Bayes Optimal Hyperplanes $\rightarrow$ Maximal Margin Hyperplanes

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

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

Abstract

Maximal margin classifiers are a core technology in modern machine learning. They have strong theoretical justifications and have shown empirical successes. We provide an alternative justification for maximal margin hyperplane classifiers by relating them to Bayes optimal classifiers that use Parzen windows estimations with Gaussian kernels. For any value of the smoothing parameter (the width of the Gaussian kernels), the Bayes optimal classifier defines a density over the space of instances. We define the *Bayes optimal hyperplane* to be the hyperplane decision boundary that gives lowest probability of classification error relative to this density. We show that, for linearly separable data, as we reduce the smoothing parameter to zero, the Bayes optimal hyperplane converges to the maximal margin hyperplane. We also analyze the behavior of the Bayes optimal hyperplane for linearly non-separable data, showing that it satisfies a very natural optima. We explore the idea of using the hyperplane that is optimal relative to a density with some small non-zero kernel width and compare the resulting hyperplane with the maximal margin and the soft margin hyperplanes.

1 Introduction

Maximal margin classifiers are a core technology in modern machine learning. They have strong theoretical justifications and have shown empirical successes. We provide an alternative justification for maximal margin hyperplane classifiers by relating them to Bayes optimal classifiers.[1]

Bayes optimal classifiers perform classification by estimating the class priors and class conditional densities and then classifying a sample as belonging to the most likely class according to the estimated densities. The Bayes optimal classifier is known to minimize the probability of misclassification relative to the estimated density. Most density representations tend to have a large number of parameters 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, when doing Bayes optimal classification in high-dimensional domains, one rarely uses anything but the simplest density estimator (e.g., the common use of the Naive Bayes classifier in text classification [10]).

We propose an alternative approach to dealing with the problem of variance in Bayes optimal classification in a spirit similar to that mentioned in [6; 9]. 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. As we mentioned, the Bayes optimal classifier minimizes this error among all possible hypothesis; we choose the hypothesis that minimizes it within the restricted class. 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 one particular instantiation of this approach, and show that it is equivalent to choosing a maximal margin hyperplane. Consider the problem of classifying vectors in $\mathbb{R}^d$ into two classes C_0 and C_1. We estimate the class conditional densities using Parzen windows estimation with Gaussian kernels. For a given value σ, the density for each class C_i is defined as a mixture of Gaussian kernels of width σ, centered on the data points in class C_i. 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. For a finite number of training instances, 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.

Our main result in this paper is the following: for linearly separable data, as σ tends to zero, the Bayes optimal hyperplane converges to the maximal margin hyperplane. We also show that for linearly non-separable data, the Bayes optimal

[1] Cristianini *et al.* [5] also provide links between Bayesian classifiers and large margin hyperplanes. Their analysis is based on viewing the resulting posterior distribution as a hyperplane in a Hilbert space, and is quite different from ours.

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.

2 The setting

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 $X = \mathbb{R}^d$ 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 sample $\mathbf{x}_i$. 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 2.1 *A* classifier h *is a mapping from* X *to the set* $\{C_0, C_1\}$. *Let* $H_0 = \{\mathbf{x} : h(\mathbf{x}) = C_0\}$ *and* $H_1 = \{\mathbf{x} : h(\mathbf{x}) = C_1\}$.

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

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

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 2.3 *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*$(h : \hat{P})$. *If* $\mathcal{H}$ *is the set of hyperplane classifiers, we call* h^* *the* Bayes optimal hyperplane.

One restricted set of classifiers that has received a lot of attention is the set of hyperplane classifiers.

Definition 2.4 *We say that a classifier* h *is a* hyperplane classifier *if we can write it in the following form:*

$$h(\mathbf{x}) = \begin{cases} C_0 & \text{if } \sum_{j=1}^{d} w_j x_j - b < 0 \\ C_1 & \text{otherwise} \end{cases}$$

for some set of weights $b, w_1, \ldots, w_d$. *The set*

$$\{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\} = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} - b = 0\}$$

is called the decision boundary. *We will use* h *to denote both the classifier and the associated hyperplane decision boundary. When* $\mathcal{H}$ *is the set of hyperplane classifiers, we call the restricted Bayes optimal classifier relative to* $\mathcal{H}$ *the* Bayes optimal hyperplane.

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.

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 $P(C_1)$. There are many techniques for estimating the class conditional densities [1; 8; 13; 14]. 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 $\sigma^2 I$; $\sigma > 0$ is called the *smoothing parameter*. More precisely, we define for $i = 0, 1$

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

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*$(h : P_\sigma)$.

3 Linearly separable data

In this section, 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* [15].

Definition 3.1 *The* margin *of a hyperplane* h, *denoted by margin*(h), *is the smallest Euclidean distance from the hyperplane to a training instance.*

The expression representing the estimated Bayes error can be re-written as:

$$\begin{aligned} &\mathit{error}_\sigma(h) \\ &= \hat{P}(C_1) \int_{\mathbf{x}\in H_0} p_\sigma(\mathbf{x} \mid C_1)\, d\mathbf{x} + \hat{P}(C_0) \int_{\mathbf{x}\in H_1} p_\sigma(\mathbf{x} \mid C_0)\, d\mathbf{x}. \end{aligned}$$

As σ becomes smaller, $p_\sigma(\mathbf{x} \mid C_i)$ becomes a mixture of sharper and sharper Gaussian peaks. Hence, the probability mass of regions that do not include the peaks goes down to zero. In particular, the probability mass in H_0 according to $p_\sigma(\mathbf{x} \mid C_1)$ — the first of the two integrals — goes to zero. The same holds for the second integral. Thus, as we reduce σ, the estimated Bayes error of any separating hyperplane tends to zero. However, it does not follow that the estimated error of all hyperplanes decreases at the same rate. Indeed, it may be the case that the estimated error of some hyperplanes will decrease much more slowly than that of others. Our main result is that, as σ tends to zero, the maximal margin hyperplane will have better estimated Bayes error than any other. We also show the converse.

Theorem 3.2 *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.* $\mathit{error}_\sigma(h^*) < \mathit{error}_\sigma(h)$ *whenever* $\sigma < S$.
- h^* *has maximal margin.*

The intuition behind this result is that 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 we will have contributing to the integral contained within the expression for the estimated 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 $-\mathit{margin}(h)^2/(2\sigma^2)$. 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 candidate hyperplane, 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 3.2 shows that this is not the case.

Corollary 3.3 *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 $\mathit{margin}(h^*) - \delta$. Then there exists $S > 0$ such that, for all $\sigma < S$ and all $h \in \mathcal{H}_\delta$, $\mathit{error}_\sigma(h^*) < \mathit{error}_\sigma(h)$.*

Thus, as σ tends to zero, 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.

4 Data that are not linearly separable

We now analyze the behavior of the Bayes optimal hyperplane as the smoothing parameter tends to zero for the more general case of data that are not necessarily linearly separable.

Corollary 4.1 *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 $f_\sigma(\mathcal{D}, h)$ is a linear combination of integrals of Gaussian densities and has the following properties:

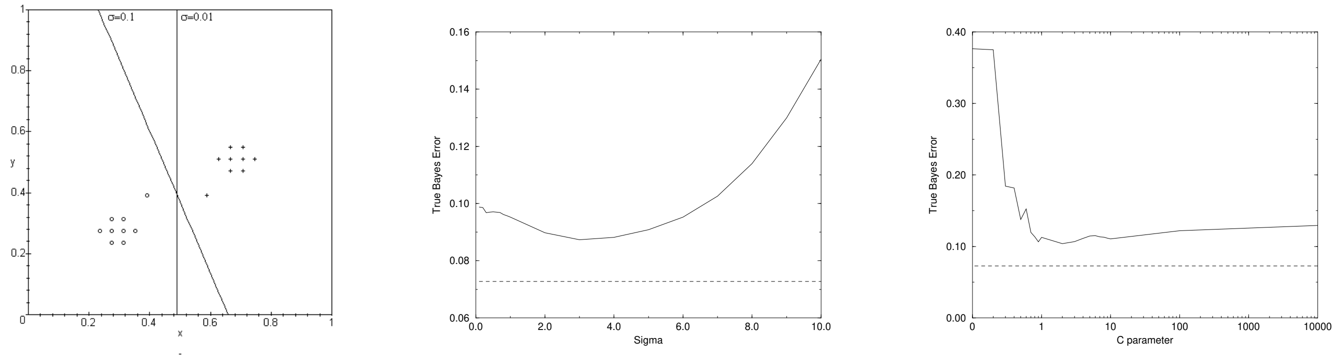
- $f_\sigma(\mathcal{D}, h) \to 0$ *as* $\sigma \to 0$
- *if h^* and h both minimize #incorrect and h^* has larger margin than h then there exists $S > 0$ such that $f_\sigma(\mathcal{D}, h^*) < f_\sigma(\mathcal{D}, h)$ whenever $\sigma < S$.*

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. If we have data that are not linearly separable, one way to pick a hyperplane is to look at all of the hyperplanes with the lowest number of training errors. Restricting ourselves to these hyperplanes, we have already accounted for the penalty of the misclassified data. We can then, for each such hyperplane, ignore the misclassified training instances. Then, for each hyperplane, we will have made our remaining data linearly separable and so we could just pick the hyperplane that maximizes the margin.

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.

5 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 3.2, 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. As argued in [12], it might be reasonable to consider a more general notion of *margin distribution*, which also considers the distances of other points from the margin. If we take σ to be small but non-zero, classifying using the estimated Bayes error will have precisely this effect. The larger we make σ, the larger the effect that points further from the margin have on the estimated Bayes error and therefore on the choice of hyperplane. Figure 1(a) illustrates one simple example where a larger value of σ leads to a choice of hyperplane which is arguably more reasonable. In the nonseparable case, we also get a similar tradeoff. For σ larger than zero, the f_σ term in $\mathit{error}_\sigma(h)$ may by non-negligible.



(a) (b) (c)

Figure 1: (a) Bayes optimal hyperplane for different σ. (b) Parzen Window Bayes Optimal Hyperplane for different σ. (c) Soft margin SVM for different values of C.

Thus, we may be prepared to sacrifice classification error to make f_σ smaller.

It is interesting to note that the form of the error function in Corollary 4.1 resembles the form of the *soft margin* error function often used in support vector machines (SVMs) [4]: $\|\mathbf{w}\|^2/2 + C(\sum_i \xi_i)$. 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 using a non-zero value of σ would achieve a similar effect to that of the soft margin error function, causing the Parzen window Bayes optimal hyperplane to has a similar accuracy to that of the soft margin hyperplane.[2] We used the "Pima Indian Diabetes" UC Irvine data set [2] and a synthetic data set. The Diabetes data set has eight features, with 576 training samples 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 Diabetes data set classification error on a separate 192 sample 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. Unfortunately the search space is not convex and contains local minima. Furthermore, for small σ, the space consists of numerous very large gently sloping plateaus. To cope with the large plateaus we used *bold driving* [1] and, motivated by Corollary 3.3, we seeded the search with the maximal margin hyperplane wherever possible. The gradient descent algorithm was then run until convergence.

Figures 1 (b) & (c) show respectively the true Bayes error of the Parzen window Bayes optimal hyperplane and of the soft margin hyperplane when using a training set of size nine for the synthetic data. The curves are the average error rates of the classifiers produced from ten, equally sized, training sets. The Bayes optimal error (the lowest attainable probability of error) was 7.27%. and is indicated on both of the graphs. The curve for the Parzen window hyperplane is noticeably smoother than the curve for the soft margin hyperplane. This is partially due to the fact that the optimum values of C for each of the ten different training sets tended to occur at quite different values.

The table below 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.

Parzen σ	Diabetes Error	Soft Margin C	Diabetes 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%	8.0%	7.6%	8.5%
Soft Margin	8.7%	7.9%	7.6%	8.7%
Max Margin	9.8%	9.4%	—	—

These results indicate that the Parzen windows Bayes optimal hyperplane, optimizing its somewhat different but arguably more natural error function, achieves very similar perfor-

[2] We used T. Joachim's SVMlight: www-ai.informatik.uni-dortmund.de/thorsten/svm_light.html and the SVM package [11] produced at Royal Holloway and AT&T.

mance to that of a soft margin hyperplane. This observation is supported by a paired t-test on the cross-validation folds of the Diabetes 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.

6 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. 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 classifier.

Our result has several implications. From one perspective, it can be viewed as providing a new probabilistic justification for maximal margin hyperplanes. 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. Conversely, 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.

This last observation suggests a new perspective on the long-standing debate between direct regression versus density estimation for classification. In many domains, direct regression empirically achieves higher classification accuracy than the Bayes optimal classifier. The usual explanation is that density estimation spends too much "effort" on minimizing "irrelevant" errors in $P(\mathbf{x})$, and not enough on reducing classification errors. Our approach provides an alternative solution, where the estimated density is not used directly in the form of the Bayes optimal decision boundary, but rather to evaluate classifiers in a restricted hypothesis class. Of course, it remains to verify in practice whether the resulting classifier is really competitive with standard direct regression classifiers in a wide variety of domains.

Recently, an interesting equivalence between the Gaussian kernel instantiation of the restricted Bayes optimal approach and and a sigmoid classifier with a size constraint on the weight vector (which is a case of structural risk minimization [15]) has been shown [3]. Together with the original motivation for maximal margin hyperplanes [15], these give three quite separately motivated approaches leading to maximal margin hyperplanes.

The equivalence with a direct regression classifier begs the question as to whether it was the specific choice of nonparametric density estimation and Gaussian kernels that lead the restricted Bayes optimal classifier to produce a high accuracy classifier or whether the restricted Bayes optimal classifier will also work well with other forms of density estimation. This is a question that we intend to investigate further, and preliminary experiments with other forms of density estimation have produced promising results. One interesting direction is in large discrete domains, where the Naive Bayes classifier often seems to outperform more sophisticated density estimators such as full Bayesian networks [7]. The common explanation is that a Bayesian network is too complex, and not tailored to the classification task. We plan to investigate the use of Bayesian networks as density estimators, combined with a restricted Bayes optimal classifier. In general, we want to experiment with this approach for a variety of density estimation approaches. We hope that it will allow us to combine the benefits of density estimation using realistically expressive representations with the high accuracy classification often associated with direct regression learning.

References

[1] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] C. Blake, E. Keogh, and C.J. Merz. UCI repository of machine learning databases, 1998.

[3] L. Bottou. Sigmoid classifiers and Bayes optimal hyperplanes. Personal communication, 1999.

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

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

[6] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.

[7] N. Friedman and M. Goldszmidt. Building classifiers using Bayesian networks. In *Proc. AAAI*, 1996.

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

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

[10] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.

[11] C. Saunders, M. O. Stitson, J. Weston, L. Bottou, B. Schölkopf, and A.Smola. Support vector machine — reference manual. Technical report, Royal Holloway, 1997.

[12] R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proc. 14th Int. Conf. on Machine Learning (ICML)*, 1997.

[13] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.

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

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