

# Bayes Optimal Hyperplanes → Maximal Margin Hyperplanes

**Simon Tong**

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

**Daphne Koller**

*koller@cs.stanford.edu*  
Computer Science Department  
Stanford University

## 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, a hyperplane is the Bayes optimal hyperplane if and only if it is the maximal margin hyperplane. We also analyze the behavior of the Bayes optimal hyperplane for non-linearly-separable data, showing that it has a very natural form. We explore the idea of using the hyperplane that is optimal relative to a density with some small non-zero kernel width, and present some promising preliminary results.

## 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.<sup>1</sup>

Bayes optimal classifiers use density estimation to 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 the training data, as is the associated

---

<sup>1</sup>Cristianini *et al.* [4] 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.

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 [11]).

We propose an alternative approach to dealing with the problem of variance in Bayes optimal classification in a spirit similar to that mentioned in [5; 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  $\sigma$ , the density for each class  $C_i$  is defined as a mixture of Gaussian kernels of width  $\sigma$ , centered on the data points in class  $C_i$ . Different values for  $\sigma$  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  $\sigma$  is often crucial for the accuracy of the Bayes optimal classifier. We can eliminate the bias induced by the smoothing effect of  $\sigma$  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  $\sigma$ .

Our main result in this paper is to show that, for linearly separable data, as  $\sigma$  tends to zero, a hyperplane is the Bayes

optimal hyperplane if and only if it is the maximal margin hyperplane. We also show that for non-linearly-separable data, the Bayes optimal hyperplane has a very natural interpretation: it minimizes the classification error, and among all the hyperplanes that have the same classification error, it is the one with the smallest margin relative to correctly classified points. Thus, the Bayes optimal hyperplane minimizes a different error function than the standard one used in support vector machines [3], but one which is arguably quite natural.

## 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  $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)$ , 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 sampled from a joint distribution  $P(C, \mathbf{x})$ . We can define the probability that a classifier  $h$  makes a classification error:

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

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

The classifier that minimizes this error is the *Bayes optimal classifier*.

**Definition 2.3** The Bayes optimal classifier relative to a distribution  $P$  is defined as:

$$h^*(\mathbf{x}) = \begin{cases} C_0 & \text{if } P(C_0 | \mathbf{x}) > P(C_1 | \mathbf{x}) \\ C_1 & \text{otherwise} \end{cases}$$

In order to use the Bayes optimal classifier, we need  $P(C_0 | \mathbf{x})$  and  $P(C_1 | \mathbf{x})$ . In general, these quantities are not known. The generative approach to classification uses the training data to estimate an approximate joint distribution  $\hat{P}(C, \mathbf{x})$ , 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 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 2.4** Given a joint distribution  $\hat{P}(C, \mathbf{x})$  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.

**Definition 2.5** 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  $b, w_1, \dots, w_D$ . 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  $\hat{P}(C, \mathbf{x}) = \hat{P}(C) \cdot \hat{p}(\mathbf{x} | C)$  where  $\hat{p}(\mathbf{x} | C)$  is the *class-conditional density* of the feature vectors  $\mathbf{x}$  within the class  $C$ .

We take the maximum likelihood estimates for  $P(C_0)$  and  $P(C_1)$  —  $\hat{P}(C_0) = \frac{n_0}{n_0+n_1}$  and  $\hat{P}(C_1) = \frac{n_1}{n_0+n_1}$ , where  $n_i$  is the number of training samples in class  $C_i$ . There are many techniques for estimating the class conditional densities [8; 12; 14; 15]. We choose a very simple variant of *non-parametric* density estimation: Parzen windows estimation with Gaussian kernels. To estimate  $p(\mathbf{x} | C_i)$ , we place a Gaussian kernel over each training instance  $\mathbf{x}_i$  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  $\Sigma = \sigma^2 I$  ( $\sigma > 0$ ). More precisely, we define for  $i = 0, 1$

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

The parameter  $\sigma$  is called the *smoothing parameter*.

Together,  $\hat{P}(C_0)$ ,  $\hat{P}(C_1)$  and  $p_\sigma(\mathbf{x} | C_i)$  define a joint density  $P_\sigma(C, \mathbf{x})$ , as required. We use  $error_\sigma(h)$  to denote  $error(h : P_\sigma)$ .

## 3 Linearly separable data

We now present our main result, showing the strong connection between our approach above and *maximal margin classifiers*.

**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.

In this section, we assume that the training data is 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.)

Consider the expression representing the estimated Bayes error:

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

As  $\sigma$  becomes smaller,  $p_\sigma(\mathbf{x} | C_i)$  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  $p_\sigma(\mathbf{x} | C_1)$  — the first of the two integrals — goes to zero. The same holds for the second integral. Thus, as we reduce  $\sigma$ , 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 the smoothing parameter 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 two statements are equivalent:*

- For each  $h \in \mathcal{H}$  such that  $h \neq h^*$ , there exists  $S > 0$  such that  $\text{error}_\sigma(h^*) < \text{error}_\sigma(h)$  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  $p_\sigma(\mathbf{x} | C_1)$  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 comprising the estimated Bayes error. The closer a kernel is to the boundary, the more it contributes to the total error. As  $\sigma$  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  $-\text{margin}(h)^2 / (2\sigma^2)$ . Thus, as  $\sigma$  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  $\sigma$  is small enough,  $h^*$  beats  $h$ . However, this does not suffice to show that, as we reduce  $\sigma$ , the Bayes optimal hyperplane  $h_\sigma^*$  for  $P_\sigma$  “converges” to the maximal margin hyperplane  $h^*$ . It could, perhaps, be the case that for each  $\sigma > 0$ ,  $h_\sigma^*$  is arbitrarily far away from  $h^*$ . In fact, we can show that this is not the case.

**Corollary 3.3** *Let  $h^* \in \mathcal{H}$  be the maximal margin hyperplane. Let  $\delta > 0$ . Let  $\mathcal{H}_\delta$  be the set of hyperplanes in  $\mathcal{H}$  with margins less than  $\text{margin}(h^*) - \delta$ . Then there exists  $S > 0$  such that, for all  $\sigma < S$  and all  $h \in \mathcal{H}_\delta$ ,  $\text{error}_\sigma(h^*) < \text{error}_\sigma(h)$ .*

Thus, as  $\sigma$  tends to zero, the margin of  $h_\sigma^*$  tends to the maximal margin. In other words, as the smoothing parameter tends to zero, the Bayes optimal hyperplane converges to the maximal margin hyperplane in terms of margin.

## 4 Data that are not linearly separable

The estimated Bayes error is perfectly well defined for data that are not linearly separable. We now analyze the behavior of the Bayes optimal hyperplane as the smoothing parameter tends to zero for this more general case.

**Proposition 4.1** *Given  $\sigma > 0$  and hyperplane  $h \in \mathcal{H}$  we can write the estimated Bayes error in the following form:*

$$\text{error}_\sigma(h) = \frac{1}{n} (f_\sigma(\text{training data}, h) + \#incorrect)$$

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

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

In other words, as  $\sigma$  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 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 data linearly separable and so will could just pick the hyperplane that maximizes the margin.

This proposition 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 Experimenting with different $\sigma$

Until now, we have focused on the behavior of the Bayes optimal hyperplane as  $\sigma$  gets arbitrarily close to zero. As we

discussed in the informal justification for Theorem 3.2, as  $\sigma$  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 that focusing only on the margin is necessarily the optimal approach. As argued in [13], 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  $\sigma$  to be small but non-zero, the estimated Bayes error will have precisely this effect. The larger we make  $\sigma$ , 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 example where a larger value of  $\sigma$  leads to a choice of hyperplane which is arguably more reasonable.

In the non-separable case, we also get a similar tradeoff. For  $\sigma$  bigger than zero, the  $f$  term may be non-negligible. Thus, we may be prepared to sacrifice classification error to make  $f$  smaller. In general, the Bayes optimal hyperplane will look at more than just the margin and misclassifications.

To test whether this behavior allows us to generalize better to test data, we conducted experiments with various settings of the smoothing parameter. For various values of  $\sigma$ , we tried to find the hyperplane that minimizes the estimated Bayes error for  $P_\sigma$ . The error function is a differentiable function of the weights of the hyperplane, so it is possible to use gradient descent techniques to find the Bayes optimal hyperplane for any given smoothing parameter. The update rules for naive gradient descent are straightforward to compute; we omit details for lack of space.

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  $\sigma$ , the space consists of numerous very large gently sloping plateaus. Thus, naive gradient descent converges to suboptimal solutions and very slowly. One technique that helps deal with this problem is based on Proposition 4.1. This result suggests that, for linearly separable data, the Bayes optimal hyperplane is within the same region of the space as the maximal margin hyperplane. We experimented with seeding the search with the maximal margin hyperplane and this seemed to improve the speed of convergence and quality of results.

We also experimented with a few methods to speed up the convergence. To cope with the large plateaus we used *bold driving* [1]. This is a technique whereby, whenever a step in the search is made that causes the error to decrease, the learning rate is increased by a constant factor. If a step is taken that increases the error then the step is retracted and the learning rate is reduced. This technique helps to make progress across gently sloping plateaus faster. We also found that keeping a separate independently-adjusting learning rate for each dimension of the search space aids convergence. We also tried other techniques such as conjugate gradient, but they were not as successful, possibly due to roundoff errors (particularly for small  $\sigma$ ).

For our experiment we considered a domain in which Support Vector Machines (SVMs) [16; 3; 2] have had good empirical successes [6; 10]: the task of text classification. The instances in the domain were articles from the Reuters collec-

tion<sup>2</sup>, with the feature vector representing a stemmed TFIDF-weighted word frequency vector for the document.<sup>3</sup> Our goal was to classify which articles were in the topic “corporate acquisitions.” The vectors have dimension of around 10,000. We deliberately tried to exacerbate the variance problem, using only 600 instances for the training. Even with such limited training data, the linear SVM<sup>4</sup> (i.e., the support vector machine which uses the simple Euclidean inner product as the kernel function) obtains a classification error of 100 out of 2000 on a separate, independently selected data set. We note that this training set was almost linearly separable: the hyperplane produced by the SVM misclassified one training sample out of the 600.

We then experimented with our gradient descent algorithm, for different values of  $\sigma$ . We used the hyperplane produced by the linear SVM to seed the search, and then ran the algorithm until the average learning rate reduced itself to a very small value. Figure 1(b) and (c) show the change in the classification error on the test set over the trajectory of the gradient descent. Both curves show the presence of numerical overfitting. For  $\sigma = 0.2$ , for example, the error on the 2000 data set decreased to 74 before climbing up to 121. Similar behavior occurred for smaller values of  $\sigma$ , although the curves tended to dip then climb back up in a shorter number of iterations.

Thus, it may be the case that better performance could be achieved if early stopping, e.g., using a validation set, were to be used. We have not yet had time to experiment with this idea. However, to indicate what these results might look like, our results below list both the test set classification error at convergence and at their lowest point in the process.

The following table shows the final and lowest error on the 2000 test cases for different values of  $\sigma$ .

$\sigma$	0.05	0.1	0.12	0.17	0.19	0.2	0.21	0.23	0.25	0.3
final err	100	100	99	116	114	121	94	125	132	133
min err	100	100	99	89	86	74	94	100	100	100

## 6 Conclusions and future work

We have defined 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 the use of a maximal margin hyperplane. 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

<sup>2</sup>Obtained from [www.research.att.com/~lewis/](http://www.research.att.com/~lewis/).

<sup>3</sup>The training data was obtained from [ftp-ai.cs.uni-dortmund.de/pub/Users/thorsten/](http://ftp-ai.cs.uni-dortmund.de/pub/Users/thorsten/).

<sup>4</sup>We used the SVMlight package by Thorsten Joachims with the PR\_LOQO optimizer by A. Smola.

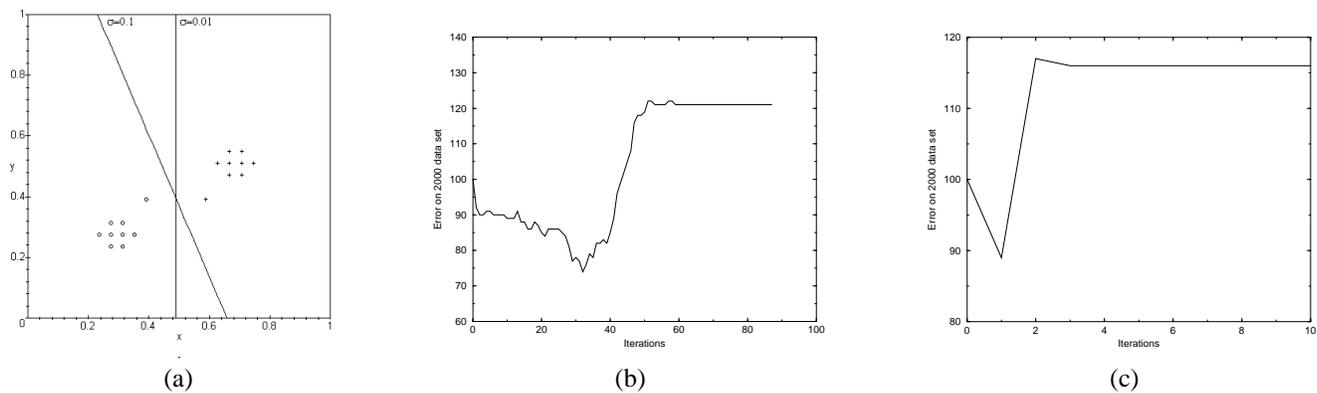


Figure 1: (a) Bayes optimal hyperplane for different  $\sigma$ . (b) Convergence curve for  $\sigma = 0.2$ . (c) Convergence curve for  $\sigma = 0.17$ .

a density — a nonparametric 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 extremely well even in high-dimensional spaces. Conversely, our result suggests that finding a hyperplane optimal relative to a complex density can be better than finding one which is optimal relative to a simpler one: 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 justification for this behavior 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 errors in the classification are explicitly reduced in a second phase, based on the results of the density estimation. 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.

We also want to use our approach to deal with a related issue. In large discrete domains, the Naive Bayes classifier seems to perform surprisingly well. Attempts to use a more sophisticated density estimator, such as a full Bayesian network, have not led to significantly increased accuracy [7]. Again, the proposed explanation is that algorithms for learning a full Bayesian network do not focus their efforts on the classification task. Perhaps if we estimate each of the two class conditional densities using a Bayesian network, and then find the optimal hyperplane relative to these densities, the resulting classifier will achieve higher performance than the simplistic Naive Bayes model.

In general, we want to experiment with this approach for a variety of density estimation approaches, in both continuous and discrete domains. 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] Christopher J.C. Burges. A tutorial on support vector machines for pattern recognition. In *Data Mining and Knowledge Discovery*, volume 2:2, 1998.
- [3] C. Cortes and V. Vapnik. Support vector networks. In *Machine Learning*, volume 20, pages 273–297, 1995.
- [4] N. Cristianini, J. Shawe-Taylor, and P. Sykacek. Bayesian classifiers are large margin hyperplanes in a Hilbert space. In *Proc. NeuroCOLT2*, 1998.
- [5] R.O. Duda and P.E. Hart. *Pattern Classification and Scene Analysis*. Wiley, New York, 1973.
- [6] S.T. Dumais, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representations for text categorization. In *Proc. Seventh International Conference on Information and Knowledge Management*, 1998.
- [7] Nir Friedman and Moises Goldszmidt. Building classifiers using bayesian networks. In *Proc. AAAI*, 1996.
- [8] K. Fukunaga. *Introduction to Statistical Pattern Recognition*. Boston: Academic Press, second edition, 1990.
- [9] W.H. Highleyman. Linear decision functions, with application to pattern recognition. In *Proc. IRE*, volume 49, pages 31–48, 1961.
- [10] T. Joachims. Text categorization with support vector machines. In *European Conference on Machine Learning (ECML)*, 1997.
- [11] T. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
- [12] B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996.
- [13] R. Schapire, Y. Freund, P. Bartlett, and W. Lee. Boosting the margin: A new explanation for the effectiveness of voting methods. In *Proc. 14th International Conference on Machine Learning (ICML)*, 1997.
- [14] D.W. Scott. *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley, 1992.
- [15] B.W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London, 1986.
- [16] V. Vapnik. *Estimation of Dependences Based on Empirical Data*. Nauka, Moscow, 1979. In Russian. English Translation: Springer Verlag, New York, 1982.