Maximal Margin Hyperplane Classifiers using Non-Parametric Density Estimation

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

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

Technical Report

December 1998

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 a notion of the score of a hyperplane relative to a given density, measuring how "far" the hyperplane is from the Bayes optimal decision boundary. We show that, as we reduce the smoothing parameter to zero, a hyperplane is the "best" approximation to the Bayes optimal decision boundary if and only if it is the maximal margin hyperplane.

1 Introduction

The notion of the "margin" of a classifier — the smallest distances between the decision boundary and training data — is a fundamental one in modern machine learning. The idea of explicitly maximizing the margin was introduced by Vapnik [5], who provides formal justification for the idea in terms of *risk minimization*. Recent empirical results show impressive success on difficult practical problems.

More recently, Schapire *et al.* [3] showed that the Adaboost algorithm is effective at producing ensembles with large minimum margins. Here the minimum margin of an ensemble classifier is essentially the smallest vote it gives to any correct training label. However, Grove and Schuurmans [2] showed that when the minimum margin is explicitly maximized via an algorithm called "LP-Adaboost," the learned ensemble classifier empirically does not necessarily lead to better generalization performance.

In this paper, we provide an alternative motivation for the idea of maximizing the margin,

by showing a tight connnection to Bayes optimal classifiers.[1] Bayes optimal classifiers use density estimation to perform classification by estimating the class priors and class conditional densities and then using the Bayes optimal classifier with our estimates of the distributions. The Bayes Optimal Classifier is known to have the nice property that if we have estimated the class priors and class conditional density functions perfectly then we are guaranteed to have the classifier that has the lowest expected test error.

One way of estimating the class conditional densities is nonparametric density estimation using Parzen windows [4]. Here, a Gaussian kernel is placed on each of the data instances in a given class; the mixture of these densities is the estimate for the associated class conditional density. The variance associated with the Gaussian kernels is the *smoothing parameter*. The choice of the smoothing parameter corresponds to a choice along the bias-variance spectrum. Smaller values for the smoothing parameter (sharper peaks for the kernels) correspond to higher variance but lower bias estimates of the density. As the smoothing parameter goes to zero, the variance of the estimator goes to infinity, providing no generalization. Thus, the choice of a smoothing parameter is often crucial for the accuracy of the Bayes optimal classifier (for a finite number of training instances).

We take a different approach to resolving the bias-variance tradeoff. We reduce the variance by restricting our hypothesis space to be hyperplanes. We then use non-parametric density estimation with Gaussian kernels and see which hyperplane best approximates the Bayes optimal classifier. As we reduce the smoothing parameter within the Gaussian kernels we reduce the bias of the Bayes optimal classifier; however, the restriction on our hypothesis space prevents the variance of our classifier from growing unboundedly. We show that, as we reduce the smoothing parameter to zero, a hyperplane has maximal margin if and only if a hyperplane is the "best" approximation to the Bayes optimal classifier when using non-parametric density estimation.

2 The Setting

Suppose we have a feature space $X \subseteq \mathbb{R}^D$ and linearly separable training data $(\mathbf{x_1}, y_1)$, $(\mathbf{x_2}, y_2), \ldots (\mathbf{x_n}, y_n)$, where $y_i \in \{C_0, C_1\}$. 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 $y_j = C_i$.

We place Gaussian kernels with diagonal covariance matrices $\Sigma = \sigma^2 I$ ($\sigma > 0$) on each training data point. That is, we define for $i = 0, 1$

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

Definition 2.1 *Given $\sigma > 0$ we define Bayes optimal classifier $h^*_\sigma = P(C_1|\mathbf{x}) - P(C_0|\mathbf{x})$.*

Now, h^*_σ classifies $\mathbf{x} \in C_0$

$$\Leftrightarrow P(C_0|\mathbf{x}) - P(C_1|\mathbf{x}) > 0 \tag{2}$$

[1] Cristianini *et al.* [1], 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.

$$\Leftrightarrow p(\mathbf{x}|C_0)P(C_0) > p(\mathbf{x}|C_1)P(C_1). \tag{3}$$

We take the maximum likelihood estimates for $P(C_0)$ and $P(C_1)$,

$$P(C_0) = \frac{n_0}{n_0 + n_1} \tag{4}$$

$$P(C_1) = \frac{n_1}{n_0 + n_1}. \tag{5}$$

Thus from equation (1) and equation (3) h^*_σ classifies $x \in C_0$ iff

$$\sum_{\mathbf{x_j} \in C_0} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{x_j})^T(\mathbf{x}-\mathbf{x_j})} > \sum_{\mathbf{x_j} \in C_1} e^{-\frac{1}{2\sigma^2}(\mathbf{x}-\mathbf{x_j})^T(\mathbf{x}-\mathbf{x_j})}. \tag{6}$$

Definition 2.2 *Let $h : X \rightarrow I\!R$. We say that h* admits a continuous decision boundary *if $\exists \phi : [0,1]^{D-1} \rightarrow X$ s.t. $\phi \in \mathcal{C}^0$ (i.e., ϕ is continuous) and $Im\,\phi = \{x \in X \mid h(x) = 0\}$ where $Im\,\phi$ denotes the image of ϕ.*

Definition 2.3 *$H = \{h : X \rightarrow I\!R \mid h \in \mathcal{C}^0$, h admits a continuous decision boundary and if $\mathbf{x} \in C_1$ then $h(\mathbf{x}) > 0$ and if $\mathbf{x} \in C_0$ then $h(\mathbf{x}) < 0$.*

Given an $h \in H$ and $\sigma > 0$ we want a way of seeing how "close" h is to h^*_σ. Intuitively h is a good approximation to h^*_σ if for each $\mathbf{x}$ on h's decision surface $P_\sigma(C_0|\mathbf{x}) \approx P_\sigma(C_1|\mathbf{x})$. We also want h's decision boundary to be a closer approximation to that of h^*_σ when $p(\mathbf{x})$ is high (in other words we want to be close where most of the data are likely to be). This leads us to the following definitions.

Definition 2.4 *Let $\mathbf{x} \in X$. Then define*

$$score_\sigma(\mathbf{x}) = p(\mathbf{x})|P_\sigma(C_0|\mathbf{x}) \ - \ P_\sigma(C_1|\mathbf{x})|.$$

Definition 2.5 *Let $h \in H$ and let ϕ be a continuous decision boundary admitted by h. Let $\sigma > 0$. Then define*

$$score_\sigma(h) = \max_{\mathbf{x} \in Im\,\phi} score_\sigma(\mathbf{x}).$$

By Bayes Rule,

$$score_\sigma(h) = \max_{\mathbf{x} \in Im\,\phi} |p(\mathbf{x}|C_0)P(C_0) - p(\mathbf{x}|C_1)P(C_1)|. \tag{7}$$

Notice that since ϕ is continuous on a compact set and since $p(\mathbf{x}|C_i)$ is continuous for $i = 0, 1$ then the *max* in equation (7) exists and hence definition 2.5 is well defined. Note that $score_\sigma(h)$ is the infinity norm of $p(\mathbf{x})|P(C_0|\mathbf{x}) - P(C_1|\mathbf{x})|$ for $\mathbf{x} \in Im\,\phi$.

Suppose $h \in H$ and h admits ϕ. Let $\sigma > 0$. Suppose further that $score_\sigma(h) = 0$. Then notice that $Im\,\phi \subseteq \{\mathbf{x} \in X \mid P(C_0|\mathbf{x}) = P(C_1|\mathbf{x})\}$. Hence the decision boundary of h is a subset of the (not necessarily continuous) decision boundary of h^*_σ, the Bayes optimal classifier.

Thus $score_\sigma(\ .\ )$ is a useful measure of how close a hypothesis in H is to h^*_σ.

Lemma 2.6 *$\exists S > 0$ s.t. $\forall \sigma \in (0, S)$ h^*_σ correctly classifies all training instances.*

Proof. Omitted.

We now introduce a few concepts that will be important for seeing what determines the score of a point and a surface.

Definition 2.7 *Let $d > 0$. Let $\mathbf{x} \in X$. We define $\tau(\mathbf{x}, d, C_i)$ as being the number of training samples in class C_i that are squared Euclidean distance d from $\mathbf{x}$.*

Definition 2.8 *For each $\mathbf{x} \in X$ consider the set of all squared distances from $\mathbf{x}$ to each of the n training samples: $\{d_1, d_2, \ldots d_m\}$, $m \leq n$. Now, for each element d of this set we see if $\tau(\mathbf{x}, d, C_0) = \tau(\mathbf{x}, d, C_1)$ and remove d from the set if this equality holds.*

This eventually results in a reduced set of distances that we call the unbalanced squared distances for $\mathbf{x}$. *We denote the i^{th} smallest unbalanced squared distance from $\mathbf{x}$ by $d^{\mathbf{x}}_i$, with $d^{\mathbf{x}}_1$ being the smallest unbalanced squared distance for $\mathbf{x}$.*

Definition 2.9 *Let $\lambda^{\mathbf{x}}_i = \tau(\mathbf{x}, d^{\mathbf{x}}_i, C_0) - \tau(\mathbf{x}, d^{\mathbf{x}}_i, C_1)$ i.e., $\lambda^{\mathbf{x}}_i$ tells us how many more (or less) training data of class C_0 we have at distance d_i from $\mathbf{x}$ than of class C_1. Notice that $\lambda^{\mathbf{x}}_i$ is never zero since $d^{\mathbf{x}}_i$ is an* unbalanced *distance.*

3 Proof Outline

The aim of this paper is to show that, as we reduce the smoothing parameter σ, a hyperplane is the best fit to the Bayes optimal classifier if and only if it is the hyperplane that maximizes the margin.

We prove this result in four steps.

1. We show that, as we reduce σ, the score of a point is dominated by its smallest unbalanced distance.

2. We show that, as we reduce σ, the score of a hypothesis is asymptotically equivalent to the score of a special *critical* point on the decision boundary. Of all the points on the decision boundary, the critical point has the minimal smallest unbalanced distance (which we denote by the *critical* distance).

3. We then turn our attention to hypotheses that have hyperplanes as decision boundaries. We show that the hypothesis that minimizes the score is the hypothesis that has maximal critical distance and vice versa.

4. Finally we show that a hyperplane has maximal critical distance if and only if it has maximal margin (where the margin is defined to be the smallest Euclidean distance from the hyperplane to a training sample).

4 What dominates the score of a point?

First, we prove upper and lower bounds on the score of a point in X. This provides us with some intuitions as to what eventually determines the score of a point as σ tends to zero.

Lemma 4.1 *Let $\mathbf{z} \in X$ and let $d_1^{\mathbf{z}}$ exist. For convenience, let $N_\sigma = \frac{1}{\sigma(2\pi)^{\frac{d}{2}}}$. Then*

$$N_\sigma \left\{ |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}} \right\} \leq score_\sigma(\mathbf{z}) \leq N_\sigma \left\{ |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}} \right\}. \tag{8}$$

(If $d_2^{\mathbf{z}}$ does not exist then we can drop the $n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}$ term within each bound.)

Proof.

$$\frac{1}{N_\sigma} score_\sigma(\mathbf{z}) = \left| \sum_{\mathbf{x_j} \in C_0} e^{-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{x_j})^T(\mathbf{z}-\mathbf{x_j})} - \sum_{\mathbf{x_j} \in C_1} e^{-\frac{1}{2\sigma^2}(\mathbf{z}-\mathbf{x_j})^T(\mathbf{z}-\mathbf{x_j})} \right|. \tag{9}$$

Notice that $(\mathbf{z}-\mathbf{x_j})^T(\mathbf{z}-\mathbf{x_j})$ is the squared Euclidean distance from $\mathbf{z}$ to $\mathbf{x_j}$. So, instead of summing over the training data one by one, we now rearrange the summation in terms of distances from training data to the point $\mathbf{z}$. Notice that any balanced distances for $\mathbf{z}$ (i.e., any distances d for which $\tau(\mathbf{z}, d, C_0) = \tau(\mathbf{z}, d, C_1)$) will be canceled out in the subtraction in equation (9). Thus we only need to sum over the unbalanced distances. Recall that $\lambda_i^{\mathbf{z}}$ is the excess (or deficit if $\lambda_i^{\mathbf{z}}$ is negative) of training data points of class C_0 we have at distance $d_i^{\mathbf{z}}$. $\lambda_i^{\mathbf{z}} > 0$ only when the number of training data of class C_0 at distance $d_i^{\mathbf{z}}$ from $\mathbf{z}$ exceeds the number of training data of class C_1 at that distance from $\mathbf{z}$.
Thus equation (9) becomes

$$\left| \sum_{i:\lambda_i^{\mathbf{z}}>0} \lambda_i^{\mathbf{z}} e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} - \sum_{i:\lambda_i^{\mathbf{z}}<0} (-\lambda_i^{\mathbf{z}}) e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} \right|. \tag{10}$$

Now we use this to prove the lower bound.
Suppose that $\lambda_1^{\mathbf{z}} > 0$. Then equation (10) becomes

$$\geq \lambda_1^{\mathbf{z}} e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + \sum_{i \geq 2:\lambda_i^{\mathbf{z}}>0} \lambda_i^{\mathbf{z}} e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} - \sum_{i \geq 2:\lambda_i^{\mathbf{z}}<0} (-\lambda_i^{\mathbf{z}}) e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} \tag{11}$$

$$\geq \lambda_1^{\mathbf{z}} e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - \sum_{i \geq 2:\lambda_i^{\mathbf{z}}>0} \lambda_i^{\mathbf{z}} e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} - \sum_{i \geq 2:\lambda_i^{\mathbf{z}}<0} (-\lambda_i^{\mathbf{z}}) e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}}. \tag{12}$$

Now, $\forall i \geq 2$ $d_i^{\mathbf{z}} \geq d_2^{\mathbf{z}}$ thus equation (12) is

$$\geq \lambda_1^{\mathbf{z}} e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}. \tag{13}$$

By a similar argument for the case where $\lambda_1^{\mathbf{z}} < 0$ we obtain the general bound

$$\frac{1}{N_\sigma} score_\sigma(\mathbf{z}) \geq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}. \tag{14}$$

We now prove the upper bound. Equation (10) is

$$\leq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + \left| \sum_{i \geq 2 : \lambda_i^{\mathbf{z}} > 0} \lambda_i^{\mathbf{z}} e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} - \sum_{i \geq 2 : \lambda_i^{\mathbf{z}} < 0} (-\lambda_i^{\mathbf{z}}) e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} \right| \tag{15}$$

$$\leq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + \left| \sum_{i \geq 2 : \lambda_i^{\mathbf{z}} > 0} \lambda_i^{\mathbf{z}} e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} \right| + \left| \sum_{i \geq 2 : \lambda_i^{\mathbf{z}} < 0} (-\lambda_i^{\mathbf{z}}) e^{-\frac{d_i^{\mathbf{z}}}{2\sigma^2}} \right| \tag{16}$$

$$\leq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}. \tag{17}$$

□

This lemma shows us that the score of a point is dominated by its smallest unbalanced distance. Note that a simple limit argument shows us that $\exists S_{\mathbf{z}}$ s.t. $\forall \sigma \in (0, S_{\mathbf{z}})$, $|\lambda_1^{\mathbf{z}}| \, e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n \, e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}} > 0$. In other words, for sufficiently small σ the expression for the lower bound is greater than zero.

We now wish to see which points in the space X will have higher scores than others as $\sigma \xrightarrow{+} 0$ i.e., given $\mathbf{x} \in X$ we wish to see what happens to $score_\sigma(\mathbf{x})$ relative to other points in the space X. Clearly, this quantity will tend to 0, but, for example, will it dominate over some other points for all σ less than a certain value?

Lemma 4.2 *Let $\mathbf{z} \in X$ s.t. $\forall d \geq 0 \;\; \tau(\mathbf{z}, d, C_0) = \tau(\mathbf{z}, d, C_1)$, that is, all distances of $\mathbf{z}$ are balanced. Then $score_\sigma(\mathbf{z}) = 0$.*

Proof. Omitted.

Lemma 4.3 *Let $\mathbf{u}, \mathbf{v} \in X$ s.t. $d_1^{\mathbf{u}} < d_1^{\mathbf{v}}$. Then*

$$\frac{score_\sigma(\mathbf{v})}{score_\sigma(\mathbf{u})} \to 0 \;\; as \;\; \sigma \xrightarrow{+} 0. \tag{18}$$

Proof. Consider the ratio

$$0 \leq \frac{score_\sigma(\mathbf{v})}{score_\sigma(\mathbf{u})}. \tag{19}$$

By lemma 4.1 for sufficiently small σ we see that equation (19) is

$$\leq \frac{N_\sigma \left\{ |\lambda_1^{\mathbf{v}}| e^{-\frac{d_1^{\mathbf{v}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{v}}}{2\sigma^2}} \right\}}{N_\sigma \left\{ |\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}} \right\}} \tag{20}$$

$$= \frac{|\lambda_1^{\mathbf{v}}| e^{-\frac{d_1^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| - n e^{-\frac{d_2^{\mathbf{u}} - d_1^{\mathbf{u}}}{2\sigma^2}}}. \tag{21}$$

Notice that for all $i \geq 2$, $d_i^{\mathbf{u}} - d_1^{\mathbf{u}} > 0$. Also, for all $j \geq 1$, $d_j^{\mathbf{v}} - d_1^{\mathbf{u}} > 0$ by the assumptions of the lemma. Thus equation (21) $\rightarrow 0$ as $\sigma \xrightarrow{+} 0$.

□

Notice that a trivial consequence of this lemma is that $\exists S > 0$ s.t. $\forall \sigma \in (0, S)$ $\frac{score_\sigma(\mathbf{v})}{score_\sigma(\mathbf{u})} < 1$. Hence we have the following corollary.

Corollary 4.4 *Let* $\mathbf{u}, \mathbf{v} \in X$ *s.t.* $d_1^{\mathbf{u}} < d_1^{\mathbf{v}}$. *Then* $\exists S > 0$ *s.t.* $\forall \sigma \in (0, S)$

$$score_\sigma(\mathbf{u}) > score_\sigma(\mathbf{v}). \tag{22}$$

Corollary 4.5 *Let* $\mathbf{u}, \mathbf{v} \in X$ *such that* $d_1^{\mathbf{u}} \leq d_1^{\mathbf{v}}$ *and if* $d_1^{\mathbf{u}} = d_1^{\mathbf{v}}$ *then* $|\lambda_1^{\mathbf{u}}| \geq |\lambda_1^{\mathbf{v}}|$. *Then*

$$\frac{score_\sigma(\mathbf{v})}{score_\sigma(\mathbf{u})} \rightarrow l \text{ as } \sigma \xrightarrow{+} 0 \text{ for some } l \in [0, 1]. \tag{23}$$

Proof. From lemma 4.1 we have that for sufficiently small σ,

$$\frac{|\lambda_1^{\mathbf{v}}| e^{-\frac{d_1^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| + n e^{-\frac{d_2^{\mathbf{u}} - d_1^{\mathbf{u}}}{2\sigma^2}}} \leq \frac{score_\sigma(\mathbf{v})}{score_\sigma(\mathbf{u})} \leq \frac{|\lambda_1^{\mathbf{v}}| e^{-\frac{d_1^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{v}} - d_1^{\mathbf{u}}}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| - n e^{-\frac{d_2^{\mathbf{u}} - d_1^{\mathbf{u}}}{2\sigma^2}}}. \tag{24}$$

The result of the corollary then follows.

□

These series of lemmas and corollaries essentially tell us the score of a point u relative to the scores of other points is predominantly determined by its minimal unbalanced distance (and $\lambda_1^{\mathbf{u}}$ whenever there are ties).

5 What determines the score of a decision surface?

We now have a good idea of how the scores of different points in X behave as σ tends to zero. Given an $h \in H$ the next step is to find points on the decision boundary $Im(\phi)$ that will dominate the score of h as σ tends to zero. That is, we want to find points on the decision boundary $Im(\phi)$ whose score will eventually be larger than that of all of the other points in $Im(\phi)$.

Definition 5.1 *Let $Z \subseteq X$. Suppose $\exists \mathbf{u} \in Z$ s.t. $d_1^{\mathbf{u}}$ exists and s.t. for each $\mathbf{z} \in Z$ either:*
- *$d_1^{\mathbf{z}}$ does not exist*

or
- *$d_1^{\mathbf{u}} < d_1^{\mathbf{z}}$*

or
- *$d_1^{\mathbf{z}} = d_1^{\mathbf{u}}$ and $\lambda_1^{\mathbf{u}} > \lambda_1^{\mathbf{z}}$.*

We then say that $\mathbf{u}$ *is a* critical point of Z *and say that $d_1^{\mathbf{u}}$ is the* critical squared distance of Z. *If Z is the decision boundary ($Im\,\phi$) of some hypothesis $h \in H$ we also say that* $\mathbf{u}$ *is a* critical point of h *and $d_1^{\mathbf{u}}$* is the critical squared distance of h.

Note that it is easy to show that if Z has a critical point then the critical distance is unique. Intuitively, if our decision surface has a critical point $\mathbf{u}$ then for almost every other point $\mathbf{z}$ on that surface the score of $\mathbf{u}$ is eventually greater than the score of $\mathbf{z}$. That is, by corollary 4.4, there exists some S s.t. $\forall \sigma < S$ the score of $\mathbf{u}$ is greater than the score of $\mathbf{z}$. So does that mean that we are done? Does that mean to show that a decision surface's score (definition 2.5) is determined by its critical point all we have to do is take the minimum of all of these S's? Not quite. Generally we have an infinite number of points to compare with our critical point and it could be that the infimum of our set of S's is exactly 0. We need a couple of stronger conditions to hold on our decision surface. While these technical conditions may seem at first obscure, they do actually capture the two cases in which the above intuitive argument will not hold. Essentially, since we require some form of uniform convergence, the two conditions ensure that $d_1^{\mathbf{u}}$ and $\lambda_1^{\mathbf{u}}$ are sufficiently isolated.

Lemma 5.2 *Let $Z \subseteq X$ and let Z have a critical point $\mathbf{u} \in Z$. Also suppose that*

$$\inf\{d_2^{\mathbf{z}} \mid z \in Z\} > d_1^{\mathbf{u}} \tag{25}$$

and that

$$\exists \delta > 0 \text{ s.t. } \forall \mathbf{z} \in Z \ |\lambda_1^{\mathbf{z}}| \leq |\lambda_1^{\mathbf{u}}| \text{ whenever } d_1^{\mathbf{z}} \leq d_1^{\mathbf{u}} + \delta. \tag{26}$$

Then

$$\lim_{\sigma \stackrel{+}{\to} 0} \max_{\mathbf{z} \in Z} \frac{score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} = 1. \tag{27}$$

Proof. Let $d_2 = \inf\{d_2^{\mathbf{z}} \mid \mathbf{z} \in Z\}$ and let $\mathbf{z} \in Z$.
We first provide upper bounds for the numerator of equation (27).
- *Case: $d_1^{\mathbf{z}}$ doesn't exist.* Then by lemma 4.2 $score_\sigma(\mathbf{z}) = 0$.
- *Case: $d_1^{\mathbf{z}} > d_1^{\mathbf{u}} + \delta$.* Then by lemma 4.1

$$\frac{1}{N_\sigma} score_\sigma(\mathbf{z}) \leq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}. \tag{28}$$

(If $d_2^{\mathbf{z}}$ doesn't exist then we can drop the second term of this expression.)

Now $d_1^{\mathbf{z}} > d_1^{\mathbf{u}} + \delta$, $|\lambda_1^{\mathbf{z}}| \leq n$ and $d_2^{\mathbf{z}} \geq d_2$ so equation (28) becomes

$$\leq n e^{-\frac{d_1^{\mathbf{u}}+\delta}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}. \tag{29}$$

- *Case:* $d_1^{\mathbf{z}} \leq d_1^{\mathbf{u}} + \delta$. Then by lemma 4.1 we have

$$\frac{1}{N_\sigma} score_\sigma(\mathbf{z}) \leq |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} + n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}}. \tag{30}$$

$\mathbf{u}$ is a critical point so $d_1^{\mathbf{z}} \geq d_1^{\mathbf{u}}$. Also $d_2^{\mathbf{z}} \geq d_2$ and by the assumptions of the lemma $|\lambda_1^{\mathbf{z}}| \leq |\lambda_1^{\mathbf{u}}|$. Thus equation (30) becomes

$$\leq |\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}. \tag{31}$$

We can now put everything together. By the above, and by applying lemma 4.1 on the denominator we know that $\forall \mathbf{z} \in Z$ and $\forall \sigma \in (0, S_{\mathbf{u}})$

$$\frac{score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} \tag{32}$$

$$\leq \frac{\max\left\{0, n e^{-\frac{d_1^{\mathbf{u}}+\delta}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}, |\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}\right\}}{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}}} \tag{33}$$

$$= \max\left\{\frac{n e^{-\frac{d_1^{\mathbf{u}}+\delta}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}}}, \frac{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}}}\right\}. \tag{34}$$

Notice that $S_{\mathbf{u}}$ and all of the terms in equation (34) are independent of $\mathbf{z}$. Thus $\forall \sigma \in (0, S_{\mathbf{u}})$

$$1 \leq \max_{\mathbf{z} \in Z} \frac{score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} \leq (34). \tag{35}$$

Now,

$$\frac{n e^{-\frac{d_1^{\mathbf{u}}+\delta}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}}} = \frac{n e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}}\left(e^{-\frac{\delta}{2\sigma^2}} + e^{-\frac{d_2-d_1^{\mathbf{u}}}{2\sigma^2}}\right)}{e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}}\left(|\lambda_1^{\mathbf{u}}| - n e^{-\frac{d_2^{\mathbf{u}}-d_1^{\mathbf{u}}}{2\sigma^2}}\right)} \to 0 \text{ as } \sigma \xrightarrow{+} 0. \tag{36}$$

This tends to zero since $d_2^{\mathbf{u}} > d_2 > d_1^{\mathbf{u}}$ by the assumption of the lemma.
Also,

$$\frac{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} + n e^{-\frac{d_2}{2\sigma^2}}}{|\lambda_1^{\mathbf{u}}| e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{u}}}{2\sigma^2}}} = \frac{e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}}\left(|\lambda_1^{\mathbf{u}}| + n e^{-\frac{d_2-d_1^{\mathbf{u}}}{2\sigma^2}}\right)}{e^{-\frac{d_1^{\mathbf{u}}}{2\sigma^2}}\left(|\lambda_1^{\mathbf{u}}| - n e^{-\frac{d_2-d_1^{\mathbf{u}}}{2\sigma^2}}\right)} \to 1 \text{ as } \sigma \xrightarrow{+} 0. \tag{37}$$

Thus (34) $\rightarrow 1$ as $\sigma \rightarrow 0$. Hence by equation (35),

$$\max_{\mathbf{z} \in Z} \frac{score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} \rightarrow 1 \text{ as } \sigma \stackrel{+}{\rightarrow} 0. \tag{38}$$

□

What we have shown is that, under certain technical conditions, the score of a set Z of points is asymptotically equal to the score of the critical point of Z.

6 Hyperplanes

We now will concentrate on hypotheses whose decision surfaces are hyperplanes in a hyper-rectangle $X \subseteq \mathbb{R}^D$.

Definition 6.1 *Let $H_p \subseteq H$ s.t. $\forall h \in H_p$ $\exists \mathbf{w} \in I\!R^D$, $b \in I\!R$ s.t.*

$$Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\} = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} + b = 0\}. \tag{39}$$

Definition 6.2 *Let Z be a hyperplane in X. We can write $Z = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} + b = 0\}$ where $\mathbf{w}$ is unit. We define the* margin *of Z (denoted by margin(Z)) to be the closest Euclidean distance that Z is to any training sample. Formally,*

$$margin(Z) = \min\{|d| \mid d = (\mathbf{w} \cdot \mathbf{x_i}) + b,\ i = 1 \ldots n\}.$$

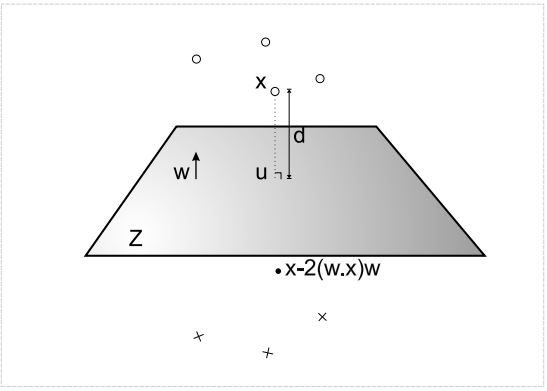
We denote the critical distance of Z by cdist(Z).

Lemma 6.3 *Let $h \in H_p$. Then* either *there are no $\mathbf{z} \in Z$ with unbalanced distances* or *h has a critical point $\mathbf{u}$ with $|\lambda_1^{\mathbf{u}}| = 1$.*

Proof. For ease of exposition assume that $Z = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} + b = 0\}$ where $\mathbf{w}$ is unit and $b = 0$ (the proof easily extends to the general case).
Notice for any point $\mathbf{x} \in X$ the perpendicular distance (and hence shortest distance) to Z is $\mathbf{w} \cdot \mathbf{x}$. Consider the subset U of the training data:

$$\begin{aligned} U \quad = \quad & \{\mathbf{x_i} \mid \text{there is not a training instance } \mathbf{x_j} \\ & \text{for which } Z \text{ is the perpendicular bisector of } \mathbf{x_i} \text{ and } \mathbf{x_j}\ \}. \end{aligned} \tag{40}$$

That is, U is the set of training samples $\mathbf{x_i}$ for which $\mathbf{x_i} - 2(\mathbf{w} \cdot \mathbf{x_i})\mathbf{w}$ is not a training sample.
Case: U is empty. Then we can split the training data into pairs such that Z is the perpendicular bisector of each pair. Hence no point of Z has an unbalanced distance.
Case: U is non-empty. U is a finite non-empty set so we can pick a training instance whose distance to Z is minimal (i.e., pick a $\mathbf{x} \in U$ s.t. $d = \mathbf{x} \cdot \mathbf{w}$ is minimal).



Consider $\mathbf{u} = \mathbf{x} - (\mathbf{x} \cdot \mathbf{w})\mathbf{w} \in Z$. By considering $\mathbf{u} \cdot \mathbf{w}$ it is straightforward to show that $\mathbf{u} \in Z$. We will show that $\mathbf{u}$ is a critical point of Z. Clearly $\mathbf{u}$ is distance d away from $\mathbf{x}$. Notice that there are no training samples in U that are even closer to Z than $\mathbf{x}$. Thus there are no points on Z with unbalanced distances strictly smaller than d. It is now enough to show that for all other points $\mathbf{z}$ on Z that have unbalanced distances d, $\lambda_1^{\mathbf{z}} = \lambda_1^{\mathbf{u}} = 1$. We defer the proof of this until the proof of lemma 6.4.

□

The previous lemma shows us that hyperplanes fall into one of two categories. In general we will be dealing with hyperplanes that do have critical distances, but we also have to take care of the degenerate cases in which they do not.

We now show that hyperplanes that possess a critical distance satisfy the two technical conditions that are required for the use of lemma 5.2.

Lemma 6.4 *Let $h \in H_p$ s.t. h has a critical point $\mathbf{u}$ and let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$. Then $\exists \delta > 0$ s.t. $\forall \mathbf{z} \in Z$ $|\lambda_1^{\mathbf{z}}| \leq |\lambda_1^{\mathbf{u}}|$ whenever $d_1^{\mathbf{z}} \leq d_1^{\mathbf{u}} + \delta$.*

Proof. Let $Z = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} + b = 0\}$ for some $\mathbf{w} \in \mathbb{R}^D$, $|\mathbf{w}| = 1$, $b \in \mathbb{R}$. Notice that if Z is the perpendicular bisector of $\mathbf{x_i}$ and $\mathbf{x_j}$ then we know that for all $\mathbf{z} \in Z$ $(\mathbf{z} - \mathbf{x_i}) \cdot (\mathbf{z} - \mathbf{x_i}) = (\mathbf{z} - \mathbf{x_j}) \cdot (\mathbf{z} - \mathbf{x_j})$ so we can disregard $\mathbf{x_i}$ and $\mathbf{x_j}$ when we examine unbalanced distances. Thus, we assume for the rest of the proof that there are no pairs of training data for which Z acts as a perpendicular bisector.

Note that $cdist(Z) = \sqrt{d_1^{\mathbf{u}}}$. Now consider:

$$e^* = \min \{ e_{ij} \mid e_{ij} \text{ is the distance of the shortest path between } \mathbf{x_i} \text{ and } \mathbf{x_j} \text{ via } Z \text{ and } i \neq j \}. \tag{41}$$

Now $e^* > 2cdist(Z)$ since $i \neq j$ and Z is not a perpendicular bisector of any $\mathbf{x_i}$ and $\mathbf{x_j}$. Let $\delta = \left(\frac{\frac{e^*}{2} - cdist(Z)}{2} \right) > 0$. Let $\mathbf{z} \in Z$ s.t. $\sqrt{d_1^{\mathbf{z}}} < cdist(Z) + \delta$. Then it is enough to show that $|\lambda_1^{\mathbf{z}}| = 1$. Notice that this will also prove that $|\lambda_1^{\mathbf{u}}| = 1$, concluding the proof of lemma 6.3.

Suppose for a contradiction that $|\lambda_1^{\mathbf{z}}| \geq 2$. Then this must mean that there are at least two distinct training data instances $\mathbf{x_i}$ and $\mathbf{x_j}$ that are distance $\sqrt{d_1^{\mathbf{z}}}$ from $\mathbf{z}$. But then this means that the path $\mathbf{x_i} \rightarrow \mathbf{z} \rightarrow \mathbf{x_j}$ has length

$$\sqrt{d_1^{\mathbf{z}}}+\sqrt{d_1^{\mathbf{z}}} < 2(cdist(Z)+\delta) = 2\left(cdist(Z) + \frac{\frac{e^*}{2} - cdist(Z)}{2}\right) = \frac{e^*}{2} + cdist(Z) < e^* \quad (42)$$

which contradicts the definition of e^*.
Thus for all $\mathbf{z} \in Z$ $|\lambda_1^{\mathbf{z}}| \leq |\lambda_1^{\mathbf{u}}|$ whenever $\sqrt{d_1^{\mathbf{z}}} \leq \sqrt{d_1^{\mathbf{u}}} + \delta$ and hence the result of the lemma follows easily.

□

Lemma 6.5 *Let $h \in H_p$ s.t. h has a critical point $\mathbf{u}$. Let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$. Then* $\inf\{d_2^{\mathbf{z}} \mid z \in Z\} > d_1^{\mathbf{u}}$.

Proof. By contradiction. The critical distance $\sqrt{d_1^{\mathbf{u}}}$ is denoted by $cdist(Z)$. Suppose that $\inf\{d_2^{\mathbf{z}} \mid z \in Z\} = d_1$. Then $\forall\delta > 0\ \exists\mathbf{z} \in Z$ s.t. $d_2^{\mathbf{z}} - d_1 < \delta$. Hence $\forall\delta > 0\ \exists\mathbf{z} \in Z$ s.t. $\sqrt{d_2^{\mathbf{z}}} - cdist(Z) < \delta$.
As before, we can disregard $\mathbf{x_i}$ and $\mathbf{x_j}$ for which Z is a perpendicular bisector. We define e^* as in lemma 6.4 and choose $\delta = \left(\frac{\frac{e^*}{2} - cdist(Z)}{2}\right) > 0$
So, we have some $\mathbf{z} \in Z$ s.t. $\sqrt{d_2^{\mathbf{z}}} - cdist(Z) < \delta$. This must mean that there are at least two distinct training data instances $\mathbf{x_i}$ and $\mathbf{x_j}$ that are distance $\sqrt{d_1^{\mathbf{z}}}$ and $\sqrt{d_2^{\mathbf{z}}}$ from $\mathbf{z}$ respectively. But then, similarly to before, the path $\mathbf{x_i} \rightarrow \mathbf{z} \rightarrow \mathbf{x_j}$ has length

$$\sqrt{d_1^{\mathbf{z}}} + \sqrt{d_2^{\mathbf{z}}} < 2(cdist(Z) + \delta) < e^*, \quad (43)$$

which contradicts the definition of e^*.

□

We can now use these results to show the connection between the score and the critical distance for hyperplanes. The next three lemmas show that, as we reduce the smoothing parameter σ, the hyperplane with minimal score is the hyperplane with maximal critical distance (or a hyperplane with no unbalanced distances) and vice versa.

Lemma 6.6 *Let $h \in H_p$ and let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$. Then the following two statements are equivalent:*

- *There are no $\mathbf{z} \in Z$ with unbalanced distances.*
- $\forall\sigma > 0$ $score_\sigma(h) = 0$.

Proof.
"$\Rightarrow$"

Let $\sigma > 0$. Clearly if there are no $\mathbf{z} \in Z$ with unbalanced distances, then $\forall \mathbf{z} \in Z$, by lemma 4.2 $score_\sigma(\mathbf{z}) = 0$. Thus $score_\sigma(h) = 0$.
"$\Leftarrow$"

Suppose for a contradiction that $\forall \sigma > 0$ $score_\sigma(h) = 0$ but that there is some $\mathbf{z} \in Z$ with an unbalanced distance. Thus $d_1^{\mathbf{z}}$ exists. Now by lemma 4.1,

$$score_\sigma(h) \geq score_\sigma(\mathbf{z}) \geq N_\sigma \left\{ |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}} \right\}. \tag{44}$$

A simple limit argument shows us that $\exists S_{\mathbf{z}}$ s.t. $\forall \sigma \in (0, S_{\mathbf{z}})$, $N_\sigma \left\{ |\lambda_1^{\mathbf{z}}| e^{-\frac{d_1^{\mathbf{z}}}{2\sigma^2}} - n e^{-\frac{d_2^{\mathbf{z}}}{2\sigma^2}} \right\} > 0$.

□

Lemma 6.7 *Let $h^* \in H_p$ and let $Z^* = \{\mathbf{x} \in X \mid h^*(\mathbf{x}) = 0\}$. Suppose that $\forall h \in H_p$ there exists at least one point on the hyperplane $\{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$ with an unbalanced distance (i.e., there is no way to perfectly bisect the training data). Furthermore, suppose that $\forall h \in H_p$ $\exists l \in [0,1]$ s.t. $\frac{score_\sigma(h^*)}{score_\sigma(h)} \to l$ as $\sigma \xrightarrow{+} 0$. Then the critical distance of h^* exists and is maximal.*

Proof. By lemma 6.3 either there are no $\mathbf{z} \in Z^*$ with unbalanced distances or h^* has a critical point. This together with the assumptions of the lemma implies that h^* has a critical point. Thus, let h^* have a critical point $\mathbf{u}^*$ with critical distance $d_1^{\mathbf{u}^*}$. Suppose for a contradiction that $d_1^{\mathbf{u}^*}$ is not maximal. Hence $\exists h \in H_p$ with critical distance $d_1^{\mathbf{u}} > d_1^{\mathbf{u}^*}$.
But then let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$ and so

$$\frac{score_\sigma(h^*)}{score_\sigma(h)} = \frac{\max_{\mathbf{z} \in Z^*} score_\sigma(\mathbf{z})}{\max_{\mathbf{z} \in Z} score_\sigma(\mathbf{z})} = \frac{\left(\frac{\max_{\mathbf{z} \in Z^*} score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u}^*)} \right)}{\left(\frac{\max_{\mathbf{z} \in Z} score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} \right)} \cdot \frac{score_\sigma(\mathbf{u}^*)}{score_\sigma(\mathbf{u})} \tag{45}$$

$\to \left(\frac{1}{1} \cdot \infty\right) = \infty$ as $\sigma \xrightarrow{+} 0$ by lemma 5.2 and lemma 4.3 since $d_1^{\mathbf{u}} > d_1^{\mathbf{u}^*}$. So, in particular, $\exists S^* > 0$ s.t. $\forall \sigma \in (0, S^*)$ $\frac{score_\sigma(h^*)}{score_\sigma(h)} > 1$, contradicting the assumption of the lemma.

□

Lemma 6.7 tells us that by minimizing the score we maximize the critical distance. We now prove the converse.

Lemma 6.8 *Let $h^* \in H_p$ and let $Z^* = \{\mathbf{x} \in X \mid h^*(\mathbf{x}) = 0\}$. Suppose that $\forall h \in H_p$ there exists at least one point in $\{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$ with an unbalanced distance. Also, suppose that the critical distance of h^* exists and is maximal. Then for each $h \in H_p$ $\exists l \in [0,1]$ s.t. $\frac{score_\sigma(h^*)}{score_\sigma(h)} \to l$ as $\sigma \xrightarrow{+} 0$.*

Proof. Let $h \in H_p$. By the assumption of the lemma let h have a critical point $\mathbf{u}$ and let $\mathbf{u}^*$ be a critical point of h^*. Let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$

$$\frac{score_\sigma(h^*)}{score_\sigma(h)} = \frac{\left(\frac{\max_{\mathbf{z} \in Z^*} score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u}^*)} \right)}{\left(\frac{\max_{\mathbf{z} \in Z} score_\sigma(\mathbf{z})}{score_\sigma(\mathbf{u})} \right)} \cdot \frac{score_\sigma(\mathbf{u}^*)}{score_\sigma(\mathbf{u})} \tag{46}$$

By lemma 6.3 $|\lambda_1^{\mathbf{u}}| = |\lambda_1^{\mathbf{u}^*}| = 1$. Lemma 4.3 and corollary 4.5 then imply the result.

□

7 Maximum margin hyperplanes

We have shown how critical distances and scores are related. We have also shown that the score is exactly and always zero if and only if the critical distance of a hyperplane does not exist. We will now concentrate on showing how the existence and value of the critical distance of a hyperplane is related to the margin.

For the next three lemmas it will be useful to have the following definition.

Definition 7.1 *Let Z be a hyperplane in X. Suppose for each training instance $\mathbf{x_i}$ that is margin(Z^*) away from Z there is another training instance $\mathbf{x_j}$ such that Z is the perpendicular bisector of $\mathbf{x_i}$ and $\mathbf{x_j}$. More formally, suppose that for each training instance $\mathbf{x_i}$ s.t. $(\mathbf{x_i} \cdot \mathbf{w} + \mathbf{b}) =$ margin(Z), the point $\mathbf{x_i} - 2(\mathbf{x_i} \cdot \mathbf{w} + \mathbf{b})\mathbf{w}$ is also a training instance. We then say that Z* has a balanced margin.

Lemma 7.2 *Let $h^* \in H_p$. Let $Z^* = \{\mathbf{x} \in X \mid h^*(\mathbf{x}) = 0\}$ Suppose that Z^* has a balanced margin. Then Z^* is the unique hyperplane with maximal margin. Also Z^* either is the unique hyperplane with maximal critical distance or $\forall \sigma > 0$ score$_\sigma(h^*) = 0$.*

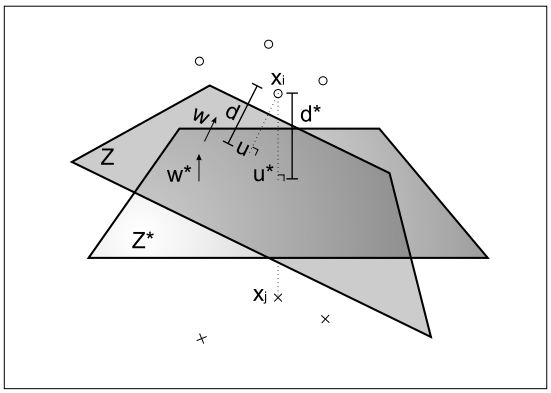
Proof. We first show that Z^* has maximal margin. Suppose that Z^* does not have maximal margin or is not unique. So there is another hyperplane Z with equal or larger margin. Since both Z^* and Z are hyperplanes we may write:

$$Z^* = \{\mathbf{x} \in X \mid \mathbf{w}^* \cdot \mathbf{x} + b^* = 0\}, \tag{47}$$

$$Z = \{\mathbf{x} \in X \mid \mathbf{w} \cdot \mathbf{x} + b = 0\}, \tag{48}$$

where $\mathbf{w}^*, \mathbf{w} \in \mathbb{R}^D$ are unit, and $b^*, b \in \mathbb{R}$.

By the assumptions of the lemma, there exists two training data instances $\mathbf{x_i}$ and $\mathbf{x_j}$ s.t. Z^* is their perpendicular bisector and $\mathbf{x_i}$ and $\mathbf{x_j}$ are a minimal distance away from Z^*. So we have the following situation:



Now the larger margin hyperplane Z must cross the line $r\mathbf{x_i} + (1-r)\mathbf{x_j}$, $r \in [0,1]$ since it must separate the data. Clearly Z cannot cross the line anywhere but at $r = \frac{1}{2}$ (which corresponds to the point $\mathbf{u}^*$ where Z^* crosses the line) since otherwise one of $\mathbf{x_i}$ or $\mathbf{x_j}$ would be closer to Z than Z^* and hence Z would not have a larger margin than Z^*.

Now $Z \neq Z^*$ so Z must pass through $\mathbf{u}^*$ at a different angle than Z^* (i.e., $\mathbf{w} \neq \pm\mathbf{w}^*$). The distance d^* from $\mathbf{x_i}$ to $\mathbf{u}^*$ is given by $d^*\mathbf{w} = (\mathbf{u}^* - \mathbf{x_i})$. Note that $|d^*|$ is the margin of Z^*. To complete this part of the proof we just need to find a point on Z that is closer to $\mathbf{x_i}$ than $|d^*|$. Let $d = (\mathbf{u}^* - \mathbf{x_i}) \cdot \mathbf{w}$. (It turns out that the distance $|d|$ is the minimum distance from $\mathbf{x_i}$ to Z although we do not have to prove this.) Consider the vector $\mathbf{u} = \mathbf{x_i} + d\mathbf{w}$.

Now, $\mathbf{u}$ lies on Z since

$$\mathbf{u} \cdot \mathbf{w} + b = \mathbf{x_i} \cdot \mathbf{w} + ((\mathbf{u}^* - \mathbf{x_i}) \cdot \mathbf{w})\, \mathbf{w} \cdot \mathbf{w} + b = \mathbf{u}^* \cdot \mathbf{w} + b = 0. \tag{49}$$

But now the squared distance from $\mathbf{x_i}$ to $\mathbf{u} \in Z$ is given by

$$(\mathbf{u} - \mathbf{x_i}) \cdot (\mathbf{u} - \mathbf{x_i}) \tag{50}$$

$$= ((\mathbf{u}^* - \mathbf{x_i}) \cdot \mathbf{w})\, \mathbf{w} \cdot ((\mathbf{u}^* - \mathbf{x_i}) \cdot \mathbf{w})\, \mathbf{w}. \tag{51}$$

Now $(\mathbf{u}^* - \mathbf{x_i}) = d^*\mathbf{w}^*$ so equation (51) becomes

$$= (d^*)^2(\mathbf{w}^*.\mathbf{w})^2, \tag{52}$$

and $\mathbf{w}^* \neq \pm\mathbf{w}$ so $(\mathbf{u} - \mathbf{x_i}) \cdot (\mathbf{u} - \mathbf{x_i}) < (d^*)^2$. But $(d^*)^2$ was the minimal squared distance a training datum was from Z^* so Z has a smaller margin than Z^*. Thus we have proved the first part of the lemma.

By lemma 6.3 and lemma 6.6 Z^* either has zero score for all $\sigma > 0$ or has a critical point. If it is the former then we are done so suppose Z^* has a critical point. Now let us consider Z. It is easy to show that the distance d for $\mathbf{u}$ must be unbalanced (essentially, if there were a training datum of the opposite class to balance $\mathbf{x_i}$, by the triangle inequality it must lie at most $2d$ away from $\mathbf{x_i}$, but the closest training datum of the opposite class lies $2d^*$ from $\mathbf{x_i}$). Thus $cdist(Z) \leq d < d^* = margin(Z^*) \leq cdist(Z^*)$. Hence Z^* must have maximal critical distance.

□

Lemma 7.3 *Let $h \in H_p$ and let $Z = \{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$. If Z has an unbalanced margin, then $cdist(Z) = margin(Z)$. .*

Proof. Since Z has an unbalanced margin, there exists an $\mathbf{x_i}$ that is $margin(Z)$ from Z and for which its reflection in Z is not another training datum. Let $\mathbf{u}^*$ be the closest point on Z to $\mathbf{x_i}$, i.e., the point on Z that is $margin(Z)$ away from $\mathbf{x_i}$. It is easy to show that $margin(Z)$ is an unbalanced distance for $\mathbf{u}^*$. This together with the general fact that $margin(Z) \leq cdist(Z)$ must mean that $cdist(Z) = margin(Z)$.

□

Lemma 7.4 *Suppose that* $\forall h \in H_p$ *there exists at least one point in* $\{\mathbf{x} \in X \mid h(\mathbf{x}) = 0\}$ *with an unbalanced distance. Let* $h^* \in H_p$ *and let* $Z^* = \{\mathbf{x} \in X \mid h^*(\mathbf{x}) = 0\}$. *Then the following two statements are equivalent:*

- Z^* *has a maximal critical distance.*
- Z^* *has a maximal margin.*

Proof. "$\Rightarrow$"
Let Z^* have maximal critical distance. If Z^* has a balanced margin then the result follows from lemma 7.2. So suppose Z^* has an unbalanced margin. We have from lemma 7.3 that $cdist(Z^*) = margin(Z^*)$.
Suppose for a contradiction that there is a different hyperplane Z with strictly greater margin.
Case: Z *has a balanced margin.* Then lemma 7.2, lemma 6.3 and the assumptions of the lemma imply that Z is the unique hyperplane with maximal critical distance which is a contradiction.
Case: Z *has an unbalanced margin.* Then we have from lemma 7.3 that that $cdist(Z) = margin(Z)$. But then we have that

$$cdist(Z) = margin(Z) > margin(Z^*) = cdist(Z^*), \tag{53}$$

which contradicts the maximality of $cdist(Z^*)$.
"$\Leftarrow$"
Proceed similarly.

□

We are now ready to prove the main theorem.

Theorem 7.5 *Let* $h^* \in H_p$. *The following two statements are equivalent:*

- *Using the convention that* $\frac{0}{0} = 1$, *for each* $h \in H_p$ $\exists l \in [0,1]$ *s.t.* $\frac{score_\sigma(h^*)}{score_\sigma(h)} \rightarrow l$ *as* $\sigma \xrightarrow{+} 0$.
- h^* *has maximal margin.*

Proof. Let $h^* \in H_p$ and let $Z^* = \{\mathbf{x} \in X \mid h^*(\mathbf{x}) = 0\}$. By lemma 6.3 either there are no $\mathbf{z} \in Z^*$ with unbalanced distances or h^* has a critical point.
"$\Rightarrow$"
Let $\exists h^* \in H_p$ s.t. for each $h \in H_p$ $\exists l \in [0,1]$ s.t. $\frac{score_\sigma(h^*)}{score_\sigma(h)} \rightarrow l$ as $\sigma \xrightarrow{+} 0$.
Case: h^* *has no unbalanced distances.* Then by lemma 7.2 Z^* has maximal margin.
Case: h^* *has a critical point.* We first show that there cannot exist a hyperplane h in H_p with no unbalanced distances. Since h^* has a critical distance we know that there is a $z \in Z^*$ with an unbalanced distance. Lemma 6.6 implies that $\neg\forall\sigma > 0$ $score_\sigma(h^*) = 0$. Thus, since we have asserted that h^* has the smallest score, there cannot exist an $h \in H_p$ which has constant 0 score for all $\sigma > 0$. Lemma 6.6 then implies that there is not a hyperplane h in H_p will no unbalanced distances.
It follows from lemma 6.7 h^* has maximal critical distance. We can then conclude from lemma 7.4 that Z^* has maximal margin.

"$\Leftarrow$"

Case: h^* has no unbalanced distances. Then by lemma 6.6 $\forall \sigma > 0$ $score_\sigma(h^*) = 0$ and hence, using the convention that $\frac{0}{0} = 1$, for each $h \in H_p$ $\exists l \in [0,1]$ s.t. $\frac{score_\sigma(h^*)}{score_\sigma(h)} \rightarrow l$ as $\sigma \xrightarrow{+} 0$.

Case: h^* has a critical point. We first show that there cannot exist a hyperplane h in H_p will no unbalanced distances. If there were an $h \in H_p$ with no unbalanced distances then lemma 7.2 would imply that h was the unique hyperplane with maximal margin and so $h = h^*$. But h^* is the hyperplane with maximal margin and lemma 6.3 tells us that a hyperplane cannot both have a critical point and no unbalanced distances.

Lemma 7.4 allows us to deduce that h^* has maximal critical distance. By applying lemma 6.8 it then follows that for each $h \in H_p$ $\exists l \in [0,1]$ s.t. $\frac{score_\sigma(h^*)}{score_\sigma(h)} \rightarrow l$ as $\sigma \rightarrow 0$.

□

8 Conclusions and Future Work

So what have we shown? Restricting the hypothesis space to hyperplanes, we have shown that, as we reduce the smoothing parameter, a hyperplane is the best fit to the Bayes optimal decision boundary if and only if it is the hyperplane that maximizes the margin.

Among other results we have also introduced the notion of a critical distance and have shown that, under two conditions, the fit to the Bayes optimal classifier is entirely determined by the score of the hypothesis at the critical point (if it exists). Extensions to hypothesis spaces other than hyperplanes are also being explored.

The notion of the score of a hyperplane, although natural, is rather arbitrary. We are investigating whether other scoring methods still lead to maximal margin hyperplanes as the lowest scoring hyperplanes.

Maximal margin hyperplanes arise as best approximations to the Bayes Optimal Classifier as we reduce the smoothing parameter to zero. As we reduce the smoothing parameter we increase the statistical variance of the Bayes optimal classifier. It is not clear that reducing the smoothing parameter all the way to zero is the optimal action to take. Among other things we are currently investigating the results of using different sizes of smoothing parameters. The larger the smoothing parameter the larger the influence data further from the hyperplane has in determining the lowest scoring hyperplane. Thus, this approach may be related to the notion of a *margin distribution* [2], which considers the entire set of distances between the decision boundary and the data, and not just the distance to the closest point.

References

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

[2] A. Grove and D. Schuurmans. Boosting in the limit: Maximizing the margin of learned ensembles. In *Proc. 15th National Conference on Artificial Intelligence (AAAI)*, 1998.

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

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

[5] V. Vapnik. *Estimation of Dependences Based on Empirical Data.* Nauka, Moscow, 1979. In Russian. English Translation: Springer Verlag, New York, 1982.