\mathbf{x}, CVA(\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_i))}{\sum_{i=1}^{\ell} p_i} \tag{A.52}$$

S tem smo pa tudi definirali osnovno idejo Kikuči-Bayesovega klasifikatorja. Četudi se zdi opis zaradi poskusa zaokroženosti tega opisa zapleten, gre v resnici za zaporedje popolnoma standarnih korakov Bayesove statistike.



Slika A.15: Plošča pri igrici križcev in krožcev ima 9 polj. Vsako polje opišemo s 3-vrednostnim atributom, ki ima zalogo vrednosti $\{\times, \circ, \_\}$. Razred označuje, ali je z atributi opisan položaj zmagovalen za $\times$ ali ne: to je dvovrednostni atribut. Izbrane interakcije so: 2-interakcije (5 zelenih krogov), 3-interakcije (4 modre črte s črtico), in 4-interakcije (6 rdečih črtkanih črt). Vsaka interakcija vsebuje razredni atribut.

Algoritem na sliki A.13 daje točne (exact) rezultate v primerjavi z verižnim pravilom takrat, ko preseki presekov in nadaljne korekcije niso potrebne (Yedidia et al., 2001). Poleg Kikučijevega obstaja tudi Bethejev približek, ki smo ga uporabili že prej in ki temelji na tem, da korekcijo prekrivanja opravimo kar na posamičnih atributih. Podobne približke v malo drugačnem kontekstu omenja tudi Kononenko (1990), kjer so številnosti kar realna števila. V splošnem pa moramo včasih neugodne ciklične strukture interakcij poenostaviti tako, da cikel združimo v eno samo veliko interakcijo. Nesmiselno pa je to početi vnaprej, če tudi neugodne prinesejo včasih korist. Če obstaja korist v združitvi, bo to opravil že algoritem učenja strukture sam s tem, da bo ustvaril večjo interakcijo. Pomen odpravljanja neugodnih topologij je zato smiselno le takrat, ko želimo zmanjšati kompleksnost preiskovanja.

Eksperimentalni rezultati

Kikuči-Bayesov klasifikator daje presenetljivo dobre rezultate na velikem številu domen in deluje precej hitro. Rezultati na 46 domenah s petkrat ponovljenim 5-kratnim prečnim preverjanjem so prikazani v tabeli A.2. Aposteriorno najbolj gotov model na domeni 'Tic-Tac-Toe' je prikazan na sliki A.15 in lepo prikazuje to, da je algoritem popolnoma avtomatsko našel smiselne strukture: center in kote igralne plošče, povezave med koti in centrom, ter dve diagonali in štiri robove. Četudi doseže logistična regresija boljšo klasifikacijsko točnost z obteženo vsoto na podlagi 2-interakcij med posamičnimi atributi in razredom, se zdi človeku interakcijski model intuitivno veliko bližje.

V dodatnih eksperimentih, ki v tem povzetku niso omenjeni, se je izkazalo, da je varčna apriorna gotovost konservativna glede na 5-kratno ali 10-kratno prečno preverjanje. To pomeni, da lahko postopek učenja izboljšamo, če vemo, kakšno prečno preverjanje bo uporabljeno. Ugotovili smo, da varčna apriorna gotovost dokaj učinkovito preprečuje preveliko prileganje podatkom: v povprečju so se rezultati z dovoljevanjem večje velikosti interakcij izboljšali. Kikuči-Bayes premaga tudi postopke indukcije odločitvenih dreves (C4.5) in se uspešno kosa tudi s postopki podpornih vektorjev.

				logaritmična napaka na primer					klasifikacijska napaka				
domain	t_K	n	df	NB	TAN	LR	kMAP	kBMA	NB	TAN	LR	kMAP	kBMA
horse-colic	1.89	369	228	1.67	*·5.97*	1.81	√0.83	**0.83**	**25.7**	*·67.3*	√35.0	√30.1	√30.6
hepatitis	0.47	155	48	0.78	*·1.31*	√0.77	√0.48	**0.43**	√15.6	√17.5	*·19.1*	**14.5**	√15.0
ionosphere	3.71	351	129	√0.64	*·0.74*	0.69	√0.39	**0.33**	**7.4**	√8.2	*·13.6*	√9.6	√9.6
vehicle	0.42	846	205	*·1.78*	1.14	0.93	√0.69	**0.66**	*·39.6*	**29.7**	√33.4	√31.4	√31.3
voting	0.23	435	48	*·0.60*	0.53	0.37	√0.21	**0.15**	*·9.3*	√7.9	√6.7	√4.6	**4.6**
monk2	0.01	601	17	0.65	0.63	*·0.65*	√0.45	**0.45**	38.2	36.2	*·39.6*	√27.6	**26.8**
p-tumor*	0.39	339	552	√3.17	*·4.76*	√2.76	2.65	**2.61**	**54.7**	√61.5	√63.6	*·71.3*	*·71.3*
heart	0.15	920	167	1.25	*·1.53*	1.24	√1.11	**1.10**	**42.8**	√44.1	*·46.2*	√44.8	√44.8
post-op	0.01	88	19	√0.93	*·1.78*	√0.81	√0.79	**0.67**	√33.4	√32.7	*·34.5*	**28.4**	√28.6
wdbc	0.57	569	61	0.26	0.29	*·0.42*	√0.15	**0.13**	√4.2	√4.4	*·7.8*	**4.0**	√4.1
promoters*	37.5	106	227	√0.60	*·3.14*	√0.70	√0.59	**0.54**	√13.4	30.4	*·57.4*	**10.4**	√10.6
lymph	0.39	148	94	√1.10	*·1.25*	√0.91	√0.98	**0.86**	√20.1	**16.1**	√23.1	*·26.5*	√25.7
cmc	0.04	1473	55	1.00	*·1.03*	0.97	√0.93	**0.92**	47.8	√45.8	*·49.7*	√43.6	**43.4**
adult	1.11	32561	134	*·0.42*	0.33	0.35	0.30	**0.30**	*·16.4*	14.3	**13.6**	√13.9	√13.9
crx	0.19	690	58	√0.49	*·0.93*	√0.39	√0.37	**0.36**	√14.1	*·17.1*	√14.1	**13.3**	√13.8
krkp	6.52	3196	69	*·0.29*	0.19	0.08	√0.06	**0.05**	*·12.4*	7.8	√2.5	**1.6**	√1.7
glass	0.03	214	90	√1.25	*·1.76*	√1.07	1.12	**1.05**	**28.3**	√29.2	√32.0	*·32.1*	√31.4
australian	0.16	690	49	√0.46	*·0.94*	√0.39	√0.41	**0.38**	√14.3	*·17.6*	√15.4	√14.3	**14.3**
titanic	0.01	2201	8	*·0.52*	√0.48	0.50	√0.48	**0.48**	*·22.3*	**21.1**	√22.2	**21.1**	√21.1
segment	0.74	2310	617	0.38	*·1.06*	0.45	**0.17**	**0.17**	√6.5	*·14.2*	√7.7	**5.4**	**5.4**
lenses	0.00	24	14	√2.44	*·2.99*	√0.89	**0.34**	0.39	√28.3	*·35.8*	√26.7	**12.5**	√15.0
monk1	0.01	556	16	0.50	0.09	*·0.50*	**0.01**	√0.02	25.4	**0.0**	*·25.5*	**0.0**	**0.0**
breast-LJ	0.03	286	24	√0.62	*·0.89*	**0.58**	√0.67	√0.58	**27.8**	√28.4	√28.3	*·29.0*	√28.7
monk3	0.01	554	17	*·0.20*	√0.11	**0.10**	√0.11	√0.11	*·3.6*	√1.6	√1.7	√1.1	**1.1**
bupa	0.01	345	12	*·0.62*	√0.60	**0.60**	√0.62	√0.61	√33.9	√32.8	*·34.5*	√33.2	**32.8**
tic-tac-toe	0.03	958	27	*·0.55*	0.49	**0.06**	√0.08	√0.07	*·29.8*	23.8	**2.0**	√3.1	√2.9
pima	0.02	768	19	√0.50	√0.49	**0.46**	*·0.51*	√0.48	√22.1	√22.1	**21.8**	*·22.4*	√22.0
iris	0.00	150	15	√0.27	*·0.32*	**0.21**	√0.27	√0.23	*·6.3*	√6.0	√5.6	**5.2**	**5.2**
spam	39.9	4601	156	*·0.53*	0.32	**0.16**	0.19	√0.19	*·9.7*	√6.9	**5.9**	√6.2	√6.2
breast-wisc	0.03	683	28	√0.21	*·0.23*	**0.13**	√0.21	√0.18	**2.6**	√3.4	√3.9	√3.9	*·4.0*
german	0.64	1000	68	√0.54	*·1.04*	**0.52**	0.65	√0.59	√24.5	*·27.3*	**24.4**	√26.3	√26.3
anneal	6.16	898	204	√0.07	*·0.17*	**0.02**	0.11	0.11	√1.3	*·2.9*	**0.3**	2.4	2.5
ecoli	0.01	336	92	√0.89	*·0.94*	**0.68**	√0.85	√0.83	**15.3**	√15.4	*·16.8*	√16.4	√16.2
hayes-roth	0.00	160	24	0.46	*·1.18*	**0.26**	0.45	0.45	√14.9	*·29.9*	√17.0	**13.5**	**13.5**
balance-scale	0.00	625	40	0.51	*·1.13*	**0.28**	0.51	0.51	√9.3	*·15.0*	**8.5**	√9.3	√9.3
soy-large*	5.95	683	822	√0.57	√0.47	**0.37**	*·0.68*	0.68	√9.0	√8.4	**7.7**	*·27.0*	27.0
o-ring	0.00	23	7	√0.83	√0.76	**0.66**	*·1.41*	√1.00	**13.0**	*·22.6*	√17.4	√22.6	√19.1
lung-cancer*	35.0	32	233	5.41	*·6.92*	**1.24**	√2.37	√1.62	**51.9**	√63.8	*·70.6*	√60.6	√61.9
audiology*	81.2	226	1783	3.55	*·5.56*	**1.40**	2.24	2.23	√40.8	62.7	**26.0**	*·68.6*	*·68.6*
soy-small*	5.29	47	115	√0.00	**0.00**	*·0.15*	0.00	0.00	**0.0**	**0.0**	*·2.1*	**0.0**	**0.0**
mushroom	1.33	8124	72	*·0.01*	**0.00**	0.00	0.00	0.00	*·0.4*	**0.0**	**0.0**	√0.0	√0.0
shuttle	0.01	253	15	*·0.16*	**0.06**	√0.10	√0.07	√0.07	*·6.7*	√2.8	**2.5**	√3.6	√2.9
car	0.02	1728	48	0.32	**0.18**	*·0.33*	0.19	0.19	14.6	**5.9**	*·16.7*	√6.5	√6.5
zoo*	0.23	101	124	**0.38**	*·0.46*	√0.38	√0.40	√0.40	**3.6**	√6.3	√7.5	*·12.9*	√12.1
wine	0.10	178	50	**0.06**	*·0.29*	√0.09	√0.19	√0.14	**0.9**	√3.1	√2.2	*·4.3*	√3.6
yeast-class*	138	186	376	**0.01**	√0.03	*·0.90*	0.25	0.23	**0.1**	√0.3	*·34.9*	√2.9	√2.9
pov.mesto				3.68	·3.99	√2.54	2.84	**1.95**	2.98	3.20	·3.34	√2.87	**2.62**

Tabela A.2: Primerjava Kikuči-Bayesovega klasifikatorja z najverjetnejšo aposteriorno strukturo (kMAP) ter z Bayesovskim povprečenjem modelov (kBMA), logistično regresijo (LR), naivnim Bayesovim klasifikatorjem (NB) in drevesnim Bayesovim klasifikatorjem (TAN, (Friedman et al., 1997)). Najboljši rezultat je v mastnem tisku, rezultati metod, ki so v vsaj dveh od 25 eksperimentov na domeni bili boljši od v povprečju najboljšega pa so označeni z $\surd$. t_K označuje čas učenja Kikuči-Bayesovega klasifikatorja v sekundah za strukture s stopnjo interakcije do vključno 4. Razvidno je predvsem to, da sta NB in TAN potisnjena na rob: zelo težko konkurirata postopkoma Kikuči-Bayesa s sposobnostjo obravnave kompleksnih struktur, ter logistične regresije s sposobnostjo pravilnega obravnavanja negativnih interakcij. Kikuči-Bayes z istim modelom zmaga po obeh kriterijih, za druge klasifikatorje to ne velja: NB je slab glede na logaritmično napako in soliden pri klasifikacijski točnosti, LR pa je solidna glede na logaritmično napako in najslabša pri klasifikacijski točnosti. Nasploh pa je uspešnosti pri logaritmčni napaki slabo korelirana z uspešnostjo pri klasifikacijski točnosti.

A.6 Zaključek in prispevki

Disertacija se je ukvarjala s splošnim problemom interakcij v strojnem učenju. Koncept interakcije smo definirali v splošnem kot potrebo po hkratnem poznavanju vrednosti več atributov. Predlagali smo nove preizkuse značilnosti interakcije. Predstavili smo načine razumevanja podatkov s pomočjo interakcij. Pokazali smo, kako lahko interakcije koristijo obstoječim postopkom strojnega učenja in kako lahko zgradimo učinkovite postopke strojnega učenja na podlagi interakcij. V splošnem se naučeni modeli in vizualizacije zelo dobro ujemajo s človeško intuicijo o podatkih.

Utrdili smo pogled na strojno učenje v kontekstu funkcije koristnosti, algoritma, prostora hipotez in podatkov. Ta kontekst velja tako za hevristike kot tudi za postopek učenja. V nadaljevanju se odpirajo nove priložnosti pri povezovanju postopkov strojnega učenja: boljše kombiniranje interakcij, boljše iskanje interakcij in boljša obravnava posamične interakcije. V preteklosti je vsak algoritem počel vse hkrati.

Prispevki k znanosti

- Posplošili smo koncepte povezanosti in korelacije v en sam pojem interakcije na k atributih, ki ga lahko prilagodimo specifičnemu modelu ter specifični funkciji koristnosti.
- Predstavili smo množico novih postopkov vizualizacije, ki dokazujejo, da se pojem interakcije zelo dobro ujema s človeško intuicijo in pogosto nudi nepričakovana a razumljiva spoznanja.
- Predstavili smo več novih preskusov statistične značilnosti, ki omogočajo, da se statistično ovrednoti smiselnost zaključka o interakciji na podlagi omejene količine podatkov.
- Predstavili smo nov in učinkovit postopek strojnega učenja, Kikuči-Bayes, katerega model se sestoji iz množice potencialno prekrivajočih se interakcij. Postopek se uspešno kosa z najboljšimi na področju strojnega učenja, posplošuje Bayesovske mreže in omogoča nedisjunktno dekompozicijo atributov.
- Uporaba varčnih apriornih gotovosti, ki avtomatsko poskrbijo za kaznovanje v skladu s povečanjem kompleksnosti danega modela. Ugotovitev, da lahko algoritem učenja prilagodimo lastnostim prečnega preverjanja.
- Predstavili smo novo hevristiko za izbiro atributov, ki temelji na interakcijskem prispevku.
- V splošnem smo problem strojnega učenja razbili na problem iskanja strukture interakcij ter na problem obravnave posamične interakcije.
- V temeljitem pregledu literature smo odkrili več neodvisnih odkritij in poimenovanj interakcijskega prispevka, kar nakazuje vsestranskost tega pojma.
- Pokazali smo obvladljivost problema iskanja interakcij na domenah z zelo veliko atributi ter nelinearne obravnave interakcij med zveznimi atributi.

- V slovenskem prostoru smo predstavili nekatere pojme moderne Bayesove statistike in med prvimi uporabili elemente statistične mehanike (približka Kikučija in Kirkwooda) v strojnem učenju.

Nadaljnje delo

- Kikuči-Bayesova obravnava interakcij izključno s kartezičnim produktom je dokaj primitivna. Eksperimenti dokazujejo uporabnost pravil, skupin in taksonomij za opisovanje interakcije. Ti elementi pa še niso integrirani kot del učenja. Te rešitve ponujajo novo perspektivo problema konstruktivne indukcije v strojnem učenju.
- Navkljub pomenu strukture se iz rezultatov vidi pomembnost numeričnih postopkov, ki učinkovito obravnavajo veliko število manj kompleksnih negativnih interakcij: metoda podpornih vektorjev z linearnim jedrom deluje boljše od Kikuči-Bayesovega klasifikatorja. V zadnjem času se je postavilo nekaj izhodišč, kjer lahko na podlagi strukture interakcij definiramo jedro, s katerim potem dobimo napovedni model. Obetavno se zdi uporabiti algoritme Kikuči-Bayesa za avtomatsko izgradnjo jeder.
- Obstajajo tudi drugi približki za združevanje interakcij, zelo malo pa je znanega o njihovi kvaliteti. Vsekakor je treba upoštevati, da je Kikučijev približek mišljen za združevanje podmodelov v skupni (joint) model in ne za združevanje v napovedni (class-predictive) model. Tu bi se s primernim pristopom lahko še kaj pridobilo.
- Čeprav smo opravili začetne preizkuse obravnave interakcij pri domenah z zelo veliko atributi, je težko ovrednotiti kvaliteto obstoječega postopka.
- Naše vizualizacije ne upoštevajo negotovosti pri stopnji interakcije in je ne prikažejo učinkovito.
- Zanimivo bi bilo omogočiti človeku interaktiven vpogled v gradnjo modela in mu omogočiti aktivno vlogo pri učenju. To omogoča in poenostavlja intuitivna razumljivost interakcij.

Izjava

Izjavljam, da sem doktorsko disertacijo izdelal samostojno pod mentorstvom akad. prof. dr. Ivana Bratka. Ostale sodelavce, ki so mi pri nalogi pomagali, sem navedel v razdelku Acknowledgments na strani vii.

Sežana, junij 2005 Aleks Jakulin

BIBLIOGRAPHY

A. Abu-Hanna and N. de Keizer. Integrating classification trees with local logistic regression in Intensive Care prognosis. *Artificial Intelligence in Medicine*, 29(1-2):5–23, Sep-Oct 2003.

A. Agresti. *Categorical data analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2nd edition, 2002.

Y. Altun, A. Smola, and T. Hofmann. Exponential families for conditional random fields. In M. Chickering and J. Halpern, editors, *Proc. of 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 2–9, Banff, Alberta, Canada, July 2004.

S.-i. Amari. Information geometry on hierarchical decomposition of stochastic interactions. *IEEE Trans. on Information Theory*, 47(5):1701–1711, July 2001.

S.-i. Amari and H. Nagaoka. *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, Rhode Island, 2000.

J. H. Badsberg and F. M. Malvestuto. An implementation of the iterative proportional fitting procedure by propagation trees. *Computational Statistics & Data Analysis*, 37: 297–322, 2001.

R. M. Baron and D. A. Kenny. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology*, 51:1173–1182, 1986.

M. S. Bartlett. Contingency table interactions. *Journal of the Royal Statistical Society*, Suppl. 2:248–252, 1935.

J. Baxter. The canonical distortion measure for vector quantization and approximation. In *ICML 1997*, pages 39–47. Morgan Kaufmann, 1997.

S. D. Bay. Multivariate discretization for set mining. *Knowledge and Information Systems*, 3(4):491–512, 2001.

A. J. Bell. The co-information lattice. In *ICA 2003*, Nara, Japan, April 2003.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57:289–300, 1995.

A. Berger, S. Della Pietra, and V. Della Pietra. A maximum entropy approach to natural language processing. *Computational Linguistics*, 22(1):39–71, March 1996.

J. Berger. Could Fisher, Jeffreys and Neyman have agreed upon testing? *Statistical Science*, 18:1–32, 2003.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory*. Wiley, Chichester, 2000.

D. R. Billinger. Some data analyses using mutual information. *Brazilian J. Probability and Statistics*, (0), 2004. to appear.

C. M. Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Computation*, 7(1):108–116, 1995.

J. L. Boes and C. R. Meyer. Multi-variate mutual information for registration. In C. Taylor and A. Colchester, editors, *MICCAI 1999*, volume 1679 of *LNCS*, pages 606–612. Springer-Verlag, 1999.

I. Borg and P. Groenen. *Modern Multidimensional Scaling*. Springer-Verlag, New York, 1997.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

L. Breiman. Random forests – random features. Technical Report 567, University of California, Statistics Department, Berkeley, 1999.

N. Brenner, S. P. Strong, R. Koberle, W. Bialek, and R. R. de Ruyter van Steveninck. Synergy in a neural code. *Neural Computation*, 12:1531–1552, 2000.

C. A. Brewer, G. W. Hatchard, and M. A. Harrower. ColorBrewer in print: A catalog of color schemes for maps. *Cartography and Geographic Information Science*, 30(1):5–32, 2003.

G. W. Brier. Verification of forecasts expressed in terms of probability. *Weather Rev*, 78: 1–3, 1950.

W. Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *ECML 2002*, volume 2430 of *LNCS*. Springer-Verlag, August 2002.

W. Buntine. Classifiers: A theoretical and empirical study. In *Int. Joint Conf. on AI*, Sydney, Australia, 1991.

W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In M. Chickering and J. Halpern, editors, *Proc. of 20th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 59–66, Banff, Alberta, Canada, 2004. `http://www.hiit.fi/u/buntine/uai2004.html`.

W. L. Buntine. A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8:195–210, 1996.

K. P. Burnham and D. R. Anderson. *Model Selection and Multimodel Inference*. Springer, 2nd edition, 2002.

R. Caruana, A. Niculescu, G. Crew, and A. Ksikes. Ensemble selection from libraries of models. In *Proc. of 21st International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004.

A. Caticha. Maximum entropy, fluctuations and priors. In A. Mohammad-Djafari, editor, *MaxEnt 2000, the 20th International Workshop on Bayesian Inference and Maximum Entropy Methods*, Maximum Entropy and Bayesian Methods in Science and Engineering (A.I.P. Vol. 568, 2001), page 94, Gif-sur-Yvette, France, 2000.

N. J. Cerf and C. Adami. Entropic Bell inequalities. *Physical Review A*, 55(5):3371–3374, May 1997.

J. Cerquides and R. López de Màntaras. Tractable Bayesian learning of tree augmented naive Bayes classifiers. In *Proc. of the 20th International Conference on Machine Learning*, pages 75–82, 2003.

B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *ECAI 1990*, pages 147–149, 1990.

C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2005. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

G. Chechik, A. Globerson, M. J. Anderson, E. D. Young, I. Nelken, and N. Tishby. Group redundancy measures reveal redundancy reduction in the auditory pathway. In T. G. Dietterich, S. Becker, and Z. Ghahramani, editors, *NIPS 2002*, pages 173–180, Cambridge, MA, 2002. MIT Press.

P. Cheeseman and J. Stutz. On the relationship between Bayesian and maximum entropy inference. In *24th International Workshop on Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, volume 735 of *AIP Conference Proceedings*, pages 445–461, Garching (Germany), July 2004.

J. Cheng, R. Greiner, J. Kelly, D. A. Bell, and W. Liu. Learning Bayesian networks from data: an information-theory based approach. *Artificial Intelligence Journal*, 137:43–90, 2002.

U. Cherubini, E. Luciano, and W. Vecchiato. *Copula methods in finance*. John Wiley & Sons, Chichester, England, 2004.

C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, 14(3):462–467, May 1968.

J. D. Clinton, S. Jackman, and D. Rivers. The statistical analysis of roll call voting: A unified approach. *American Political Science Review*, 98(2):355–370, 2004. `http://www.princeton.edu/~clinton/CJR_APSR2004.pdf`.

T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley Series in Telecommunications. Wiley, New York, 1991.

K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, December 2001.

I. Csiszár. Why least squares and maximum entropy? an axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

I. Cziszár. Information theoretic methods in probability and statistics. *IEEE Information Theory Society Newsletter*, 48:21–30, 1998.

J. N. Darroch. Multiplicative and additive interaction in contingency tables. *Biometrika*, 61(2):207–214, 1974.

J. N. Darroch and D. Ratcliff. Generalised iterative scaling and maximum likelihood. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

J. N. Darroch, S. L. Lauritzen, and T. P. Speed. Markov fields and log-linear interaction models for contingency tables. *The Annals of Statistics*, 8:522–539, 1980.

O. A. Davis, M. J. Hinich, and P. C. Ordeshook. An expository development of a mathematical model of the electoral process. *American Political Science Review*, 64:426–448, 1970.

A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application.* Cambridge University Press, 1997.

A. P. Dawid. Statistical theory: the prequential approach (with discussion). *J. R. Statist. Soc. A*, 147:278–292, 1984.

J. de Leeuw. Principal component analysis of binary data: Applications to roll-call-analysis. Technical Report 364, UCLA Department of Statistics, 2003. `http://preprints.stat.ucla.edu/download.php?paper=364`.

J. de Leeuw. Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, and B. van Cutsem, editors, *Recent developments in statistics*, pages 133–145, Amsterdam, The Netherlands, 1977. North Holland Publishing Company.

M. H. DeGroot and M. J. Schervish. *Probability and Statistics.* Addison-Wesley, third edition, 2002.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(4):380–393, April 1997.

J. Demšar. *Constructive Induction by Attribute Space Reduction.* PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 2002.

J. Demšar and B. Zupan. Orange: From experimental machine learning to interactive data mining, 2004. White Paper (`http://www.ailab.si/orange`) Faculty of Computer and Information Science, University of Ljubljana, Slovenia.

P. Diaconis, S. Holmes, and R. Montgomery. Dynamical bias in the coin toss. 2004.

T. G. Dietterich and E. B. Kong. Machine learning bias, statistical bias, and statistical variance of decision tree algorithms. Technical report, Oregon State University, Department of Computer Science, 1995.

P. Domingos and M. Pazzani. On the optimality of the simple Bayesian classifier under zero-one loss. *Machine Learning*, 29:103–130, 1997.

B. Efron and G. Gong. A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37:36–48, February 1983.

N. Eriksson, S. E. Fienberg, A. Rinaldo, and S. Sullivant. Polyhedral conditions for the nonexistence of the MLE for hierarchical log-linear models, May 2004. arxiv:math.CO/0405044.

S. Esmeir and S. Markovitch. Lookahead-based algorithms for anytime induction of decision trees. In R. Greiner and D. Schuurmans, editors, *Proc. of 21st International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004.

R.-E. Fan, P.-H. Chen, and C.-J. Lin. Working set selection using the second order information for training SVM. Technical report, Department of Computer Science, National Taiwan University, 2005.

R. M. Fano. *The Transmission of Information: A Statistical Theory of Communication*. MIT Press, Cambridge, Massachussets, March 1961.

U. M. Fayyad and K. B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI 1993*, pages 1022–1027. AAAI Press, 1993.

D. H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2:139–172, 1987.

R. A. Fisher. On an absolute criterion for fitting frequency curves. *Messenger of Mathematics*, 41:155–160, 1912.

R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85:87–94, 1922.

R. A. Fisher. Inverse probability. *Proceedings of the Cambridge Philosophical Society*, 26: 528–535, 1930.

R. A. Fisher. The fiducial argument in statistical inference. *Annals of Eugenics*, 6:391–398, 1935.

F. Fleuret. Fast binary feature selection with conditional mutual information. *Journal of Machine Learning Research*, 5:1531–1555, November 2004.

M. R. Forster and E. Sober. How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, 45:1–35, 1994.

E. Frank and S. Kramer. Ensembles of nested dichotomies for multi-class problems. In R. Greiner and D. Schuurmans, editors, *Proc. of 21st International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004. ACM Press.

E. Frank and I. H. Witten. Using a permutation test for attribute selection in decision trees. In J. Shavlik, editor, *Proc. of the Fifteenth International Conference on Machine Learning*, pages 152–160, Madison, Wisconsin, 1998. Morgan Kaufmann Publishers, San Francisco, CA.

Eibe Frank, July 2004. Personal communication.

J. J. Freeman. Note on approximating discrete probability distributions. *IEEE Transactions on Information Theory*, 17(4):491–493, July 1971.

A. A. Freitas. Understanding the crucial role of attribute interaction in data mining. *Artificial Intelligence Review*, 16(3):177–199, November 2001.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

Y. Freund, Y. Mansour, and R. E. Schapire. Generalization bounds for averaged classifiers (how to be a Bayesian without believing). *The Annals of Statistics*, 32(4):1698–1722, August 2004.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.

I. Gat. *Statistical Analysis and Modeling of Brain Cells' Activity*. PhD thesis, Hebrew University, Jerusalem, Israel, 1999.

G. Gediga and I. Düntsch. On model evaluation, indices of importance, and interaction values in rough set analysis. In S. K. Pal, L. Polkowski, and A. Skowron, editors, *Rough-Neuro Computing: Techniques for computing with words*. Physica Verlag, Heidelberg, 2003.

D. Geiger and D. Heckerman. Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82:45–74, April 1996.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL, 2nd edition, 2004a.

A. Gelman, J. N. Katz, and J. Bafumi. Standard voting power indexes don't work: an empirical analysis. *British Journal of Political Science*, 34(4):657–674, 2004b.

S. Geman, E. Bienenstock, and R. Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4:1–58, 1992.

A. Globerson and N. Tishby. The minimum information principle in discriminative learning. In *Uncertainty in Artificial Inteligence (UAI)*, pages 193–200, Banff, Canada, 2004.

A. Goldenberg and A. Moore. Tractable learning of large Bayes net structures from sparse data. In R. Greiner and D. Schuurmans, editors, *Proc. of 21st International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004. ACM Press.

G. H. Golub, P. C. Hansen, and D. P. O'Leary. Tikhonov regularization and total least squares. *SIAM J. Matrix Anal. Appl.*, 21(1):185–194, 1999.

I. J. Good. Maximum entropy for hypothesis formulation. *The Annals of Mathematical Statistics*, 34:911–934, 1963.

I. J. Good. *The Estimation of Probabilities: An Essay on Modern Bayesian Methods*, volume 30 of *Research Monograph*. M.I.T. Press, Cambridge, Massachusetts, 1965.

I. J. Good. *Good Thinking: The Foundations of Probability and Its Applications*. University of Minnesota Press, Minneapolis, 1983.

M. Grabisch and M. Roubens. An axiomatic approach to the concept of interaction among players in cooperative games. *International Journal of Game Theory*, 28(4):547–565, 1999.

D. Grossman and P. Domingos. Learning Bayesian network classifiers by maximizing conditional likelihood. In *Proc. 21st International Conference on Machine Learning*, pages 361–368, Banff, Canada, 2004. ACM Press.

P. Grünwald. *The Minimum Description Length Principle and Reasoning Under Uncertainty*. PhD dissertation, Universiteit van Amsterdam, Institute for Logic, Language, and Computation, 1998.

P. D. Grünwald and A. P. Dawid. Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. *Annals of Statistics*, 32(4), August 2004.

M. A. Hall. Correlation-based feature selection for discrete and numeric class machine learning. In *ICML 2000*, Stanford University, CA, 2000. Morgan Kaufmann.

T. S. Han. Linear dependence structure of the entropy space. *Information and Control*, 29:337–368, 1975.

T. S. Han. Lattice-theoretic formulation of multifactor interactions and minimality of cell-symmetric decomposition of χ^2-statistics. *Rep. Stat. Appl. Res. JUSE*, 24(2):55–66, June 1977.

T. S. Han. Nonnegative entropy measures of multivariate symmetric correlations. *Information and Control*, 36:133–156, 1978.

T. S. Han. Multiple mutual informations and multiple interactions in frequency data. *Information and Control*, 46(1):26–45, July 1980.

P. Harremoës and F. Topsøe. Maximum entropy fundamentals. *Entropy*, 3:191–226, 2001.

W. H. Harris. Traumatic arthritis of the hip after dislocation and acetabular fractures: Treatment by mold arthroplasty: end result study using a new method of result evaluation. *J Bone Joint Surg*, 51-A:737–55, 1969.

T. Hastie and R. Tibshirani. Discriminant adaptive nearest neighbor classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, 18(6):607–616, 1996. ISSN 0162-8828. doi: http://dx.doi.org/10.1109/34.506411.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: data mining, inference, and prediction.* Springer, 2001.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–1415, October 2004.

D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

Hawaii State Department of Health. Vital statistics report, 2002. `http://www.hawaii.gov/health/statistics/vital-statistics/vr_02/`.

S. Hettich and S. D. Bay. The UCI KDD archive. Irvine, CA: University of California, Department of Information and Computer Science, 1999. URL `http://kdd.ics.uci.edu`.

J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky. Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–417, 1999.

L. Hunt and M. Jorgensen. Mixture model clustering using the MULTIMIX program. *Australian & New Zealand Journal of Statistics*, 41(2):153–171, June 1999.

M. Hutter and M. Zaffalon. Distribution of mutual information from complete and incomplete data. *Computational Statistics & Data Analysis*, 48(3):633–657, 2005. URL `http://www.hutter1.de/ai/mifs.htm`.

A. Jakulin. Modelling modelled. *S.E.E.D. Journal*, 4(3):58–77, 2004.

A. Jakulin. Symmetry and information theory. *Symmetry: Culture and Science*, 2005. in press.

A. Jakulin. Attribute interactions in machine learning. Master's thesis, University of Ljubljana, Faculty of Computer and Information Science, January 2003.

A. Jakulin. Extensions to the Orange data mining framework, January 2002. `http://-www.ailab.si/aleks/orng/`.

A. Jakulin and I. Bratko. Analyzing attribute dependencies. In N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, editors, *Proc. of Principles of Knowledge Discovery in Data (PKDD)*, volume 2838 of *LNAI*, pages 229–240. Springer-Verlag, September 2003.

A. Jakulin and I. Bratko. Quantifying and visualizing attribute interactions: An approach based on entropy. `http://arxiv.org/abs/cs.AI/0308002` v3, March 2004a.

A. Jakulin and I. Bratko. Testing the significance of attribute interactions. In *Proc. of 21st International Conference on Machine Learning (ICML)*, pages 409–416, Banff, Alberta, Canada, 2004b.

A. Jakulin and G. Leban. Interaktivna interakcijska analiza. In M. Bohanec, B. Filipič, M. Gams, D. Trček, and B. Likar, editors, *Proceedings A of the 6th International Multi-Conference Information Society IS 2003*, pages 57–60, Ljubljana, Slovenia, October 2003. Jožef Stefan Institute.

A. Jakulin, I. Bratko, D. Smrke, J. Demšar, and B. Zupan. Attribute interactions in medical data analysis. In M. Dojat, E. Keravnou, and P. Barahona, editors, *Proc. of Artificial Intelligence in Medicine Europe (AIME)*, volume 2780 of *LNAI*, pages 229–238. Springer-Verlag, October 2003.

E. T. Jaynes. In G. L. Bretthorst, editor, *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK, 2003.

E. T. Jaynes. Information theory and statistical mechanics. *The Physical Review*, 106(4): 620–630, May 1957.

T. Jebara. *Machine Learning: Discriminative and Generative*. Kluwer International Series in Engineering and Computer Science. Kluwer, 2003.

D. Jenul. Uporaba metod strojnega učenja za izdelavo napovednih modelov bolnikovega dolgoročnega kliničnega statusa po vgraditvi kolčne endoproteze : diplomsko delo, 2003. Fakulteta za računalnistvo in informatiko, Univerza v Ljubljani.

T. Joachims. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*. MIT Press, 1999.

J. D. Jobson. *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*. Springer-Verlag, New York, July 1992.

Harry Joe. *Multivariate models and dependence concepts*. Chapman & Hall, London, 1997.

C. M. Kadie. *Seer: Maximum Likelihood Regression for Learning-Speed Curves*. PhD thesis, University of Illinois at Urbana-Champaign, 1995.

D.-K. Kang, A. Silvescu, J. Zhang, and V. Honavar. Generation of attribute value taxonomies from data for data-driven construction of accurate and compact classifiers. In *Proceedings of the IEEE International Conference on Data Mining*, 2004.

J. N. Kapur. *Maximum Entropy Models in Science and Engineering*. John Wiley & Sons, 1990.

S. Kaski and J. Sinkkonen. Metrics that learn relevance. In *Proceedings of the International Joint Conference on Neural Networks (IJCNN-2000)*, volume 5, pages 547–552, Piscataway, NJ, 2000. IEEE.

L. Kaufman and P. J. Rousseeuw. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York, 1990.

E. Keogh and M. Pazzani. Learning the structure of augmented Bayesian classifiers. *International Journal on Artificial Intelligence Tools*, 11(4):587–601, 2002.

R. Kikuchi. A theory of cooperative phenomena. *Physical Review*, 81(6):988–1003, 1951.

K. Kira and L. A. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *ICML 1992*, pages 249–256. Morgan Kaufmann, 1992.

W. Kirchherr, M. Li, and P. M. B. Vitányi. The miraculous universal distribution. *Mathematical Intelligencer*, 19(4):7–15, 1997.

J. G. Kirkwood and E. M. Boggs. The radial distribution function in liquids. *Journal of Chemical Physics*, 10(6):394–402, 1942.

R. Kohavi. Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *KDD 1996*, pages 202–207, 1996.

R. Kohavi. *Wrappers for Performance Enhancement and Oblivious Decision Graphs*. PhD dissertation, Stanford University, September 1995.

I. Kojadinovic. Modeling interaction phenomena using fuzzy measures: on the notions of interaction and independence. *Fuzzy Sets and Systems*, 135(3):317–340, May 2003.

D. Koller and M. Sahami. Toward optimal feature selection. In L. Saitta, editor, *ICML 1996*, Bari, Italy, 1996. Morgan Kaufmann.

I. Kononenko. Semi-naive Bayesian classifier. In Y. Kodratoff, editor, *EWSL 1991*, volume 482 of *LNAI*. Springer Verlag, 1991.

I. Kononenko. *Strojno učenje*. Fakulteta za računalništvo in informatiko, Ljubljana, Slovenija, 1997.

I. Kononenko. *Bayesovske nevronske mreže*. PhD thesis, Univerza v Ljubljani, 1990.

I. Kononenko and I. Bratko. Information based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80, 1991.

I. Kononenko, B. Cestnik, and I. Bratko. *Assistant Professional User's Guide*. Jožef Stefan Institute, Ljubljana, Slovenia, 1988.

I. Kononenko, E. Šimec, and Marko Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with RELIEFF. *Applied Intelligence*, 7(1):39–55, 1997.

E. Koutsofios and S. C. North. *Drawing Graphs with dot*, 1996. URL `research.att.com/dist/drawdag/dotguide.ps.Z`.

K. Krippendorff. *Information Theory: Structural Models for Qualitative Data*, volume 07–062 of *Quantitative Applications in the Social Sciences*. Sage Publications, Inc., Beverly Hills, CA, 1986.

F. R. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Information Theory*, 47(2):498–519, 2001.

M. Kukar. Drifting concepts as hidden factors in clinical studies. In M. Dojat, E. T. Keravnou, and P. Barahona, editors, *AIME*, volume 2780 of *Lecture Notes in Computer Science*, pages 355–364. Springer, 2003.

S Kullback. Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236–1243, 1968.

S. Kullback and R. Leibler. On information and sufficiency. *Ann. Math. Stat.*, 22:76–86, 1951.

J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. of the International Conference on Machine Learning (ICML)*, pages 282–289, 2001.

J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: representation and clique selection. In R. Greiner and D. Schuurmans, editors, *Proc. of 21st International Conference on Machine Learning (ICML)*, Banff, Alberta, Canada, 2004.

H. O. Lancaster. *The Chi-Squared Distribution*. J. Wiley & Sons, New York, 1969.

G. R. G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

N. Landwehr, M. Hall, and E. Frank. Logistic model trees. In N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, editors, *ECML 2003*, volume 2837 of *LNAI*, pages 241–252. Springer-Verlag, September 2003.

P. Langley and S. Sage. Induction of selective Bayesian classifiers. In *Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406, Seattle, WA, 1994. Morgan Kaufmann.

B. L. Lawson, M. E. Orrison, and D. T. Uminsky. Noncommutative harmonic analysis of voting in committees, April 2003. `http://homepages.uc.edu/~lawsonb/research/noncommutative.pdf`.

H. K. H. Lee and M. A. Clyde. Lossless online Bayesian bagging. *Journal of Machine Learning Research*, 5:143–151, 2004.

L. Leydesdorff and M. Meyer. The triple helix of university-industry-government relations. *Scientometrics*, 58(2):191–203(13), October 2003.

M. Li, X. Chen, X. Li, B. Ma, and P. M. B. Vitányi. The similarity metric. *IEEE Transactions on Information Theory*, 50(12):3250–3264, December 2004.

C. X. Ling and H. Zhang. The representational power of discrete Bayesian networks. *Journal of Machine Learning Research*, 3:709–721, 2002.

R. López de Màntaras. A distance based attribute selection measure for decision tree induction. *Machine Learning*, 6(1):81–92, 1991.

H. S. Lynn. Suppression and confounding in action. *The American Statistician*, 57(1): 58–61, February 2003.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, October 2003.

R. Malouf. A comparison of algorithms for maximum entropy parameter estimation. In D. Roth and A. van den Bosch, editors, *Proceedings of the 6th Conference on Natural Language Learning (CoNLL-2002)*, pages 49–55, Taipei, Taiwan, 2002.

C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, May 1999.

D. J. Margolis, A. C. Halpern, T. Rebbeck, L. Schuchter, R. L. Barnhill, J. Fine, and M. Berwick. Validation of a melanoma prognostic model. *Arch Dermatol.*, 134:1597–1601, 1998.

H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096–3102, September 2000.

F. Matúš. Conditional independences among four random variables III: final conclusion. *Combinatorics, Probability & Computing*, 8:269–276, 1999.

A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classification. In *AAAI-98 Workshop on 'Learning for Text Categorization'*, 1998.

N. McCarty, K. T. Poole, and H. Rosenthal. The hunt for party discipline in congress. *American Political Science Review*, 95:673–687, 2001. `http://voteview.uh.edu/d011000merged.pdf`.

G. H. McClelland and C. M. Judd. Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin*, 114:376–390, 1993.

W. J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.

W. J. McGill and H. Quastler. Standardized nomenclature: An attempt. In H. Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 83–92, Glencoe, Illinois, 1955. The Free Press.

G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, New York, 2000.

M. Meilă. Comparing clusterings by the variation of information. In *COLT 2003*, 2003.

M. Meilă and T. Jaakkola. Tractable Bayesian learning of tree belief networks. Technical Report CMU-RI-TR-00-15, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2000.

M. Meilă and M. I. Jordan. Learning with mixtures of trees. *Journal of Machine Learning Research*, 1:1–48, October 2000.

R. Meo. Maximum independence and mutual information. *IEEE Trans. on Information Theory*, 48(1):318–324, January 2002.

D. Michie and A. Al Attar. Use of sequential Bayes with class probability trees. In J. Hayes, D. Michie, and E. Tyugu, editors, *Machine Intelligence 12*, pages 187–202. Oxford University Press, 1991.

G. A. Miller. Note on the bias of information estimates. In H. Quastler, editor, *Information Theory in Psychology: Problems and Methods*, pages 95–100, Glencoe, Illinois, USA, 1955. The Free Press Publishers.

T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, USA, 1997.

D. Mladenić. *Machine Learning on non-homogeneous, distributed text data*. PhD thesis, Univerza v Ljubljani, Slovenija, Faculty of Computer and Information Science, 1998.

S. Monti and G. F. Cooper. A Bayesian network classifier that combines a finite mixture model and a naïve Bayes model. In *UAI 1999*, pages 447–456, 1999.

J. N. Morgan and J. A. Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58:415–435, 1963.

A. Murphy. A new vector partition of the probability score. *J. Appl. Meteor.*, 12:595–600, 1973.

P. Myllymaki, T. Silander, H. Tirri, and P. Uronen. B-Course: A web-based tool for Bayesian and causal data analysis. *International Journal on Artificial Intelligence Tools*, 11(3):369–387, 2002.

R. M. Neal. Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 9:249–265, 2000.

R. M. Neal and G. E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 355–369. Kluwer Academic Publishers, 1998.

I. Nemenman. Information theory, multivariate dependence, and genetic network inference. Technical Report NSF-KITP-04-54, KITP, UCSB, 2004. `http://arxiv.org/abs/q-bio/0406015`.

H. S. Nguyen and S. H. Nguyen. Discretization methods for data mining. In L. Polkowski and A. Skowron, editors, *Rough Sets in Knowledge Discovery*, pages 451–482. Physica-Verlag, Heidelberg, 1998.

NIST/SEMATECH, 2002. NIST/SEMATECH e-Handbook of Statistical Methods, 2 Sep 2002. `http://www.itl.nist.gov/div898/handbook/`.

S. W. Norton. Generating better decision trees. In *IJCAI 1989*, pages 800–805, 1989.

L. Orlóci, M. Anand, and V. D. Pillar. Biodiversity analysis: issues, concepts, techniques. *Community Ecology*, 3(2):217–236, 2002.

G. Pagallo and D. Haussler. Boolean feature discovery in empirical learning. *Machine Learning*, 5(1):71–99, March 1990.

M. Paluš. Identifying and quantifying chaos by using information-theoretic functionals. In A. S. Weigend and N. A. Gerschenfeld, editors, *Time Series Prediction: Forecasting the Future and Understanding the Past*, NATO Advanced Research Workshop on Comparative Time Series Analysis, pages 387–413, Sante Fe, NM, May 1994. Addison-Wesley.

M. J. Pazzani. Searching for dependencies in Bayesian classifiers. In *Learning from Data: AI and Statistics V*. Springer-Verlag, 1996.

J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, San Francisco, CA, USA, 2000.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Francisco, CA, USA, 1988.

J. Peltonen, J. Sinkkonen, and S. Kaski. Sequential information bottleneck for finite data. In R. Greiner and D. Schuurmans, editors, *Proc. of 21st International Conference on Machine Learning (ICML)*, pages 647–654, Banff, Alberta, Canada, 2004.

J. M. Peña, J. A. Lozano, and P. Larrañaga. Learning recursive Bayesian multinets for data clustering by means of constructive induction. *Machine Learning*, 47(1):63–89, 2002.

E. Pérez. *Learning Despite Complex Attribute Interaction: An Approach Based on Relational Operators*. PhD dissertation, University of Illinois at Urbana-Champaign, 1997.

K. T. Poole. Non-parametric unfolding of binary choice data. *Political Analysis*, 8(3): 211–232, 2000. `http://voteview.uh.edu/apsa2.pdf`.

W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C*. Cambridge University Press, second edition, 1992.

H. Quastler. The measure of specificity. In H. Quastler, editor, *Information Theory in Biology*. Univ. of Illinois Press, Urbana, 1953.

J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 1993.

J. R. Quinlan. Induction of decision trees. *Machine Learning*, 1(1):81–106, 1986.

R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria, 2004. URL `http://www.R-project.org`. 3-900051-00-3.

H. Ragavan and L. A. Rendell. Lookahead feature construction for learning hard concepts. In *ICML 1993*, pages 252–259, 1993.

C. Rajski. A metric space of discrete probability distributions. *Information and Control*, 4:373–377, 1961.

I. Rish, J. Hellerstein, and T. S. Jayram. An analysis of data characteristics that affect naive Bayes performance. Technical Report RC21993, IBM, 2001.

J. Rissanen. Complexity of nonlogarithmic loss functions, 3 2001. `http://www.cs.tut.fi/~rissanen/papers/loss.ps`.

J. Rissanen. Stochastic complexity and modeling. *Annals of Statistics*, 14:1080–1100, 1986.

S. J. Roberts, R. Everson, and I. Rezek. Maximum certainty data partitioning. *Pattern Recognition*, 33(5):833–839, 1999.

M. Robnik-Šikonja. *Lastnosti in uporaba hevristične funkcije Relief v strojnem učenju*. PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 2001.

M. Robnik-Šikonja and I. Kononenko. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine Learning*, 53(1-2):23–69, 2003.

S. N. Roy and M. A. Kastenbaum. On the hypothesis of no 'interaction' in a multi-way contingency table. *The Annals of Mathematical Statistics*, 27:749–757, 1956.

D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9:130–134, 1981.

H. Rubin. The weak system of axioms for 'rational' behaviour and the non-separability of utility from prior. *Statistics and Decisions*, 5:47–58, 1987.

Y.D. Rubinstein and T. Hastie. Discriminative vs informative learning. In *SIGKDD 1997*, pages 49–53. AAAI Press, 1997.

S. N. Salthe. *Evolving Hierarchical Systems: Their Structure and Representation*. Columbia University Press, 1985.

S. N. Salthe. *Development and Evolution: Complexity and Change in Biology*. MIT Press, 1993.

E. Schneidman, S. Still, M. J. Berry II, and W. Bialek. Network information and connected correlations. `http://arxiv.org/abs/physics/0307072` v1, July 2004.

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

R. Setiono and H. Liu. Fragmentation problem and automated feature construction. In *ICTAI 1998*, pages 208–215, Taipei, Taiwan, November 1998. IEEE Computer Society.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

A. D. Shapiro. *Structured induction in expert systems*. Turing Institute Press in association with Addison-Wesley Publishing Company, 1987.

N. Slonim, N. Friedman, and N. Tishby. Unsupervised document classification using sequential information maximization. In *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2002.

D. J. Speigelhalter, N. G. Best, B. P. Carlin, and A. van der Linde. Bayesian measures of model complexity and fit (with discussion). *Journal of the Royal Statistical Society, Series B*, 64(4):583–616, 2003.

O. Sporns, G. Tononi, and G. M. Edelman. Theoretical neuroanatomy: Relating anatomical and functional connectivity in graphs and cortical connection matrices. *Cerebral Cortex*, 10(2):127–141, February 2000.

C. Stanfill and D. Waltz. Towards memory-based reasoning. *Communications of the ACM*, 29(12):1213–1228, 1986.

C. H. Stewart III. *Analyzing Congress*. W. W. Norton & Company, Inc., New York, 2001.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B*, 36:111–147, 1974.

J. D. Storey. The positive false discovery rate: A Bayesian interpretation and the q-value. *Annals of Statistics*, 31:2013–2035, 2003.

B. Streitberg. Exploring interactions in high-dimensional tables: a bootstrap alternative to log-linear models. *Annals of Statistics*, 27(1):405–413, 1999.

A. Struyf, M. Hubert, and P. J. Rousseeuw. Integrating robust clustering techniques in S-PLUS. *Computational Statistics and Data Analysis*, 26:17–37, 1997.

M. Studenỳ and J. Vejnarovà. The multiinformation function as a tool for measuring stochastic dependence. In M. I. Jordan, editor, *Learning in Graphical Models*, pages 261–297. Kluwer, 1998.

R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

M. Theus and S. R. W. Lauer. Visualizing loglinear models. *Journal of Computational and Graphical Statistics*, 8(3):396–412, 1999.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.

A. N. Tikhonov and V. Y. Arsenin. *Solution of Ill-Posed Problems*. Winston, Washington, DC, 1977.

N. Tishby, F. Pereira, and W. Bialek. The information bottleneck method. In B. Hajek and R. S. Sreenivas, editors, *Proceedings of the 37th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, Urbana: University of Illinois, 1999. arxiv.org/physics/0004057.

R. C. Tolman. *The Principles of Statistical Mechanics*. Dover, New York, 1979.

F. Topsøe. Information theory at the service of science. To appear in a special volume of the Janos Bolyai Mathematical Society, 2004.

I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector learning for interdependent and structured output spaces. In *Proc. 21st International Conference on Machine Learning*, Banff, Canada, 2004. ACM Press.

T. Tsujishita. On triple mutual information. *Advances in Applied Mathematics*, 16:269–274, 1995.

L. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, 2nd edition, 1999.

V. Vedral. The role of relative entropy in quantum information theory. *Reviews of Modern Physics*, 74, January 2002.

R. Vilalta and I. Rish. A decomposition of classes via clustering to explain and improve naive Bayes. In N. Lavrač, D. Gamberger, H. Blockeel, and L. Todorovski, editors, *ECML 2003*, volume 2837 of *LNAI*, pages 444–455. Springer-Verlag, September 2003.

R. Vilalta, G. Blix, and L. A. Rendell. Global data analysis and the fragmentation problem in decision tree induction. In *ECML 1997*, volume 1224 of *LNAI*, pages 312–326. Springer-Verlag, 1997.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical Report 649, Department of Statistics, University of California, Berkeley, September 2003.

H. M. Wallach. Conditional random fields: An introduction. Technical Report MS-CIS-04-21, Department of Computer and Information Science, University of Pennsylvania, 2004.

S. Watanabe. Information theoretical analysis of multivariate correlation. *IBM Journal of Research and Development*, 4:66–82, 1960.

T. Wennekers and N. Ay. Spatial and temporal stochastic interaction in neuronal assemblies. *Theory in Biosciences*, 122:5–18, 2003.

D. Wettschereck, D. W. Aha, and T. Mohri. A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. *Artificial Intelligence Review*, 11(1-5):273–314, 1997.

G. Widmer and M. Kubat. Learning in the presence of concept drift and hidden contexts. *Machine Learning*, 23(1):69–101, 1996.

W. Wienholt and B. Sendhoff. How to determine the redundancy of noisy chaotic time series. *Int. J. Bifurcation Chaos*, 6(1):101–117, 1996.

D. H. Wolpert and D. R. Wolf. Estimating functions of distributions from a finite set of samples. *Physical Review E*, 52(6):6841–6854, 1995.

A. K. C. Wong and T. S. Liu. Typicality, diversity, and feature pattern of an ensemble. *IEEE Trans. on Computers*, C-24(2):158–181, February 1975.

T. Yairi, S. Nakasuka, and K. Hori. Sensor fusion for state abstraction using Bayesian classifier. In *IEEE International Conference on Intelligent Engineering Systems (INES'98)*, pages 90–104, 1998.

J. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2001.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. Technical Report TR2004-040, MERL, 2004.

D. Yekutieli and Y. Benjamini. Resampling-based false discovery rate controlling multiple test procedures for correlated test statistics. *Journal of Statistical Planning and Inference*, 82:171–196, 1999.

R. W. Yeung. *A First Course in Information Theory*. Kluwer Academic/Plenum Publishers, New York, 2002.

R. W. Yeung. A new outlook on Shannon's information measures. *IEEE Trans. on Information Theory*, 37:466–474, May 1991.

L. A. Zadeh. Fuzzy sets. *Information and Control*, 8:338–353, 1965.

N. L. Zhang, T. D. Nielsen, and F. V. Jensen. Latent variable discovery in classification models. *Artificial Intelligence in Medicine*, 30(3):283–299, 2004.

C. L. Zitnick and T. Kanade. Maximum entropy for collaborative filtering. In *Proc. of UAI 2004*, pages 636–643, 2004.

B. Zupan. *Machine Learning Based on Function Decomposition*. PhD thesis, University of Ljubljana, Faculty of Computer and Information Science, 1997.

B. Zupan, M. Bohanec, J. Demšar, and I. Bratko. Learning by discovering concept hierarchies. *Artificial Intelligence*, 109:211–242, 1999.

B. Zupan, J. Demšar, D. Smrke, K. Božikov, V. Stankovski, I. Bratko, and J. R. Beck. Predicting patient's long term clinical status after hip arthroplasty using hierarchical decision modeling and data mining. *Methods of Information in Medicine*, 40:25–31, 2001.