To appear in ECML'01

Understanding Probabilistic Classifiers

Ashutosh Garg and Dan Roth

Department of Computer Science and the Beckman Institute
University of Illinois, Urbana, IL. 61801, USA
{ashutosh,danr}@uiuc.edu

Abstract. Probabilistic classifiers are developed by assuming generative models which are product distributions over the original attribute space (as in naive Bayes) or more involved spaces (as in general Bayesian networks). While this paradigm has been shown experimentally successful on real world applications, despite vastly simplified probabilistic assumptions, the question of why these approaches work is still open.
This paper resolves this question. We show that *almost all* joint distributions with a given set of marginals (i.e., all distributions that could have given rise to the classifier learned) or, equivalently, almost all data sets that yield this set of marginals, are very close (in terms of distributional distance) to the product distribution on the marginals; the number of these distributions goes down exponentially with their distance from the product distribution. Consequently, as we show, for almost all joint distributions with this set of marginals, the penalty incurred in using the marginal distribution rather than the true one is small. In addition to resolving the puzzle surrounding the success of probabilistic classifiers our results contribute to understanding the tradeoffs in developing probabilistic classifiers and will help in developing better classifiers.

1 Introduction

Probabilistic classifiers and, in particular, the archetypical naive Bayes classifier, are among the most popular classifiers used in the machine learning community and increasingly in many applications. These classifiers are derived from generative probability models which provide a principled way to the study of statistical classification in complex domains such as natural language and visual processing.

The study of probabilistic classification is the study of approximating a joint distribution with a product distribution. Bayes rule is used to estimate the conditional probability of a class label y, and then assumptions are made on the model, to decompose this probability into a product of conditional probabilities.

$$Pr(y|x) = Pr(y|x^1, x^2, \ldots x^n) = \Pi_{i=1}^{n} Pr(x^i|x^1, \ldots x^{i-1}, y) \frac{Pr(y)}{Pr(x)} = \Pi_{j=1}^{n'} Pr(y^j|y) \frac{Pr(y)}{Pr(x)},$$

where $x = (x^1, \ldots, x^n)$ is the observation and the $y^j = g_j(x^1, \ldots x^{i-1}, x^i)$, for some function g_j, are independent given the class label y.

While the use of Bayes rule is harmless, the final decomposition step introduces independence assumptions which may not hold in the data. The functions g_j encode the

probabilistic assumptions and allow the representation of any Bayesian network, e.g., a Markov model. The most common model used in classification, however, is the *naive Bayes* model in which $\forall j, g_j(x^1, \ldots x^{i-1}, x^i) \equiv x^i$. That is, the original attributes are assumed to be independent given the class label.

Although the naive Bayes algorithm makes some unrealistic probabilistic assumptions it has been found to work remarkably well in practice [4, 3]. Roth [10] develops a partial answer to this unexpected behavior using techniques from learning theory. It is shown that naive Bayes and other probabilistic classifiers are all "Linear Statistical Query" classifiers; thus, PAC type guarantees [12] can be given on the performance of the classifier on future, previously unseen data, as a function of its performance on the training data, independently of the probabilistic assumptions made when deriving the classifier. However, the key question that underlies the success of probabilistic classifiers is still open. That is, why is it even possible to get good performance on the training data, i.e., to "fit the data"[1] with a classifier that relies heavily on extremely simplified probabilistic assumptions on the data?

This paper resolves this question and develops arguments that could explain the success of probabilistic classifiers and, in particular, that of naive Bayes. We start by quantifying the optimal Bayes error as a function of the entropy of the data. We develop upper and lower bounds on this term, and discuss where do most of the distributions lie relative to these bounds. While this gives some idea as to what can be expected in the best case, we would like to quantify what happens in realistic situations, when the probability distribution is not known. Quantifying the penalty incurred due to the independence assumptions allows us to show its direct relation to the distributional distance between the true (joint) and the product distribution over the marginals used to derive the classifier. This is used to derive the main result of the paper which, we believe, explains the practical success of product distribution based classifiers. Informally, we show that *almost all* joint distributions with a given set of marginals (that is, all distributions that could have given rise to the classifier learned)[2] are very close to the product distribution on the marginals - the number of these distributions goes down exponentially with their distance from the product distribution. Consequently, the error incurred when predicting using the product distribution is small for *almost all* joint distributions with the same marginals.

There is no claim in this paper that distributions governing "practical" problems are sampled according to a uniform distribution over these marginal distributions. Clearly, there are many distributions for which the product distribution based algorithm will not perform well (e.g., see [10]) and in some situations, these could be the interesting distributions. The counting arguments developed here suggest, though, that "bad" distributions are relatively rare.

Finally, we show how these insights may allow one to quantify the potential gain achieved by the use of complex probabilistic models thus explaining phenomena observed previously by experimenters.

---

[1] We assume here a fixed feature space; clearly, by blowing up the feature space it is always possible to fit the data.

[2] Or, equivalently, as we show, almost all data sets with this set of marginals.

It is important to note that this paper ignores small sample effects. We do not attend to learnability issues but rather assume that good estimates of the statistics required by the classifier can be obtained; the paper concentrates on analyzing the properties of the resulting classifiers.

2 Preliminaries

We consider the standard binary classification problem in a probabilistic setting. In this model one assumes that data elements (x, y) are sampled according to some arbitrary distribution P on $\mathcal{X} \times \{0, 1\}$. $\mathcal{X}$ (e.g., $\mathcal{X} = \Re^M$) is the instance space and $y \in \{0, 1\}$ is the label. The goal of the learner is to determine, given a new example $x \in \mathcal{X}$, its most likely corresponding label $y(x)$, which is chosen as follows:

$$y(x) = \arg\max_{i \in \{0,1\}} P(y = i|x) = \arg\max_{i \in \{0,1\}} P(x|y = i)\frac{P(y = i)}{P(x)}.$$

We define the following distributions over $\mathcal{X}$: $P_0 \doteq P(x|y = 0)$ and $P_1 \doteq P(x|y = 1)$. With this notation, the Bayesian classifier predicts $y = 0$ iff $P_0(x) > P_1(x)$.

Throughout the paper we will use capital letters (X, Y, Z) to denote random variables and lower case (x, y, z) to denote particular instantiation of them. $P(x)$ refers to the probability of random variable X taking on value x. $P^n(\cdot)$ refers to the joint probability of observing a sequence of n i.i.d samples distributed according to P.

Definition 1. *Let $X = (X^1, X^2, ..., X^M) \in \mathcal{X}$ be a random vector and P a probability distribution over $\mathcal{X}$. The* marginal distribution *of the ith component of X (P^i) and the* product distribution*(P_m) induced by P over $\mathcal{X}$ are defined, resp. as $P^i = \sum_{X \setminus X^i} P(X)$; $P_m = \prod_i P^i$. P_m is identical to P under the assumption that the components X^i of X are independent of each other. We sometimes call P_m the* marginal distribution *of P.*

Definition 2 (Entropy; Kullback-Leibler Distance). *For probability distributions P, Q over $\mathcal{X}$ the entropy of P and the Kullback-Leibler distance between P and Q and the* conditional entropy *of a random variable x given y, are defined, resp. by*

$$H(P) = -\sum_{x \in \mathcal{X}} P(x) \log P(x) \; ; \; D(P||Q) = \sum_{x \in \mathcal{X}} P(x) \log \frac{P(x)}{Q(x)} \; ; \; H(x|y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \{0,1\}} P(x, y) \log P(x|y)$$

3 Bayes Optimal Error and Entropy

Given a sample $\{(x, y)\}_1^n$ sampled according to P we are interested in studying the optimal Bayes error achievable on it. Assuming, for simplicity, that the two classes are equally likely ($P(y = 1) = P(y = 0) = \frac{1}{2}$), the optimal Bayes error is given by $\frac{1}{2}P_0(\{x|P_1(x) > P_0(x)\}) + P_1(\{x|P_0(x) > P_1(x)\})$.

Lemma 1. *[2] The Bayes optimal error under the uniform class prob. assumption is:*

$$\epsilon = \frac{1}{2} - \frac{1}{4}\sum_x |P_0(x) - P_1(x)|. \quad (1)$$

Note that $P_0(x)$ and $P_1(x)$ are independent quantities and can be changed without influencing each other. [2] also gives the relation between the Bayes optimal error and the entropy of the class label conditioned upon the data, $P(y|x)$.

$$-\log(1-\epsilon) \leq H(P(y|x)) \leq -\epsilon \log \epsilon - (1-\epsilon)\log(1-\epsilon), \tag{2}$$

where the left hand side inequality is same as the Fano's inequality [1] and the right side follows by the direct application of the Jensen's inequality. However, $P(y|x)$ is typically not always available and thus the use of this bound depends on learning a probabilistic classifier. Now we derive a relation between the lowest achievable Bayes error and the conditional entropy of the input data given the class label thus allowing for an assessment of the optimal performance of the Bayes classifier just by looking at the given data. Naturally, the relation obtained between error and entropy is much loser, compared to the one given in Eqn 2, as has been documented in previous attempts to develop bounds of this sort[5]. We assume a domain of size M, $x \in \{0, 1, \ldots, M-1\}$ and $y \in \{0, 1\}$. Let $H_b(p)$ denotes the binary entropy $H_b(p) = -(1-p)\log(1-p) - p\log p$. Then we have:

Theorem 1. *Assuming equal class probabilities and an optimal Bayes error of ϵ, the conditional entropy $H(x|y)$ of input data conditioned upon class label is bounded by*

$$\frac{1}{2}H_b(2\epsilon) \leq H(x|y) \leq H_b(\epsilon) + \log\frac{M}{2}. \tag{3}$$

We prove the theorem using the following sequence of lemmas (For proofs, please see [8]). For simplicity, our analysis assumes that M is an even number. The general case follows similarly.

Lemma 2. *Consider two probability distributions P, Q defined over $x \in \{0, 1, \ldots, M-1\}$. Let $p_i = P(x=i)$ and $q_i = Q(x=i)$. Assume that the two distributions are constrained such that $\sum_i |p_i - q_i| = \alpha$. Then the sum of the entropy ($H(P) + H(Q)$) of two distributions is maximized when for some $0 \leq K \leq M$,*

$$\forall i: 0 \leq i \leq K \;\; p_i = c_1, \;\; q_i = d_1 \quad \forall i: K < i \leq M \;\; p_i = c_2, \;\; q_i = d_2$$

Where c_1, c_2, d_1, d_2 are some constants (which are functions of α, K, M).

Lemma 3. *The entropy $H(P) + H(Q)$ from Lemma 2 achieves maxima at K=M/2.*

When M is odd, due to the concavity and symmetry of the entropy function, maximum entropy is achieved when K is either $\frac{M+1}{2}$ or $\frac{M-1}{2}$.

The next lemma is used later to develop the lower bound on the conditional entropy.

Lemma 4. *Let P, Q be probability distributions such that $\sum_i |p_i - q_i| = \alpha$. The sum $H(P)+H(Q)$ of their entropies is minimized when for some K and some j, $0 \leq j \leq K$, $p_j = \frac{\alpha}{2}$ and $\forall i: 0 \leq i \leq K, i \neq j, p_i = 0$ and $\forall i: 0 \leq i \leq K$, $q_i = 0$. And for some $j: K < j \leq M, p_j = 1 - \frac{\alpha}{2}, q_j = 1$ and $\forall i \neq j: K < i \leq M$, $p_i, q_i = 0$. That is,*

$$P = \{p_1 = 0, .., p_j = \frac{\alpha}{2}, 0, p_K = 0, ..., p_i = 1-\frac{\alpha}{2}, .., 0\} \;\; Q = \{0, .., 0, 0, 0, ..., p_i = 1, 0, ..., 0\}$$

Now we are in a position to prove Theorem 1. Lemma 2 and 3 are used to prove the upper bound and Lemma 4 is used to prove the lower bound on the entropy.

Proof: (**Theorem 1**) We assume $P(y=0) = P(y=1) = \frac{1}{2}$ and a Bayes optimal error of ϵ. For upper bound we would like to obtain P_0 and P_1 that achieve the maximum conditional entropy, given by $H(x|y) = \frac{1}{2}H(P_0(x)) + \frac{1}{2}H(P_1(x))$. Since we are constraining the Bayes optimal error to be ϵ, we can write it as $\sum_x |P_0(x) - P_1(x)| = 4 - 2\epsilon = \alpha$. Since the distributions that maximize the conditional entropy will also maximize the sum of the entropies of the two distributions (P_0, P_1 because of equal class probability assumption), we can use the results given in Lemma 2,3 to obtain such distributions. Treating $P_1(x)$ as P and $P_0(x)$ as Q, we obtain the distributions that maximize the conditional entropy:

$$P_0 = \{\frac{1+\frac{\alpha}{2}}{M}, \frac{1+\frac{\alpha}{2}}{M}, ..., \frac{1-\frac{\alpha}{2}}{M}, \frac{1-\frac{\alpha}{2}}{M}\} \quad P_1 = \{\frac{1-\frac{\alpha}{2}}{M}, \frac{1-\frac{\alpha}{2}}{M}, ..., \frac{1+\frac{\alpha}{2}}{M}, \frac{1+\frac{\alpha}{2}}{M}\} \tag{4}$$

The conditional entropy for this distribution is $H(x|y) = H_b(\epsilon) + \log\frac{M}{2}$.

To prove the lower bound on the conditional entropy given Bayes optimal error of ϵ we use the distributions given by Lemma 4:

$$P_0 = \{0, 0, ..., 1-\frac{\alpha}{2}, ..., 0, \frac{\alpha}{2}, 0, ..., 0\} \quad P_1 = \{0, 0, ..., 1, ..., 0, 0, 0, ..., 0\} \tag{5}$$

The entropy for this distribution is given by $H(x|y) = \frac{1}{2}H_b(2\epsilon)$. ■

The results of the theorem are depicted in Figure 1 for $M = 4$. The bounds imply that the points outside the shaded area cannot be realized. It is interesting to see that the bound is tight in the sense that there are distributions on the boundary of the curves. This also addresses the common misconception that "low entropy implies low error and high entropy implies high error". Our analysis shows that while the latter is correct, the former may not be. We observe that when the entropy is zero, the error can either be 0 (no error, perfect classifier, point (A) on graph) or 50% error (point (B) on graph). Although somewhat counterintuitive, consider:

Example 1. Let $P_0(x=1) = 1$ and $P_0(x=i) = 0, \forall i \neq 1$ and $\forall x, P_1(x) = P_0(x)$. Then $H(x|y) = 0$ since $H(P_0(x)) = H(P_1(x)) = 0$ and the probability of error is 0.5.

The other critical points on this curve are also realizable. Point "D", which corresponds to the maximum entropy is achieved only when $P_0(x) = \frac{1}{M}, \forall x$ and $P_1(x) = \frac{1}{M}$. Again the error is 0.5. Point (C) corresponds to the maximum entropy with 0 achievable error. It is given by $H(P(y|x)) = \log\frac{M}{2}$. Finally, point (E) corresponds to the minimum entropy for which there exists a distribution for any value of optimal error. This corresponds to $entropy = 0.5$. Continuity arguments imply that all the shaded area is realizable. At a first glance it appears that the points (A) and (C) are very far apart, as (A) corresponds to 0 entropy where as (C) corresponds to entropy of $\log\frac{M}{2}$. One might think that most of the joint probability distributions are going to be between (A) and (C) - a range for which the bounds are vacuous. It turns out, however, that most of the distributions actually lie beyond the $\log\frac{M}{2}$ entropy point.

Theorem 2. *Consider a probability distribution over $x \in \{0, 1, ..., M-1\}$ given by $P = [p_0, ..., p_{M-1}]$ and assume that $H(p) \leq \log\frac{M}{2}$. Then, $\forall\delta : \ 0 < \delta < \frac{1}{M}$, the distribution Q defined by $q_i = \frac{1}{M} + \delta(p_i - \frac{1}{M}), \ \ \forall i$ satisfies $H(Q) > \log\frac{M}{2}$.*

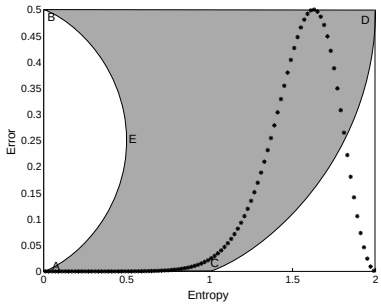


Fig. 1. The relation between the error and the conditional entropy of the data. The shaded region represents the feasible region (the distributions with the corresponding error and entropy are realizable). The dotted curve gives the empirical distribution of the joint distributions over a given set of input features.

Proof: To show that $H(Q) > \log \frac{M}{2}$ consider $H(Q) - \log \frac{M}{2} = -\sum_{i=1}^{M} q_i \log \left(q_i \frac{M}{2}\right)$. Now if $0 < \delta < 1$ then it is straightforward to see that $\forall i, q_i > 0$, and if $0 < \delta < \frac{1}{M}$ then $\forall i, \quad \frac{M}{2} q_i < 1$, implying that $H(Q) > \log \frac{M}{2}$. Since $H(P) \leq \log \frac{M}{2}$, $P \neq Q$. Hence, for each δ we have defined a 1-1 mapping of distributions with entropy below $\log \frac{M}{2}$ to those with entropy above it. ■

Consequently, the number of distributions with entropy above $\log \frac{M}{2}$ is at least as much as the number of those with entropy below it. This is illustrated using the dotted curve in Figure 1 for the case $M = 4$. For the simulations we fixed the resolution and did not distinguish between two probability distributions for which the probability assignments for all data points is within some small range. We then generated all the conditional probability distributions and their (normalized) histogram. This is plotted as the dotted curve superimposed on the bounds in Figure 1. It is evident that most of the distributions lie in the high entropy region, where the relation between the entropy and error in Thm. 1 carries useful information.

4 Classification Error

While previously we bounded the Bayes optimal error assuming the correct joint probability is known, in this section we start investigating the more interesting case – the mismatched probability distribution. We assume that the learner has estimated a probability distribution that is different from the true joint distribution. The performance measure used in our study is the *probability of error*. We can look at the probability of misclassification from the perspective of hypothesis testing. That is, this is the probability of misclassifying a sample as coming from hypothesis $H_0 \sim P_0(X^1, ..., X^M)$ when it actually came from $H_1 \sim P_1(X^1, ..., X^M)$, and vice versa. Although slightly different from the standard classification problem, the hypothesis testing framework provides two advantages. It provides tools for the analysis of the mismatched probability distributions and at the same time, allows one to analyze the asymptotic probability of error. Since it is assumed that the distributions P_0, P_1 have already been estimated from data, this perspective (i.e., looking at many samples $X^1, ..., X^M$) allows us to

obtain better bounds for the performance of these estimates. Under this framework, the probability of error can be grouped into two categories α, β. Where α (Type I error) is the probability of misclassification when the true hypothesis is $H_0 \sim P_0$ and β (Type II error) is the misclassification error when the true hypothesis is $H_1 \sim P_1$. Formally, if $A = \{x : \frac{P_0(x)}{P_1(x)} > \tau\}$ is the acceptance region for hypothesis H_0 then $\alpha = P_0(A^c)\ ;\ \beta = P_1(A)$. Now, A_n, α_n, β_n denote the corresponding terms when the decision is made for n random vectors.

Stein's lemma [1] gives asymptotic bounds on the performance of a classifier which is using Likelihood ratio test for deciding between the two hypotheses. It shows that under the condition that $\alpha_n < \epsilon$, and for $0 < \epsilon < \frac{1}{2}$, defining $\beta_n^\epsilon = \min_{\alpha_n < \epsilon} \beta_n$ gives

$$\lim_{\epsilon \to 0} \lim_{n \to \infty} \frac{1}{n} \log \beta_n^\epsilon = -D(P_0||P_1) \tag{6}$$

In practice, however, rather than the true joint distribution over the samples, the induced product distribution (derived using conditional independence assumptions) is used. The standard Stein's lemma doesn't hold in this case and we prove a modified version of it for this case.

Theorem 3. *(Modified Stein's Lemma) Let $X_1, ..., X_n$ be $i.i.d \sim Q$. Consider the hypothesis test between two hypothesis $Q \sim P_0$, and $Q \sim P_1$. Let A_n be the acceptance region for hypothesis $H_0 = Q \sim P_0$. The probabilities of error can then be written as $\alpha_n = P_0^n(A_n^c);\ \beta_n = P_1^n(A_n)$. Assume $P_0^{'}$ is used instead of P_0 for the likelihood ratio test. Then if A_n is chosen such that $\alpha_n < \epsilon$, then the type II error (β) is given by*

$$\epsilon \overset{\lim}{\to} 0 \ \ n \overset{\lim}{\to} \infty \ \frac{1}{n} \log \beta_n^\epsilon = -D_{P_0}(P_0^{'}||P_1) = -E_{P_0}(\log \frac{P_0^{'}}{P_1}). \tag{7}$$

For the proof, please refer to [8]. By writing $D_{P_0}(P_0^{'}||P_1)$ in a more recognizable form, the asymptotic bound on the error can be written as

$$\frac{1}{n} \log(error) \leq -D_{P_0}(P_0^{'}||P_1) = -D(P_0||P_1) + D(P_0||P_0^{'}) \tag{8}$$

The first term on the right hand side of Eqn 8 is the same as the one in the original Stein's Lemma. Since Stein's lemma gave the minimum error achievable by any algorithm, we can't do better than this quantity which can be viewed as a "baseline" term. Improving the approximation affects the second term - the distance between the true distribution and the approximation - which acts as the actual penalty.

Although the bound is derived under the assumption that only the distribution corresponding to one hypothesis is approximated, a similar bound can be derived for the more general case (when the distributions corresponding to both hypothesis are unknown) under the condition that $P_1(x) \leq KP_1^{'}(x)$ for some finite K. In this case, the bound will be given by $\log D_{P_0}(P_0^{'}||P_1^{'})$. The condition is fairly general and always holds for product distributions. However, the bound given by Eqn 7 highlights some basic properties of the distributions and will be analyzed in the rest of the paper. The

[2] For the purpose of the analysis of the performance, we study performance using error on the sample.

general case follows similar arguments. Eqn 8 shows that the additional penalty term is related to $D(P_0||P_0^{'})$, with P_0 being the true distribution and $P_0^{'}$ the approximation. In the special case when both P_1 and $P_0^{'}$ are product form distributions, we have

$$D_{P_0}(P_0^{'}||P_1) = \sum_x P_0(x) \log \frac{P_0^{'}(x)}{P_1(x)} = \sum_{x^1,x^2,\ldots,x^n} P_0(x^1, x^2, ..., x^n) \sum_i \log \frac{P_0(x^i)}{P_1(x^i)}$$
$$= \sum_i P_0(x^i) \log \frac{P_0(x^i)}{P_1(x^i)} = D(P_0^{'}||P_1) = D(P_0||P_1) - D(P_0||P_0^{'}) \quad (9)$$

Corollary 1. *If both $P_0^{'}$ and P_1 are product distributions then $\frac{1}{n}\log(error) \leq -D(P_0^{'}||P_1)$, i.e. the bound is independent of the joint distribution and depends just on the marginals.*

5 Density of Distributions

Let P_m be the product distribution induced by P. As mentioned before, given data sampled according to P, probabilistic algorithms estimate P_m and use it for classifying future data. In this section we explain why making classifications using an induced product distribution rather than the true joint distribution works well in practice.

Given the bound in Eq. 8, classifying data from P using P_m incurs penalty $D(P||P_m)$ (in addition to the baseline term there, which reflects the error *any* classifier must make). This section shows that given P_m, for almost all distributions P that induce P_m, the error term $D(P||P_m)$ is small; equivalently, as we show, for almost all data sets sampled according to distributions that induce P_m the error term is small.

For analysis, discrete domain $\{0,1\}^m$ is assumed (although it can be extended to the case of continuous random variables). The analysis is based on the *method of types* [1] which allows one to study the number of sequences of length n that can be observed when sampling according to distribution P. We first provide some preliminaries.

Definition 3. *Let $\mathbf{x} = \{x_1, \ldots x_n\}$ be a sequence of n symbols from an alphabet $\mathcal{H}$. The* type *$P_\mathbf{x}$ (or empirical probability distribution) of a sequence $x_1, ..., x_n$ is the relative proportion of occurrences of each symbol of $\mathcal{H}$, i.e. $P_\mathbf{x} = \frac{N(a|\mathbf{x})}{n}$ for all $a \in \mathcal{H}$, where $N(a|\mathbf{x})$ is the number of times symbol a occurs in the sequence $\mathbf{x} \in \mathcal{H}^n$.*

Definition 4. *Let $\mathcal{P}_n$ denotes the set of types with denominator n (i.e. set of empirical probability distributions derived from sequences of length n.) For $P \in \mathcal{P}_n$, the set of sequences of length n and type P is called the type class of P, denoted $T(P)$.*

Example 2. Let $\mathcal{H} = \{1,2,3\}$, $\mathbf{x} = 11321$. The type $P_\mathbf{x}$ is $P_\mathbf{x}(1) = 3/5, P_\mathbf{x}(2) = 1/5, P_\mathbf{x}(1) = 1/5$. The type class of $P_\mathbf{x}$ is the set of all sequences of length 5 with three 1's, one 2 and one 3. There are 20 such sequences, that is $|T(P_\mathbf{x})| = 20$.

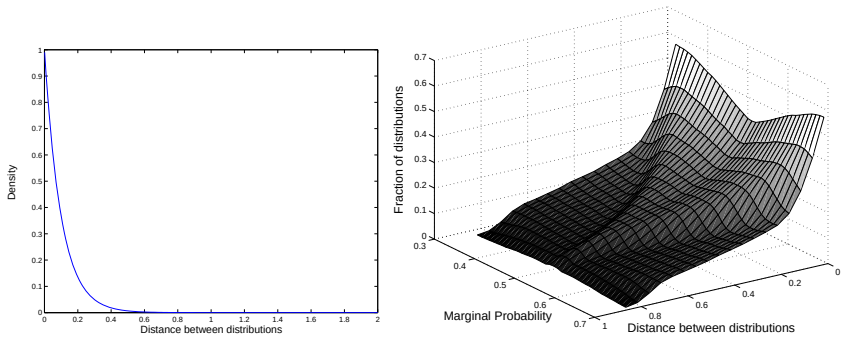
Notice the similarity to the case studied in this paper (with $\mathcal{H} = \{0,1\}^M$). All data sets in $T(P_\mathbf{x})$ induce the same product distribution $P_\mathbf{x}$*conditional entropy* of a random variable x given y

Theorem 4. *[1] For any probability distribution $P \in \mathcal{P}_n$, $\frac{1}{(n+1)^{|\mathcal{H}|}} 2^{nH(P)} \leq |T(P)| \leq 2^{nH(P)}$. That is, within a polynomial approximation $|T(P)| = 2^{nH(P)}$. (Here $|\mathcal{H}|$ is the alphabet size.)*

We note that it is possible to write an exact term for $|T(P)|$ as a multinomial; $|T(P)| = \frac{n!}{\prod_{a\in\mathcal{H}}(nP(a))!}$ where $(nP(a))$ is the expected number of time one observes symbol a in a sequence of length n. However, we will use the powerful relation to entropy and thus to prediction error. The following lemma uses the concept of sample entropy defined as $-\frac{1}{n}\log P^n(x_1, x_2, ..., x_n) = -\sum_{a\in\mathcal{H}}\frac{N(a)}{n}\log P(x = a)$. Here $N(a)$ refers to the number of time $x = a$ is observed in the sample (i.i.d) of length n. We know that as $n \to \infty$, $\frac{N(a)}{n} \to P(a)$ and hence known as sample entropy

Theorem 5. *[1] Let A^n_δ (typical set) denote the set of all the sequences with sample entropy as*

$$H(P) - \delta \leq -\frac{1}{n}\log P^n(x_1, x_2, ..., x_n) \leq H(P) + \delta \text{ Then } |A^n_\delta| \leq 2^{n(H(P)+\delta)}$$



(a) (b)

Fig. 2. (a) Density of sequences of length n. Y-axis gives the number of sequences ($K2^{-n\epsilon}$) as a function of the distance of the true joint distribution from the product distribution ($D(P_0||P_{m0}) = \epsilon$) in the X-axis. (b) shows the decay in the number of the distributions as a function of the entropy of the marginal distribution and the distance of the joint distribution and its induced product distribution. Plots are based on a two attributes case. P_m varies from [0.3 0.3] to [0.7 0.7] (i.e., the attributes have the same marginal distribution.)

This theorem gives a bound on the number of sequences with a given entropy. We now present the main result of this section:

Theorem 6. *Let P_m be a product distribution and let $\mathcal{P}$ be the collection of all probability distributions P that induce P_m, and such that $D(P||P_m) = \epsilon$. Then, the number of sequences with joint probability P, for $P \in \mathcal{P}$, is equal to (within a polynomial approximation) $K\ 2^{-n\epsilon}$, for some constant K that is independent of P.*

Proof: (**Sketch**) Consider a probability distribution P such that $D(P||P_m) = \epsilon$. We know that

$$D(P||P_m) = \sum_x P(x)\log P(x) - \sum_x P(x)\log P_m(x) = H(P_m) - H(P), \qquad (10)$$

where the second equality follows from the argument given in Eqn 9. That is, the entropy of distributions for which $D(P||P_m) = \epsilon$ is $H(P) = H(P_m) - \epsilon$ and it decays as the distance between P, P_m increases. We know from the law of large numbers that the sample entropy converges to the true entropy, as the number of sample increases. Thus the number of sequences with entropy P is given by (using Theorem 5)

$$2^{n(H(P)+\delta)} = 2^{n(H(P_m)-\epsilon+\delta)} = K2^{-n\epsilon}, \tag{11}$$

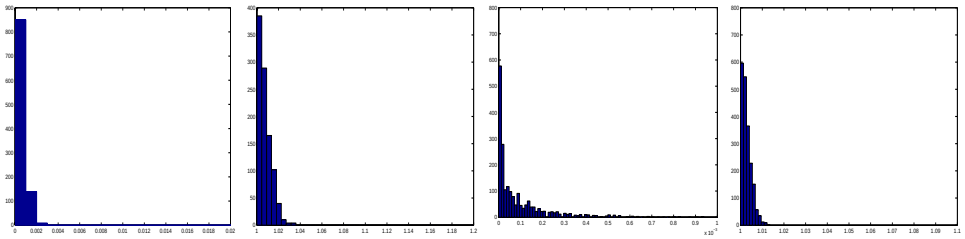
where δ (a small number) goes to zero as n increases. ■

The theorem shows that a randomly picked sequence of n elements $x \in \{0,1\}^M$ (a "data set") with a given marginal distribution over the individual features is likely to have true joint distribution that is very close to the marginal distribution. Equivalently, the probability of a data set which is ϵ away from the product distribution decays exponentially with ϵ (Figure 2(a)).

Together with the penalty results in Sec. 4 it is clear why we represent this in terms of the distance between the distributions. If, as probabilistic classifier do, classification with respect to P is done using the induced product distribution P_m, then the error incurred is related to $D(P||P_m)$ (disregarding the baseline term in Eq. 8). Therefore, Thm 6 implies that for most data sets, the classification error is going to be small. While the results above are phrased in terms of the number of data sets that are sampled according to distributions in a certain distance from the given product distribution, an equivalent result can be shown for the number of joint distributions in a certain distance from the induced product distribution. In both cases the decay is exponential in the distance. This is illustrated in Fig. 2(b). The histogram shows that the density of the joint distributions which have the same marginal distribution, as a function of the product distribution and the distance between the joint and the product distribution ($D(P||P_m)$[3]). e.g., consider two random variables $x^1, x^2 \in \{0,1\}$. Lets fix $P(x^1 = 1) = 0.8$ and $P(x^2 = 1) = 0.2$ (i.e. fixing the marginal distribution). This means that $P(x^1 = 1, x^2 = 1)$ can take only finite number of values (if we limit the resolution of the probabilities to say 0.001.) Thus it shows that the "bad" cases (when the distribution is far from marginal) are rare when considering the space of all possible distributions with a given marginal distribution (or all data sets sampled according to distributions with a given marginal). Note that, this is a upper bound analysis . Sometimes this bound is tight, as shown in Sec. 4 for the case in which P_1 is in product form. Nevertheless, there could be cases in which the bound is loose. However, the bound goes in the right direction, and in the majority of the cases the upper bound is small.

To show what happens in practice, some simulations are presented in Fig. 3. We considered a case of 2 and 3 features as input and the case of a binary classifier. In each case, 1000 sequences of fixed length were randomly sampled according to different joint distributions, all having the same induced product distribution. Plots of the number of sequences, with a joint distribution at a certain distance from the product distribution are given in Fig. 3 (a&c)(for $2\&3$ features resp.). As expected, the histogram looks very similar to the Fig. 2. Also shown (Fig. 3(b&d) for $2\&3$ features resp.) are the resulting classification errors as a function of the distance between the joint and the product distribution. The figures give the ratio of the errors made during classification when one

[3] Notice that this distance is always finite since P_m is 0 iff P is zero.



(a) (b) (c) (d)

Fig. 3. The plots (a),(b) gives the density of the joint distribution as a function of the distance from the product distribution. Plots (c),(d) gives the ratio of the errors after approximation (product distribution assumption) over the bayes optimal error (modeling complete joint distribution.)

Dataset	$D(P_0\|\|P_{m0})$	$D(P_1\|\|P_{m1})$	$D(P_0\|\|P_1)$	$D(P_1\|\|P_0)$	Avg. Diff	NB Res	TAN Res
Pima	0.0957	0.0226	0.9432	0.8105	0.8177	75.51±1.63	75.13±1.36
Breast	0.1719	0.4458	6.78	9.70	7.9311	97.36±0.50	*95.75±1.25*
Mofn-3-7-10	0.3091	0.3711	0.1096	0.1137	-0.2284	86.43±1.07	**91.70±0.86**
Diabetes	0.0228	0.0953	0.7975	0.9421	0.8108	74.48±0.89	75.13±0.98
Flare	0.5512	0.7032	0.8056	0.8664	0.2088	79.46±1.11	**82.74±1.60**

Table 1. This table compares the performance of naive Bayes classifier with the Tree augmented Bayes classifier (TAN). The results presented here are the ones published.The Avr. Diff. column is the average (over the two classes) of the distances between the TAN and the naive product distributions. It is evident that it explains the success (e.g., rows 3, 5) and failure (row 2) of TAN over the naive distribution.

uses the product distribution vs. the use of the true joint distribution. As expected the error ratio ($exp(-D(P||P_m))$) has an exponential decay.

6 Complex Probabilistic Models and Small Sample Effects

In the practice of machine learning [9, 11] the use of probabilistic classification algorithms is preceded by the generation of new features from the original attributes in the space which can be seen as using complex probabilistic classifiers. We analyze the particular case of tree augmented Bayesian (TAN) classifier introduced in [7], which is a sophisticated form of the naive Bayesian classifier modeling higher (second) order probabilistic dependencies between the attributes. They [7] conducted a number of experiments and reported improved results on some of the datasets. It is easy to see that by modeling the TAN distribution, one is essentially decreasing the distance between the true (joint) and the approximated distribution. i.e. $D(P||P_m) \geq D(P||P_{TAN})$ where P_{TAN} refers to the probability distribution modeled by TAN. Replacing P by either P_0 or P_1 reduces to the case presented in Section 4. Reduction in $D(P||P_m)$ is directly mapped to the reduction in the bound on error, thus explaining the better performance. Table 5 exhibits this result when evaluated on five data sets (chosen based on the number of attributes and training examples) studied in [7]. In addition to presenting the results

published in [7], we have computed, for each one of the classes (0, 1), the distance between the pure naive Bayes and the TAN distribution, and their average. The Avr. Diff. column is the average (over the two classes) of the distances between the TAN and the product distributions. Clearly our results predict well the success (rows 3, 5) and failure (row 2) of TAN over the naive Bayesian distribution.

As mentioned before, in this paper we have ignored small sample effects, and assumed that good estimates of the statistics required by the classifier can be obtained. In general, when the amount of data available is small, the naive Bayes classifier may actually do better than the more complex probability models because of the insufficient amount of data that is available. In fact, this has been empirically observed and discussed by a number of researchers [6, 7].

7 Conclusions

In the last few years we have seen a surge in learning work that is based on probabilistic classifiers. While this paradigm has been shown experimentally successful on many real world applications, it clearly makes vastly simplified probabilistic assumptions. This papers uses an information theoretic framework to resolve the fundamental question of: why do these approaches work. On the way to resolving this puzzle we develop methods for analyzing probabilistic classifiers and contribute to understanding the tradeoffs in developing probabilistic classifiers and thus to the development of better classifiers.

Acknowledgments: Research supported by NSF grants ITR-IIS-0085836, ITR-IIS-0085980 and IIS-9984168.

References

1. T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley and Sons, 1991.
2. L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Applications of Mathematics. Springer Verlag, 1996.
3. P. Domingos and Pazzani. M. Beyond independence: Conditions for the optimality of simple bayesian classifier. *Machine Learning*, 29:103–130, 1997.
4. C. Elkan. Boosting and naive bayesian learning. Technical Report CS97-557, Department of Computer Science, University of California, San Diego, 1997.
5. M Feder and N. Merhav. Relation between entropy and error probability. *IEEE Trans. on Information Theory*, 40:259–266, 1994.
6. J. H. Friedman. On bias, variance, 0/1-loss and curse-of-dimensionality. *Data Mining and Knowledge Discovery*, 55, 1997.
7. N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
8. A. Garg and D. Roth. Understanding probabilistic classifiers. Technical Report UIUCDCS-R-2001-2206, UIUC Computer Science Department, March 2001.
9. A. R. Golding. A Bayesian hybrid method for context-sensitive spelling correction. In *Proceedings of the 3rd workshop on very large corpora, ACL-95*, 1995.
10. D. Roth. Learning in natural language. In *Proc. of the International Joint Conference of Artificial Intelligence*, pages 898–904, 1999.
11. H. Schneiderman and T. Kanade. A statistical method for 3D object detection applied to faces and cars. In *The IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 746–751, 2000.
12. L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.