Restricted Bayes Optimal Classifiers

Simon Tong
Computer Science Department
Stanford University
simon.tong@cs.stanford.edu

Daphne Koller
Computer Science Department
Stanford University
koller@cs.stanford.edu

Abstract

We introduce the notion of *restricted Bayes optimal classifiers*. These classifiers attempt to combine the flexibility of the generative approach to classification with the high accuracy associated with discriminative learning. They first create a model of the joint distribution over class labels and features. Instead of choosing the decision boundary induced directly from the model, they restrict the allowable types of decision boundaries and learn the one that minimizes the probability of misclassification relative to the estimated joint distribution. In this paper, we investigate two particular instantiations of this approach. The first uses a non-parametric density estimator — Parzen Windows with Gaussian kernels — and hyperplane decision boundaries. We show that the resulting classifier is asymptotically equivalent to a maximal margin hyperplane classifier, a highly successful discriminative classifier. We therefore provide an alternative justification for maximal margin hyperplane classifiers. The second instantiation uses a mixture of Gaussians as the estimated density; in experiments on real-world data, we show that this approach allows data with missing values to be handled in a principled manner, leading to improved performance over regular discriminative approaches.

Introduction

We introduce the notion of *restricted Bayes optimal classifiers*. These classifiers attempt to combine the flexibility of the generative approach to classification with the high accuracy associated with discriminative learning.

Classification methods can generally be separated into two groups: generative and discriminative. Discriminative learning methods such as Support Vector Machines and logistic regression directly optimize a decision boundary and tend to create classifiers possessing higher accuracy than generative methods (Michie, Spiegelhalter, & Taylor 1994; Dumais *et al.* 1998).

In contrast, the generative approach to classification first creates a model of the joint distribution of the class label and features and then classifies a future instance as belonging to the most likely class according to the model. Such a classifier is called a Bayes optimal classifier and is known to minimize the probability of misclassification relative to the estimated density. The presence of a generative model of the joint distribution permits data with missing values, data with missing labels and encoding of prior knowledge to be handled in a principled manner. However, since a model of the full joint distribution has to be constructed, these techniques often tend to have a large number of parameters that need to be estimated. Thus, the learned density is typically quite sensitive to noise in the training data, as is the associated decision boundary. In other words, the Bayes optimal classifier typically has high variance. As a consequence, particularly when doing Bayes optimal classification in high dimensional domains, one commonly uses simple density estimator (e.g., the common use of the Naive Bayes classifier in text classification (Mitchell 1997)).

We propose an alternative approach to dealing with the problem of variance in Bayes optimal classification in a spirit similar to that mentioned in (Duda & Hart 1973; Highleyman 1961). Rather than simplifying the density, we restrict the nature of the decision boundary used by our classifier. In other words, rather than using the classification hypothesis induced by the Bayes optimal classifier, we select a hypothesis from a restricted class. The hypothesis selected is the one that minimizes the probability of error relative to our learned density. We call this error the *estimated Bayes error* of the hypothesis. For example, we can restrict to hypotheses defined by hyperplane decision boundaries. We call the hyperplane that minimizes the estimated Bayes error with respect to a given density a *Bayes optimal hyperplane*.

In this paper, we investigate two particular instantiations of this approach. The first part of the paper discusses using a non-parametric density estimator — Parzen Windows with Gaussian kernels — and hyperplane decision boundaries. We prove that the resulting classifier converges asymptotically to a maximal margin hyperplane classifiers. Maximal margin classifiers are one of the most successful discriminative classifiers, having strong theoretical justifications and empirical successes. By relating them to Bayes optimal classifiers we provide an alternative justification for maximal margin hyperplane classifiers. We also extent our results to a wider class of decision boundaries.

The second part of this paper considers using a semi-parametric density estimator — a mixture of Gaussians — and hyperplane boundaries. We perform experiments on real world data sets. Here, the benefits of having a model of the joint distribution become apparent, allowing data with missing values to be handled in a principled manner leading to improved performance over regular discriminative approaches.

General Framework

Our focus in this paper is the task of classifying real-valued data cases into two classes. More precisely, suppose we have a feature space $\mathcal{X} = \mathbb{R}^d$ and training data $\mathcal{D} = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots (\mathbf{x}_n, y_n)\}$, where $y_i \in \{C_0, C_1\}$ is called the *class label* for instance $\mathbf{x}_i$. Let n_0 and n_1 be the (non-zero) number of training data in classes C_0 and C_1 respectively. We write $\mathbf{x}_j \in C_i$ when $y_j = C_i$.

Definition 1 *A* classifier *h is a mapping from $\mathcal{X} - N$ to the set $\{C_0, C_1\}$, where N is a null set. Let $H_0 = \{\mathbf{x} : h(\mathbf{x}) = C_0\}$ and $H_1 = \{\mathbf{x} : h(\mathbf{x}) = C_1\}$.*

In other words, h is a classifier if it maps (almost all) points in $\mathcal{X}$ to one of two class labels.

Let our data instances and their labels be independent and identically distributed according to a joint distribution $P(\mathbf{x}, C)$. We can define the probability that a classifier h makes a classification error:

Definition 2 *Given a joint distributions $P(\mathbf{x}, C)$ and a classifier h we define the* expected misclassification rate *or* Bayes error *of h relative to P, error$(h : P)$, as:*

$$E_P[L(h(\mathbf{x}), C)], \text{ where } L(h(\mathbf{x}), C) = \begin{cases} 1 \text{ if } h(\mathbf{x}) \neq C \\ 0 \text{ otherwise} \end{cases}$$

The *Bayes optimal classifier relative to a distribution P* is given by: $h^*(\mathbf{x}) = C_0$ whenever $P(C_0 \mid \mathbf{x}) > P(C_1 \mid \mathbf{x})$, and $h^*(\mathbf{x}) = C_1$ otherwise. It is well-known that it minimizes the Bayes error.

In order to use the Bayes optimal classifier, we need $P(C_0 \mid \mathbf{x})$ and $P(C_1 \mid \mathbf{x})$. In general, these quantities are not known. The generative approach to classification uses the training data to estimate an approximate joint distribution $\hat{P}(\mathbf{x}, C)$, and then uses the Bayes optimal classifier relative to $\hat{P}$. Let the *estimated Bayes error* denote the Bayes error relative to an estimated distribution $\hat{P}$. The Bayes optimal classifier relative to $\hat{P}$ minimizes the estimated Bayes error.

The Bayes optimal classifier often induces decision boundaries that are fairly complex. Furthermore, the estimate of $\hat{P}$ is often quite sensitive to the noise in the training data, which often implies a similar sensitivity for the decision boundary. In other words, the variance of the Bayes optimal classifier is quite large. A possible approach for reducing this variance is to restrict the class of hypotheses that we allow ourselves to consider. That is, we select the "best" hypothesis within some restricted class $\mathcal{H}$.

Definition 3 *Given a joint distribution $\hat{P}(\mathbf{x}, C)$ and a set of classifiers $\mathcal{H}$, we say that h^* is a* restricted Bayes optimal classifier with respect to $\mathcal{H}$ and $\hat{P}$ *if $h^* \in \mathcal{H}$ and for all $h \in \mathcal{H}$, error$(h^* : \hat{P}) \leq$ error$(h : \hat{P})$.*

One restricted set of classifiers that has received a lot of attention is the set of hyperplane classifiers, where the decision boundary is a hyperplane in feature space.

The above definitions hold in a very general setting. In order to apply them, we need to choose a concrete approach to estimating $\hat{P}$. In most cases, it is easier to estimate $\hat{P}$ using the decomposition $\hat{P}(\mathbf{x}, C) = \hat{P}(C) \cdot \hat{p}(\mathbf{x} \mid C)$ where $\hat{p}(\mathbf{x} \mid C)$ is the *class-conditional density* of the feature vectors $\mathbf{x}$ within the class C. There are many techniques for estimating the class conditional densities (Bishop 1995; Fukanaga 1990; Silverman 1986). We will consider two types of density estimators in this paper: Parzen Windows, and mixtures of k Gaussians.

Parzen Windows and Maximal Margin Hyperplanes

In this section we choose a standard method to estimate the joint density that uses the above decomposition. First, we take the maximum likelihood estimates for $P(C_0)$ and $P(C_1)$. We choose a simple variant of *non-parametric* density estimation: Parzen Windows estimation with Gaussian kernels. To estimate $p(\mathbf{x} \mid C_i)$, we place a Gaussian kernel over each training instance $\mathbf{x}_j$ in class C_i; the estimated density is simply the average of these kernels. We use identical Gaussian kernels for all data cases, each with a diagonal covariance matrix $\sigma^2 I$. More precisely, we define for $i = 0, 1$

$$p_\sigma(\mathbf{x} \mid C_i) = \frac{1}{n_i} \sum_{\mathbf{x}_j \in C_i} \frac{1}{\sigma^d (2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{x}_j)^T(\mathbf{x}-\mathbf{x}_j)}$$

where n_i is the number of training instances in class C_i and σ is called the *smoothing parameter*. Together, $\hat{P}(C_0)$, $\hat{P}(C_1)$ and $p_\sigma(\mathbf{x} \mid C_i)$ define a joint density $P_\sigma(\mathbf{x}, C)$ as required. We use *error*$_\sigma(h)$ to denote *error*$(h : P_\sigma)$.

Different values for σ correspond to different choices along the bias-variance spectrum: smaller values (sharper peaks for the kernels) correspond to higher variance but lower bias estimates of the density. The choice of σ is often crucial for the accuracy of the Bayes optimal classifier. We can eliminate the bias induced by the smoothing effect of σ by making it arbitrarily close to zero. We prevent the variance of the classifier from growing unboundedly by restricting our hypotheses to the very limited class of hyperplanes. Thus, we choose as our hypothesis the Bayes optimal hyperplane relative to the estimated density induced by the data and σ.

In this section our main result is the following: for linearly separable data, as σ tends to zero, the Bayes optimal hyperplane converges to the maximal margin hyperplane. We further show that a similar result holds for a much wider class of classifiers: for small enough σ, the classifier that maximizes the margin will have a lower estimated error than a classifier with a smaller margin. We also show that for linearly non-separable data, the Bayes optimal hyperplane has a very natural interpretation: it minimizes the training set classification error, and among all the hyperplanes that have the same classification error, it is the one with the largest margin.

Linearly separable data

In this subsection, we assume that the training data are linearly separable. In other words, there exists at least one hyperplane classifier that will correctly classify all of the training data. We will also restrict the hypothesis space $\mathcal{H}$ to be the set of hyperplane classifiers that correctly classify all training data. (These restrictions will be relaxed later on.)

In this case, we can show a tight connection between Bayes optimal hyperplanes and *maximal margin classifiers* (Vapnik 1982).

Definition 4 *The* margin *of a hyperplane h, denoted by margin(h), is the smallest Euclidean distance from the hyperplane to a training instance.*

Theorem 5 *Let h^* be some hyperplane in $\mathcal{H}$. The following statements are equivalent:*

- *$\forall h \in \mathcal{H}$ s.t. $h \neq h^*$, $\exists S > 0$ s.t. error$_\sigma(h^*) <$ error$_\sigma(h)$ whenever $\sigma < S$.*
- *h^* has maximal margin.*

The intuition behind this result is based on the following alternative expression for the estimated Bayes error, *error*$_\sigma(h)$:

$$\hat{P}(C_1)\int_{\mathbf{x}\in H_0} p_\sigma(\mathbf{x} \mid C_1)\,d\mathbf{x} + \hat{P}(C_0)\int_{\mathbf{x}\in H_1} p_\sigma(\mathbf{x} \mid C_0)\,d\mathbf{x}.$$

Points that are closer, in Euclidean distance, to one of the Gaussian kernels in $p_\sigma(\mathbf{x} \mid C_1)$ have significantly higher density. Thus, the closer we move the decision boundary to the centers of these kernels, the more mass it will contribute to the estimate Bayes error. As σ shrinks, the kernels that are closest to the decision boundary dominate more and more. A careful analysis shows that the estimated Bayes error of a hyperplane h is dominated by an expression which is exponential in $-\textit{margin}(h)^2/(2\sigma^2)$. Thus, as σ tends to zero the hyperplane with the larger margin will dominate (have lower error relative to other hyperplanes).

This theorem shows that for any other hyperplane h, once σ is small enough, h^* beats h. However, this does not suffice to show that, as we reduce σ, the Bayes optimal hyperplane h^*_σ for P_σ "converges" to the maximal margin hyperplane h^*. It could, perhaps, be the case that for each $\sigma > 0$, h^*_σ is arbitrarily far away from h^*. In fact, a similar proof to that of Theorem 5 shows that this is not the case.

Corollary 6 *Let $h^* \in \mathcal{H}$ be the maximal margin hyperplane. Let $\delta > 0$ and $\mathcal{H}_\delta$ be the set of hyperplanes in $\mathcal{H}$ with margins less than margin$(h^*) - \delta$. Then there exists $S > 0$ such that, for all $\sigma < S$ and all $h \in \mathcal{H}_\delta$, error$_\sigma(h^*) <$ error$_\sigma(h)$.*

Thus, as $\sigma \to 0$, the margin of h^*_σ tends to the maximal margin. In other words, the Bayes optimal hyperplane converges to the maximal margin hyperplane in terms of margin.

General classifiers

We can generalize the previous framework to more complex classes of classifiers. Instead of letting $\mathcal{H}$ be the set of hyperplanes classifiers we will let $\mathcal{H}$ be any set of classifiers obeying the following condition:

Condition 7 *$\mathcal{H}$ is a set of classifiers such that:*

- *For each $h \in \mathcal{H}$, $H_0 = \{\mathbf{x} : h(\mathbf{x}) = C_0\}$ and $H_1 = \{\mathbf{x} : h(\mathbf{x}) = C_1\}$ are both open sets.*

This condition is fairly mild and allows for a large range of classifiers. For example, any set of classifiers which have hyperplane or polynomial decision boundaries, hinged hyperplanes or decision boundaries induced by neural networks with sigmoidal or linear activation functions satisfy the condition.

We can also generalize the notion of a margin to hold for these more general forms of classifiers. Intuitively the margin is the smallest distance between a training instance and the decision boundary.

Definition 8 *Let h be a classifier. Then margin(h) is:*

$$\min\left(\min_{\mathbf{x}_j\in C_1}\big(\inf_{\mathbf{x}\in H_0}\|\mathbf{x}-\mathbf{x}_j\|\big), \min_{\mathbf{x}_j\in C_0}\big(\inf_{\mathbf{x}\in H_1}\|\mathbf{x}-\mathbf{x}_j\|\big)\right).$$

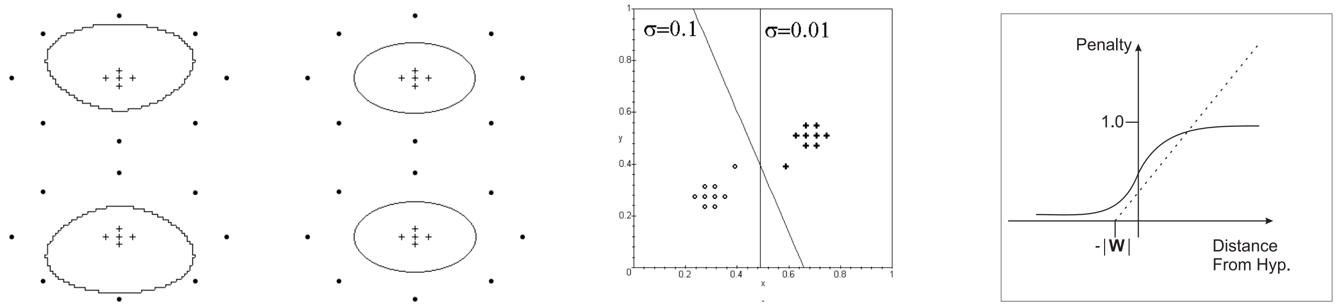
In the case of hyperplanes, this definition coincides with the original definition of the margin of a hyperplane classifier. We can now prove the following theorem.

Theorem 9 *Let $\mathcal{H}$ be a set of classifiers that satisfy condition 7. Let h^* be some classifier in $\mathcal{H}$. The following statements are equivalent:*

- *$\forall h \in \mathcal{H}$ s.t. margin$(h) \neq$ margin(h^*), $\exists S > 0$ s.t. error$_\sigma(h^*) <$ error$_\sigma(h)$ whenever $\sigma < S$.*
- *h^* has maximal margin.*

This result says that a classifier that has maximal margin will eventually have a lower estimated error than any other particular classifier with a smaller margin. The proof of the theorem (omitted) is rather more involved that that of theorem 5 but the intuition behind it is similar. Suppose we are given two classifiers one of which has a larger margin than the other. It is possible to find a region in $\mathcal{X}$ that is classified one way by the larger margin classifier but the opposite way by the smaller margin classifier and for which the estimated Bayes error incurred by wrongly classifying that area dominates as we reduce the smoothing parameter.

It is interesting to compare this instantiation of restricted Bayes optimal classification with Support Vector Machines (SVMs) (Vapnik 1982; Cortes & Vapnik 1995). An SVM finds the hyperplane that maximizes the margin. The problem of maximizing the margin can be cast as a convex optimization problem then only depends on the training data via inner products between training instances (e.g., $\mathbf{x}_i \cdot \mathbf{x}_j$). One can then apply the "kernel trick" (Cortes & Vapnik 1995), where we replace the inner products with a "kernel function" $K(\mathbf{x}_i, \mathbf{x}_j)$ that satisfies Mercer's condition. (These "kernel functions" are not to be confused with the Gaussian kernels used in Parzen Windows.) Since K satisfies Mercer's condition, we can write $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ and so by using K we are then implicitly projecting the training data into a different (often higher dimensional) feature space $\mathcal{F}$ and finding the hyperplane that maximizes the margin in that space. By choosing different kernel functions we can implicitly project the training data from $\mathcal{X}$ into spaces $\mathcal{F}$ for which hyperplanes in $\mathcal{F}$ correspond to more complex decision boundaries in the original space $\mathcal{X}$. One commonly



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

Figure 1: (a) SVM using degree 4 polynomial kernel. (b) Degree 4 polynomial restricted Bayes classifier using Parzen Windows and small σ. (c) Bayes optimal hyperplane for different σ. (d) Parzen Bayes (solid) and Soft Margin (dotted) error functions.

used kernel is $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^p$ which induces polynomial decision boundaries of degree p in the original space $\mathcal{X}$.

How does this relate to using restricted Bayes optimal classification with non-parametric density estimation? One fact is that using an SVM that maximizes the margin and that uses a kernel function $K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ corresponds to performing Parzen Window density estimation in the projected space $\mathcal{F}$ and finding the Bayes optimal hyperplane in that projected space for $\sigma \to 0$.

A second fact is that even though the SVM maximizes the margin in the higher dimensional space, the classifier it produces does not necessarily maximize the margin in the original space. So, if we decided to use the kernel $K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i \cdot \mathbf{x}_j + 1)^4$ — which induces degree four polynomial boundaries in the original space — the SVM will not correspond to the degree four polynomial with the largest margin in the original space. However, if we were to use Parzen Windows in our original space and restrict our hypotheses to degree four polynomials then the maximal margin degree four polynomial will have a lower estimated Bayes error than any other polynomial for sufficiently small σ. Figure 1(a) shows the decision boundary produced when using a SVM with a degree four polynomial kernel and figure 1(b) shows the decision boundary produced when using a restricted Bayes optimal classifier where we restrict the set of classifiers to the set of polynomials with degree at most four. The decision boundaries produced by the two approaches do not necessarily coincide.

Data that are not linearly separable

We return to considering the set of hyperplane classifiers and now analyze the behavior of the Parzen Windows Bayes optimal hyperplane for the more general case of data that are not necessarily linearly separable.

Corollary 10 *Given $\sigma > 0$ and hyperplane $h \in \mathcal{H}$ we can write the estimated Bayes error in the following form:*

$$\mathit{error}_\sigma(h) = \frac{1}{n}\left(f_\sigma(\mathcal{D}, h) + \#\mathit{incorrect}\right)$$

where #incorrect is the number of misclassifications on the training data $\mathcal{D}$, and $f_\sigma(\mathcal{D}, h)$ has the following properties:

- $f_\sigma(\mathcal{D}, h) \to 0$ *as* $\sigma \to 0$
- *if h^* and h both minimize #incorrect and h^* has larger margin than h then there exists $S > 0$ such that $f_\sigma(\mathcal{D}, h^*) < f_\sigma(\mathcal{D}, h)$ whenever $\sigma < S$.*

In other words, as σ tends to zero, the hyperplane with the lowest score will have the lowest classification error on the training data; and, of all such minimum error hyperplanes, it will have the greatest margin with respect to the correctly classified data. Intuitively, this seems a very reasonable hyperplane to pick. This corollary is also consistent with the results we obtained for the linearly separable case; in this case there are hyperplanes in which the number of misclassifications are zero. So, for sufficiently small smoothing parameters the Bayes optimal hyperplane will be a hyperplane which correctly classifies all training data. Hence, for the previous theorems for the linearly separable case we can remove the restriction of the hypothesis space to hyperplanes that correctly classify all training data — we know that for small enough smoothing parameters the Bayes optimal hyperplane will always correctly classify the training data.

The effect of nonzero σ

Until now, we have focused on the behavior of the Bayes optimal hyperplane as σ gets arbitrarily close to zero. As we discussed in the informal justification for Theorem 5, as σ shrinks, the data points closer to the decision boundary have larger and larger impact. At the limit, only the points closest to the boundary, i.e., the ones on the margin, have impact. It is not clear whether focusing only on the margin is necessarily the optimal approach. If we take σ to be non-zero, classifying using the estimated Bayes error will consider the distances of other points from the margin. The larger we make σ, the larger the effect that points further from the margin have on the estimated Bayes error and on the choice of hyperplane. Figure 1(c) illustrates one simple example where a larger value of σ leads to a hyperplane which is arguably more reasonable. In the non-separable case, we also get a similar tradeoff. It is interesting to note that the form of the error function in Corollary 10 resembles the form of the *soft margin* error function often used in SVMs to cope with linearly non-separable data (Cortes & Vapnik 1995): $\|\mathbf{w}\|^2/2 + C(\sum_i \xi_i)$. Briefly, $\mathbf{w}$ is the hyperplane weight vector (including the bias weight b), C is a tunable parameter that influences how much of a penalty to assign errors and the ξ_is are *slack variables* where $(\sum_i \xi_i)$ provides an upper bound on the number of training errors. Thus the soft error function can be decomposed into the sum of a deterministic function of the margin and a "softer" function of the training error whereas our error function can be decomposed into a "soft" function of the margin and the exact training error.

We investigated whether using a non-zero value of σ would achieve a similar effect to that of the soft margin error function.[1] We used the "Pima Indian Diabetes" UC Irvine data set (Blake, Keogh, & Merz 1998) and a synthetic data set. The Pima data set has eight features, with 576 training instances of which 198 are labeled as positive. The synthetic data were generated from two dimensional Gaussian class conditional distributions. In the synthetic case the underlying distribution was known and we could compute the true Bayes error for each of the hypotheses produced, while in the Pima data set classification error on a separate 192 instance test set was measured.

For various values of σ, we searched for the hyperplane that minimizes the estimated Bayes error for P_σ. The error is a differentiable function of the weights of the hyperplane, so we can use gradient descent techniques to find the Bayes optimal hyperplane for any given smoothing parameter. The update rules are easily derived and are omitted.

There are some practical issues to deal with in the implementation of this idea. Unfortunately, the search space is not convex and local minima exist. Furthermore, for small σ, the space consists of numerous very large gently sloping plateaus. Thus, naive gradient descent converges to suboptimal solutions and very slowly. Corollary 6 suggests seeding the search with the maximal margin hyperplane wherever possible and this seemed to improve the speed of convergence and quality of results. We also used *bold driving* (Bishop 1995) to speed up the convergence.

Table 1 lists the errors for various settings of σ and C. For the synthetic data the error is the average true Bayes error over ten data sets, where the optimal parameter was chosen for each data set separately. With the larger data sets we experimented with using cross validation for setting the σ and C parameters and these results are also listed. Many of the synthetic data sets of size nine and fifteen were actually linearly separable and for each of those the maximum margin hyperplane was also computed. These results indicate that the Parzen Windows Bayes optimal hyperplane, optimizing its somewhat different but arguably more natural error function, achieves very similar performance to that of a soft margin hyperplane. This observation is supported by a paired t-test on the cross-validation folds of the Pima data (at 5% significance). Furthermore, the synthetic data support our intuition above that, even for linearly separable data, the Bayes optimal hyperplane can produce a more appropriate classifier than the maximal margin hyperplane, and so reducing the smoothing parameter to zero is not always optimal.

[1]We used T. Joachim's SVMlight: www-ai.informatik.uni-dortmund.de/thorsten/svm_light.html

Table 1: Comparison with Soft Margin and Maximal Margin

Parzen σ	Pima Error	Soft Margin C	Pima Error
0.008	28.1	0.1	24.4
0.01	18.8	0.3	21.3
0.02	21.4	0.9	21.9
0.03	22.9	5	22.3
0.04	23.4	10	22.3
0.05	**22.3**	100	22.3
0.1	22.3	**1000**	**22.3**
0.2	27.1	10000	22.3

* Underlined rows indicate settings picked by cross validation.

Synthetic size	9	15	30	5-fold CV 30
Parzen Hyp	**8.1 ± 0.35**	8.0 ± 0.23	**7.6 ± 0.11**	**8.5 ± 0.62**
Soft Margin	8.7 ± 0.73	**7.9 ± 0.25**	**7.6 ± 0.08**	8.7 ± 0.55
Max Margin	9.8 ± 0.47	9.4 ± 0.61	—	—

Table 2: Resistance to Outliers. Classification error.

Outlier Percentage	Parzen Hyperplane	Soft Margin Hyperplane	Logistic Regression
0%	**16.5 ± 0.2**	16.6 ± 0.2	16.8 ± 0.2
1%	**16.7 ± 0.2**	25.1 ± 0.4	24.7 ± 0.2
2%	**16.7 ± 0.4**	24.0 ± 0.2	24.0 ± 0.2
3%	**17.3 ± 0.8**	25.0 ± 0.2	25.1 ± 0.1
4%	**18.9 ± 1.5**	24.7 ± 0.3	24.4 ± 0.2
5%	**17.4 ± 0.9**	24.9 ± 0.3	24.7 ± 0.2

With the restricted Bayes hyperplane, the error incurred from a misclassified training instance is eventually saturated the further the instance is from the hyperplane. In contrast, the soft margin error function penalizes a misclassified training instance proportionally to the instance's distance from the margin (see figure 1(d)). This indicates that Parzen Windows Bayes hyperplanes may be more resistant to outliers than Soft Margin SVMs. To test this hypothesis we performed a simple experiment. We sample data from two Gaussians. Each training set consisted of 100 instances and outliers were added to the training sets in various proportions. Outliers were approximately an order of magnitude distance away from the other data points. Again, cross validation was used to select σ and C. Table 2 presents the generalization performance with each row being an average over ten independent runs. The table indicates that Parzen Windows hyperplanes with Gaussian kernels tend to be more resistance to outliers than the other linear methods.

Mixtures of Gaussians

Until now, we have considered using non-parametric density estimation with Gaussian kernels as our density estimator. Clearly we can use other densities with the restricted Bayes optimal classification approach. We now consider using a more parametric density estimator that will allow us to take greater advantage of having a model of the joint distribution. The mixture of k Gaussians density estimator assumes that the class i conditional density $p(\mathbf{x} \mid C_i)$ is a mixture of k Gaussian densities. More precisely, we define for $i = 0, 1$

$$p(\mathbf{x} \mid C_i) = \sum_{j=1}^{k} m_{ij} \left(\frac{1}{\sigma_i^d (2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2\sigma_i^2}(\mathbf{x}-\mu_{ij})^T(\mathbf{x}-\mu_{ij})} \right)$$

Table 3: Average test set error using complete data.

Data Set	Linear SVM	Logistic	MoG Hyp	MoG
Breast	28.74 ± 0.43	$\mathbf{27.38 \pm 0.47}$	27.42 ± 0.50	29.16 ± 0.53
Diabetes	23.43 ± 0.17	23.37 ± 0.18	$\mathbf{23.23 \pm 0.17}$	26.50 ± 0.21
German	24.12 ± 0.23	$\mathbf{23.94 \pm 0.21}$	24.03 ± 0.24	26.35 ± 0.27
Heart	$\mathbf{16.00 \pm 0.33}$	16.97 ± 0.28	16.33 ± 0.33	17.78 ± 0.37
Hepatitis	32.53 ± 0.59	31.21 ± 0.51	$\mathbf{27.19 \pm 0.34}$	32.98 ± 0.45
Ionosphere	13.44 ± 0.22	$\mathbf{13.16 \pm 0.23}$	13.27 ± 0.23	10.55 ± 0.31
Sonar	$\mathbf{25.06 \pm 0.42}$	25.07 ± 0.41	27.62 ± 0.38	29.95 ± 0.46
Waveform	$\mathbf{12.85 \pm 0.05}$	13.44 ± 0.07	12.91 ± 0.06	10.65 ± 0.04

Table 4: Average test set error with missing data values.

Data Set	Linear SVM	Logistic	MoG Hyp	MoG
Breast	$\mathbf{29.74 \pm 0.49}$	30.91 ± 0.50	$\mathbf{29.74 \pm 0.47}$	32.22 ± 0.58
Diabetes	$\mathbf{26.09 \pm 0.23}$	26.22 ± 0.27	26.93 ± 0.24	31.50 ± 0.44
German	30.09 ± 0.37	29.46 ± 0.32	$\mathbf{28.38 \pm 0.27}$	30.91 ± 0.46
Heart	18.21 ± 0.42	18.66 ± 0.43	$\mathbf{17.94 \pm 0.40}$	20.10 ± 0.46
Hepatitis	28.63 ± 0.40	28.26 ± 0.42	$\mathbf{27.81 \pm 0.48}$	30.47 ± 0.61
Ionosphere	$\mathbf{13.38 \pm 0.21}$	13.73 ± 0.22	14.96 ± 0.30	15.46 ± 0.27
Sonar	33.15 ± 0.55	$\mathbf{31.70 \pm 0.51}$	32.95 ± 0.46	35.42 ± 0.48
Waveform	14.80 ± 0.08	15.89 ± 0.08	$\mathbf{14.59 \pm 0.10}$	13.88 ± 0.21

Here each m_{ij} is a mixture weight that determines how much the j-th Gaussians contributes towards the overall class i conditional density. Mixtures of Gaussians are semi-parametric density estimators. Notice here that the number of mixture components k is fixed and typically a small value. This is in contrast to the Parzen Windows non-parametric density estimator where the number of kernels grows with the number of training instances.

We can estimate the parameters $m_{ij}, \mu_{ij}, \sigma_i$ for $i \in \{0, 1\}, j \in \{1, \ldots k\}$ by using the Expectation Maximization (EM) algorithm (Dempster, Laird, & Rubin 1977). EM finds parameters that locally maximize the likelihood of the observed data. As before, we use the maximum likelihood estimates for the class priors $P(C_0)$ and $P(C_1)$.

We can also handle the presence of missing values in the training data in a principled way. We now use the E-step of EM to not only to compute the mixture components m_{ij}, but also the expected missing values of the data.

Given a density of the above form we can compute the Bayes optimal hyperplane using a similar gradient decent technique as in the Parzen Windows density estimation case.

Experiments

For our experiments we compared three different hyperplane classifier methods: Linear SVMs, logistic regression and Bayes optimal hyperplanes using mixtures of Gaussians (MoG Hyp). We also looked at the Bayes optimal classifier derived from using the mixture of Gaussians density directly (MoG). For the SVM the soft margin parameter, C, needed to be tuned. For the Bayes hyperplane and for the mixture of Gaussians Bayes optimal classifier the number of class mixture components, k, needed to be chosen.

We used data sets from the UC Irvine repository (Blake, Keogh, & Merz 1998). [2] We created 100 randomly generate train/test splits of the data (in a roughly 40/60 split). Each data set contained no missing values. For each realization of the data we learned a SVM, a logistic hyperplane, a Bayes optimal hyperplane using mixtures of Gaussians, and finally the Bayes optimal classifier for a mixture of Gaussians. We used five-fold cross validation on the first five realizations to choose the parameters for each of the methods.

Table 3 contains the test set error rates for the data sets averaged over the one hundred runs. Bold face figures indicate the best *hyperplane* method for each data set. The Bayes optimal hyperplane using mixtures of Gaussians actually outperforms the mixture of Gaussians Bayes optimal classifier on six of the eight data sets. This indicates that restricting the nature of the decision boundary can be better than using the density estimator directly. The Bayes optimal hyperplane is competitive with the two discriminative methods — it is the best hyperplane method on two out of the eight sets, having a lower error rate than the SVM on five sets and outperforming logistic regression on four.

We then looked at how the methods performed with data that contained missing values. For each training instance we randomly removed a feature value with probability 0.75.[3] We used EM to perform density estimation with mixtures of two Gaussians. However, the regular SVM and logistic methods do not handle missing values. For these two discriminative methods we used the common technique (Bishop 1995) of filling in the missing values with their class averages.

Table 4 contains the error rates for the data sets. Here the Bayes hyperplane performs somewhat better, being the first (or equal first) best hyperplane method for five of the eight sets while logistic regression is the best hyperplane method for only one of the data sets and SVMs are the best (or equal best) for three sets. Again, the Bayes hyperplane outperforms the plain mixture of Gaussians Bayes optimal classifier on most of the data sets (seven out of the eight). It is better or equal to the linear SVM on six out of the eight and outperforms logistic regression on five out of eight sets.

As they stand, the gains from the mixture of Gaussian hyperplanes are only suggestive rather than overwhelming. Note that we used a particularly naive form of the Mixture of Gaussian estimator — every Gaussian component within a class had to have an identical and restricted form of covariance matrix. It could well be that allowing more flexible covariance matrices would lead to further improvements in performance.

Conclusions and future work

We have introduced an alternative approach for dealing with the high variance of the Bayes optimal classifier in high dimensional spaces. Our approach is based on finding simple hypotheses that minimize the estimated Bayes error within a certain class, where the Bayes error is estimated relative to the learned distribution.

[2] The UC Irvine breast cancer data was obtained from M. Zwitter and M. Soklic at the University Medical Centre, Inst. of Oncology, Ljubljana, Yugoslavia.

[3] The hepatitis data was an exception. We only removed a value with probability 0.4 since otherwise some features consisted entirely of missing values over the whole training set.

We have shown that one very natural instantiation of our approach, where we use Parzen Windows with Gaussian kernels, converges at the limit to the maximal margin hyperplane classifier. We have further analyzed the behavior of the Parzen Windows restricted Bayes method when we consider more general forms of classifiers. While possessing desirable properties, Gaussian kernels tend to cause numerous local minima making a practical system hard to implement well. We are currently investigating choosing different kernels that reduce the jaggedness of the search space while still retaining many of the properties of Gaussian kernels.

The Parzen Windows hyperplane result has several implications. From one perspective, it can be viewed as providing a new probabilistic justification for maximal margin hyperplanes. There have been several other studies exploring probabilistic interpretations, although mainly in the context of Bayesian learning (Cristianini, Shawe-Taylor, & Sykacek 1998; Sollich 1999; Herbrich, Graepel, & Campbell 1998). From another perspective, it provides a strong justification for our intuition that the restricted Bayes optimal classifier avoids the high variance problem of the unrestricted Bayes optimal classifier, even when the representation of the density is very complex. We considered an extremely high variance representation of a density — a non-parametric density with arbitrarily low kernel width. However, the Bayes optimal hyperplane relative to this distribution is (close to) the maximum margin hyperplane, which is known to work well even in high-dimensional spaces. Furthermore, our result suggests that finding a simple classifier optimal relative to a complex density can be better than finding the unrestricted Bayes optimal classifier relative to a simpler density: the maximal margin classifier is better in many domains than most Bayes optimal classifiers.

These observations raise the obvious question as to whether it was the specific choice of non-parametric density estimation and Gaussian kernels that led to the success of restricted Bayes optimal classifier. We addressed this question to some degree by investigating the use of mixture of Gaussians as the density estimator. Our experiments strongly suggest that restricted Bayes optimal classifiers can be used in conjunction with other forms of density estimation to obtain competitive classification performance on complete data. The experiments also suggest that restricted Bayes optimal classifiers can take advantage of having a model of the joint distribution to give an edge over discriminative methods when dealing with data sets with missing values.

This last observation suggests a new perspective on the debate between discriminative learning and the generative approach for classification (Duda & Hart 1973; Rubenstein & Hastie 1997; Jaakola & Haussler 1998; Jaakkola, Meila, & Jebara 1999). In many domains, discriminative learning empirically achieves higher classification accuracy than the Bayes optimal classifier. The usual explanation is that the generative approach spends too much "effort" on minimizing "irrelevant" errors in $P(\mathbf{x}, C)$, and not enough on reducing classification errors. Our approach provides an alternative solution, where the estimated joint density is not used directly in the form of the Bayes optimal decision boundary, but rather to evaluate classifiers in a restricted class.

We intend to investigate the benefits of restricted Bayes optimal classifiers further. Having a model of the joint distribution can provide other advantages. It facilitates the encoding of prior knowledge in a principled way and EM could be used to incorporate unlabeled data. We plan to experiment with this approach for a variety of density estimation approaches. We hope that it will allow us to combine the benefits of the generative approach using realistically expressive representations with the high accuracy classification often associated with discriminative learning.

Acknowledgements

This work was supported by DARPA's *Information Assurance* program under subcontract to SRI International, and by ARO grant DAAH04-96-1-0341 under the MURI program "Integrated Approach to Intelligent Systems".

References

Bishop, C. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.

Blake, C. Keogh, E. and Merz, C. 1998. UCI repository of machine learning databases.

Cortes, C., and Vapnik, V. 1995. Support vector networks. In *Machine Learning*, volume 20.

Cristianini, N. Shawe-Taylor, J. and Sykacek, P. 1998. Bayesian classifiers are large margin hyperplanes in a Hilbert space. In *Proc. NeuroCOLT2*.

Dempster, A. Laird, N. and Rubin, D. 1977. Maximum likelihood from incomplete data via the EM algorithm. In *Journal of the Royal Statistical Society*.

Duda, R., and Hart, P. 1973. *Pattern Classification and Scene Analysis*. Wiley, New York.

Dumais, S. Platt, J. Heckerman, D. and Sahami, M. 1998. Inductive learning algorithms and representations for text categorization. In *Proc. 7th International Conference on Information and Knowledge Management*.

Fukanaga, K. 1990. *Introduction to Statistical Pattern Recognition*. Boston: Academic Press.

Herbrich, R. Graepel, T. and Campbell, C. 1998. Bayesian learning in reproducing kernel Hilbert spaces. Technical Report TR 99-11, Technical University of Berlin.

Highleyman, W. 1961. Linear decision functions, with application to pattern recognition. In *Proc. IRE*, volume 49, 31–48.

Jaakkola, T. Meila, M. and Jebara, T. 1999. Maximum entropy discrimination. Technical Report AITR-1668, MIT.

Jaakola, T. S., and Haussler, D. 1998. Exploiting generative models in discriminative classifiers. In *Ten Conf. on Advances in Neural Info. Processing Systems (NIPS)*.

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

Mitchell, T. 1997. *Machine Learning*. McGraw-Hill.

Rubenstein, Y., and Hastie, T. 1997. Discriminative vs informative learning. In *Proc. AAAI*.

Silverman, B. 1986. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Sollich, P. 1999. Probabilistic interpretation and Bayesian methods for Support Vector Machines. In *Proceedings of ICANN 99*.

Vapnik, V. 1982. *Estimation of Dependences Based on Empirical Data*. Springer Verlag.