Discriminative vs Informative Learning

Y. Dan Rubinstein and Trevor Hastie
Department of Statistics
Stanford University
Stanford, CA 94305
ruby@stat.stanford.edu
trevor@stat.stanford.edu

Abstract

The goal of pattern classification can be approached from two points of view: informative - where the classifier learns the class densities, or discriminative - where the focus is on learning the class boundaries without regard to the underlying class densities. We review and synthesize the tradeoffs between these two approaches for simple classifiers, and extend the results to modern techniques such as Naive Bayes and Generalized Additive Models. Data mining applications often operate in the domain of high dimensional features where the tradeoffs between informative and discriminative classifiers are especially relevant. Experimental results are provided for simulated and real data.[1]

KDD and Classification

Automatic classification is among the main goals of data mining systems (Fayyad, Piatetsky-Shapiro, & Smyth 1996). Given a database of observations consisting of input (predictor) and output (response, i.e. class label) variables, a classifier seeks to learn relationships between the predictors and response that allow it to assign a new observation, whose response is unknown, into one of the K predetermined classes. The goal of good classification is to minimize misclassifications or the expected cost of misclassifications if some types of mistakes are more costly than others.

Classifiers can be segmented into two groups:

1. Informative: These are classifiers that model the class densities. Classification is done by examining the likelihood of each class producing the features and assigning to the most likely class. Examples include Fisher Discriminant Analysis, Hidden Markov Models, and Naive Bayes. Because each class density is considered separately from the others, these models are relatively easy to train.
2. Discriminative: Here, no attempt is made to model the underlying class feature densities. The focus is on modeling the class boundaries or the class membership probabilities directly. Examples include Logistic Regression, Neural Networks, and Generalized Additive Models. Because this requires simultaneous consideration of all other classes, these models are harder to train, often involve iterative algorithms, and do not scale well.

As we shall see below, these two approaches are related via Bayes rule, but often lead to different decision rules, especially when the class density model is incorrect or there are few training observations relative to the number of parameters in the model.

There are tradeoffs between the two approaches in terms of ease of training and classification performance. Precise statements can only be made for very simple classifiers, but the lessons can be applied to more sophisticated techniques. In this paper we review the known statistical results that apply to simple non-discriminative classifiers and we demonstrate how modern techniques can be categorized as being discriminative or not. Using Naive Bayes and GAM applied both to simulation and real data, we exemplify that, counter-intuitively, discriminative training may not always lead to the best classifiers. We also propose methods of combining the two approaches. We focus on parametric techniques although similar results obtain in the non-parametric case. With the advent of increasingly sophisticated classification techniques, it is important to realize what category the classifier falls in, because the assumptions, problems and fixes for each type are different.

Overview of Bayesian Classification Theory

Formally, the classification problem consists of assigning a vector observation $x \in \mathcal{R}^p$ into one of K classes. The true class is denoted by $y \in \{1, \ldots, K\}$. The classifier is a mapping that assigns class labels to observations: $\gamma : x \rightarrow \{1, \ldots, K\}$. There is also a cost matrix $c(r,s),\ \ r,s = 1, \ldots, K$ which describes the cost associated with misclassifying a member of class-r to class-s. A special case is 0/1 loss, $c(r,s) = 1 - \delta_{r,s} = 1$ if $r \neq s$ and 0 otherwise.

Underlying the observations is a true joint density $p(x,y) = p(y|x)p(x) = p(x|y)p(y)$ which is unknown. The goal is to minimize the total cost of errors, known as the overall risk and this is achieved by the Bayes classifier (Duda & Hart. 1973)

$$\begin{aligned} \gamma(x) &= \min_k{}^{-1} \sum_{m=1}^{K} c(k,m)p(y=m|x) && (1) \\ &= \max_k{}^{-1} p(y=k|x) \ \ (0/1 \text{ loss}). && (2) \end{aligned}$$

For 0/1 loss this reduces to classifying x to the class k for which the class posterior probability $p(y=k|x)$ is maximum.

In practice, the true density $p(x,y)$ is unknown and all that is available is a set of training observations (x_i, y_i) for $i = 1, \ldots, n$. Many classification techniques seek to estimate the class posterior probabilities $p(y=k|x)$, since we see in (2) that optimal classification can be achieved if these are known perfectly (for a discussion on the relationship between class posteriors and neural net outputs see (Ney 1995)).

For convenience in what follows, we will make use of the discriminant function

$$\lambda_k(x) = \log \frac{p(y=k|x)}{p(y=K|x)}.$$

This discriminant preserves the ordering of the class posterior probabilities and can be used instead of them for classification.

Informative Classification

Rather than estimate the class posteriors $p(y|x)$ directly, the class densities $p(x|y)$ and priors $p(y)$ are estimated. The operative equation here is Bayes rule, which gives the class posteriors in terms of the class densities and priors

$$p(y=k|x) = \frac{p(x|y=k)p(y=k)}{\sum_m^K p(x|y=m)p(y=m)}.$$

Typically some model is chosen for the class densities, for example gaussian, $p_\theta(x|y=k) = \mathcal{N}(x; \mu_k, \Sigma)$ (here $\theta = \{\mu_1, \ldots, \mu_K, \pi_1, \ldots, \pi_K, \Sigma\}$), and the model parameters are estimated from the data by maximizing the full log likelihood

$$\begin{aligned} \theta_{MLE} &= \max_\theta{}^{-1} \sum_i^n \log p_\theta(x_i, y_i) \\ &= \max_\theta{}^{-1} \sum_i^n \log p_\theta(x_i|y_i) + \log p_\theta(y_i). \end{aligned}$$

For the gaussian case this yields the well known estimates $\hat{\pi}_k = n_k/n, \hat{\mu}_k = \bar{x}_k = \frac{1}{n_k} \sum_{y_i=k} x_i, \hat{\Sigma} = \frac{1}{n} \sum_k^K \sum_{y_i=k} (x_i - \bar{x}_k)(x_i - \bar{x}_k)'$ where n_k is the number of observations from class k and $n = \sum_k^K n_k$. The discriminant functions are

$$\begin{aligned} \lambda_k(x) &= (\log \frac{\pi_k}{\pi_K} - \frac{1}{2}(\mu_k + \mu_K)\Sigma^{-1}(\mu_k - \mu_K)) + \\ &\quad (\mu_k - \mu_K)\Sigma^{-1} x \\ &= \beta_0^k + \beta^{k'} x \end{aligned}$$

and are linear in x. Note that while $Kp + p(p+1)/2 + (K-1)$ parameters are estimated, the discriminants involve only $(K-1)(p+1)$ parameters.

The important points with informative training are

1. A model $p_\theta(x|y)$ is assumed for the class densities.
2. The parameters are obtained by maximizing the full log likelihood $\log p_\theta(x,y) = \log p_\theta(y|x)p_\theta(x)$.
3. A decision boundary is induced, and the model parameters may appear in a way that reduces the effective number of parameters in the discriminant.

Discriminative Classification

The discriminative approach models the class posteriors and hence the discriminants directly. The discriminative approach is more flexible with regard to the class densities it is capable of modeling. By only restricting the discriminant $\lambda_k(x) = \log[p(y=k|x)/p(y=K|x)] = \log[p(x|y=k)p(y=k)/p(x|y=K)p(y=K)]$ we are capable of modeling any class densities that are exponential "tilts" of each other

$$p(x|y=k) = e^{\lambda_k(x)} p(x|y=K) \frac{p(y=K)}{p(y=k)}.$$

In particular, the informative model, as regards the class densities, is seen to be a special instance of the more general discriminative model. The example above was a special case with a gaussian as the "carrier" density

$$p(x|y=k) = \mathcal{N}(x; \mu_K, \Sigma) e^{\beta_0^k + \beta^{k'} x} \left(\frac{\pi_K}{\pi_k} \right)$$

while the corresponding discriminative model allows any carrier density

$$p(x|y=k) = f_K(x; \theta) e^{\beta_0^k + \beta^{k'} x} \left(\frac{\pi_K}{\pi_k} \right)$$

so long as the discriminant is linear.

Parameter estimation in the discriminative case is carried out by maximizing the condition log likelihood

$$\theta_{DISCR} = \max_{\theta}{}^{-1} \sum_{i}^{n} \log p_\theta(y_i|x_i).$$

On the one hand, maximizing the conditional likelihood is a natural thing to do because it is directly focused on the class posteriors $p(y|x)$ which is all that is required in order to classify well. However, it is ignoring part of the data, namely, the marginal distribution $p(x)$. Compare to the full likelihood case where each observation contributes $p(x, y) = p(y|x)p(x)$. The discriminative approach, which uses only the first term on the right side, throws away the information in the marginal density of x. Thus, if the class density model is correct, the discriminative approach ignores useful information. However, ignoring the class models may be good if they are incorrect. The table below summarizes the main comparisons between the two approaches.

	Informative	Discriminative
Objective Function	Full log likelihood $\sum_i \log p_\theta(x_i, y_i)$	Conditional log likelihood $\sum_i \log p_\theta(y_i|x_i)$
Model Assumptions	Class densities $p(x|y = k)$	Discriminant functions $\lambda_k(x)$
Parameter Estimation	"Easy"	"Hard"
Advantages	More efficient if model correct, borrows strength from $p(x)$	More flexible, robust because fewer assumptions
Disadvantages	Bias if model is incorrect.	May also be biased. Ignores information in $p(x)$

Logistic Regression vs Linear Discriminant Analysis

A lot of insight can be gained from examining the two class case where the class densities are assumed to be Gaussian $p_\theta(x|y = k) = \mathcal{N}(\mu_k, \Sigma)[x] = \frac{1}{\sqrt{(2\pi)^p \det \Sigma}} \exp -(\frac{1}{2}(x - \mu^k)'\Sigma^{-1}(x - \mu^k))$ with priors $p_\theta(y = k) = \pi_k$.

When the populations are gaussian, informative classification is more efficient than discriminative, ie fewer training observations are required or for a fixed number of training observations, better classification is obtained (Efron 1975; O'Neill 1980; Ruiz-Velasco 1991). Even when the class densities are not gaussian there are circumstances - such as when the classes are well separated - when informative training does about as well as discriminative (Byth & McLachlan 1980).

The informative approach requires estimating class means and a pooled covariance which requires only a single sweep through the data. The discriminative approach requires an iterative optimization via a gradient descent of the conditional likelihood.

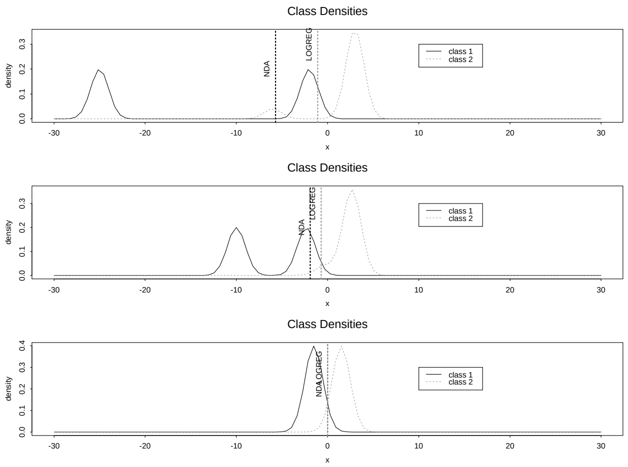

Figure 1: Class densities for 3 cases of simulation data. The class boundaries derived from many (10000) training observations for Normal Discriminant Analysis (LDA) and Logistic Regression (LOGREG) are shown: points to the left of the boundary are classified to class 1.

Figure 1 shows class densities for 3 simulation experiments. Case 3 is a gaussian class case for which we expect LDA to do better than LOGREG when the models are learned using training data. For each case, 100 training sets with 5 observations per class, i.e. $p(y = 1) = p(y = 2) = 1/2$, were drawn according to the class densities pictured. LDA and LOGREG classifiers were trained for each set and the exact probability of error was computed using integration over a grid $P(\text{error}) = \frac{1}{2}\left[\int_{\gamma(x)=1} p(x|y = 2)dx + \int_{\gamma(x)=2} p(x|y = 1)dx\right]$.

The table below provides error rates using the two procedures. Each column corresponds to a different density case as depicted in figure 1. The first two rows are "best" in the sense that the model is trained using the complete density, not a sample of training observations. The remaining rows are averages and standard errors of the error rates across 100 training sets, each of which contained 5 observations per class.

case	1	2	3
LDA - best	28.1	8.6	6.7
LOGREG - best	8.8	3.1	6.7
LDA	25.2	9.6	7.6
SE(LDA)	0.47	0.61	0.12
LOGREG	12.6	4.1	8.1
SE(LOGREG)	0.94	0.17	0.27

As expected, LDA did better than LOGREG when the classes were gaussian (case 3). An interesting result in case 1 is that LDA does significantly better (25.2% vs 28.1%) when it does not know the true distributions. In this case, it is because the true distribution is highly non-gaussian. When the number of observations are few relative to their dimensionality, informative methods may do surprisingly well even when the model is incorrect (see also the GAM/Naive Bayes example below).

StatLog data

The StatLog experiments compared several classification techniques on various datasets. For most of the datasets, logistic discrimination did better than the corresponding informative approach of LDA (Michie, Spiegelhalter, & Taylor 1994). However, there were several cases, such as the chromosome dataset, in which LDA did better than logistic discrimination. For these cases, the informative model apparently makes use of important information in the marginal density $p(x)$.

Naive Bayes and GAM

Naive Bayes classifiers are a specialized form of a Bayesian network (John & Langley 1995; Langley & Sage 1994) and fall into the informative category of classifiers. The class densities assume independence among the predictors

$$\begin{aligned} p(x|y=k) &= \prod_j^p p(x_j|y=k) \\ \Rightarrow \log p(x|y=k) &= \sum_j^p \log p(x_j|y=k) \\ &= \sum_j^p g_{k,j}(x_j), \end{aligned}$$

and are naive for this reason. Langley (John & Langley 1995) considered class densities that are products of univariate gaussians as well as "flexible" gaussian kernel densities.

The corresponding discriminative procedure is known as a Generalized Additive Model (GAM) (Hastie & Tibshirani 1990). GAM's assume that the log ratio of class posteriors is additive in each of the predictors $x_j,\ j=1,\ldots,p$

$$\log \frac{p(y=k|x)}{p(y=K|x)} = \sum_j^p f_{k,j}(x_j) + \text{const}_k.$$

Theorem 1 *Naive Bayes classifiers are a specialized case of GAM.*

Proof It suffices to show that the induced discriminant is log additive.

$$\begin{aligned} \log \frac{p(y=k|x)}{p(y=K|x)} &= \log \frac{p(x|y=k)p(y=k)}{p(x|y=K)p(y=K)} \\ &= \log p(x|y=k) - \log p(x|y=K) + \\ &\quad \log[p(y=k)/p(y=K)] \\ &= \sum_j^p [g_{k,j}(x_j) - g_{K,j}(x_j)] + \\ &\quad \log[p(y=k)/p(y=K)] \\ &= \sum_j^p f_{k,j}(x_j) + \text{const}_k \end{aligned}$$

□

In the comparisons to follow, we ensure that the same representations are possible for both procedures. In particular, for the informative case, we model the class densities with logspline densities which imply an additive spline discriminant

$$\begin{aligned} \theta_k(x) &= \sum_j^p \beta_{j,k} B_j(x) \\ P(y=k|x) &= \frac{\exp \theta_k(x)}{\sum_m \exp \theta_m(x)} \end{aligned}$$

where B is a natural cubic spline basis.

Logspline simulation study

For the simulation study shown in figure 2, the discriminant was taken to be an additive spline with 5 uniformly spaced fixed knots. Class 1 is a complicated mixture density (the outer ring), and class 2 (the two mounds in the middle) is the exponential tilt (using the logspline discriminant) of class 1. The Naive Bayes classifier assumes a logspline density (see (Stone *et al.* to appear)) separately in each dimension and in each class. Asymptotically, the GAM classifier achieves the Bayes error rate (7.2%) since the true discriminant is log additive by construction. Asymptotically, the Naive Bayes (NB) classifier does worse (9.0%) than GAM, since the class densities are not a product form. However, when only a finite sample of training observations is available, the Naive Bayes classifier does surprisingly well (this behavior has been noted by Langley

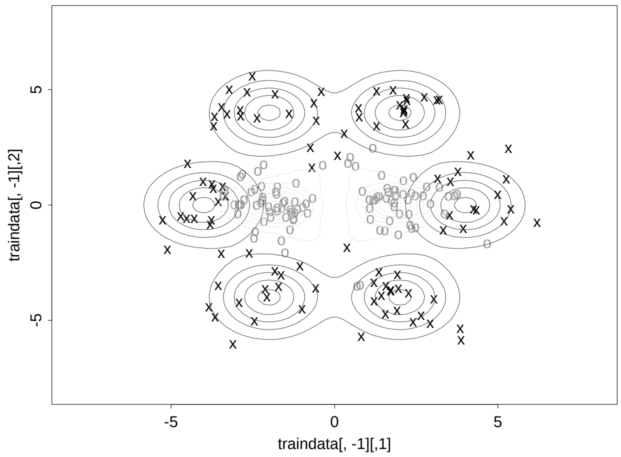

Figure 2: Contours of class densities, and training observations.

(Langley & Sage 1994)). In simulation experiments, 25 training sets each containing 25 observations from each class were used to train both NB and GAM classifiers. The average error rates were 11.1% for NB and 11.4% for GAM with standard errors of 0.05 % and 0.06% respectively. Here then is an instance where informative training actually does slightly better than discriminative training, even though the discriminative model is correct and the informative one is not!

Conclusion

Recently, Friedman (Friedman 1996) has shown that when it comes to classification, bias in the class posteriors is not so critical because of the discretization of the assignment rule. So even if the class density model is incorrect, i.e. biased, it may yet get the upper hand especially if it has lower variance estimates of the class posteriors across training sets.

It is best to use an informative approach if confidence in the model correctness is high. This suggests a promising way of combining the two approaches: partition the feature space into two. Train an informative model on those dimensions for which it seems correct, and a discriminative model on the others. Experimental results on this approach are forthcoming. We are also investigating other techniques of combining the two procedures.

Even when the goal is discrimination between classes, it pays to investigate the performance of the corresponding informative model which borrows strength from the marginal density.

References

Byth, K., and McLachlan, G. J. 1980. Logistic regression compared to normal discrimination for non-normal populations. *The Australian Journal of Statistics* 22:188–196.

Duda, R. O., and Hart., P. E. 1973. *Pattern classification and scene analysis.* Wiley.

Efron, B. 1975. The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* 70(352):892–898.

Fayyad, U. M.; Piatetsky-Shapiro, G.; and Smyth, P. 1996. From data mining to knowledge discovery: An overview. In *Advances in Knowledge Discovery and Data Mining.* The Mit Press. 1–31.

Friedman, J. 1996. On bias, variance, 0/1-loss and the curse of dimensionality. Technical report, Dept. of Statistics, Stanford University.

Hastie, T., and Tibshirani, R. 1990. *Generalized Additive Models.* Chapman Hall.

John, G. H., and Langley, P. 1995. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh Conference on Uncertainty in Artifical Intelligence.* San Mateo: Morgan Kaufmann.

Langley, P., and Sage, S. 1994. Induction of selective bayesian classifers. In *Proceedings of the Tenth Conference on Uncertainty in Artifical Intelligence.* Seattle, WA: Morgan Kaufmann.

Michie, D.; Spiegelhalter, D. J.; and Taylor, C. 1994. *Machine Learning, Neural and Statistical Classification.* Ellis Horwood.

Ney, H. 1995. On the probabilistic interpretation of neural network classifiers and discriminative training criteria. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 17(2):107–119.

O'Neill, T. J. 1980. The general distribution of the error rate of a classification procedure with application to logistic regression discrimination. *Journal of the American Statistical Association* 75(369):154–160.

Ruiz-Velasco, S. 1991. Asymptotic efficiency of logistic regression relative to linear discriminant analysis. *Biometrika* 78:235–243.

Stone, C. J.; Hansen, M.; Kooperberg, C.; and Truong, Y. K. to appear. Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics.*