EECS 553: Machine Learning (ECE) **University of Michigan**

Separating Hyperplanes

Winter 2023 **Clayton Scott**

We have previously seen two examples of linear classifiers, LDA and logistic regression. Those were based on probabilistic models for the data generating distribution. In this section we look at a different approach to learning a linear classifier based on geometric principles. These notes will culminate in a description of the *optimal soft-margin hyperplane*, which is a special case of a more general classifier that we will study later called the *support vector machine.*

1 Vapnik's Maxim

There is a mantra in machine learning attributed to Vladimir Vapnik, a pioneer of machine learning:

> *"Don't solve a harder problem than you have to."*

Plug-in methods require estimation of (conditional) densities or mass functions, which can be more difficult than estimating a decision boundary.

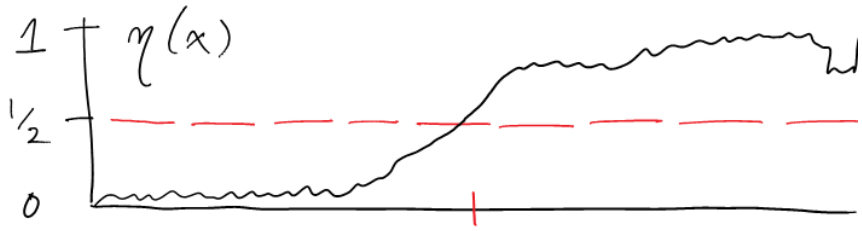

Figure 1: $\eta(x)$ is quite complicated but the decision regions are simple and η is smooth near $1/2$.

In these notes we will look at a method for linear classification that estimates the classifier more directly.

2 Hyperplanes

A *hyperplane* is a subset of $\mathbb{R}^d$ of the form

$$\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\} \tag{1}$$

for some $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$. See Figure 2 for an illustration when $d = 3$. In general, a hyperplane is an *affine subspace* of dimension $d - 1$.

Remark 1. An affine subspace is essentially a conventional linear subspace, but where all elements have been shifted by the same vector. An affine subspace is not closed under vector addition and scalar multiplication unless it contains the origin, in which case it is a linear subspace.

The vector $\boldsymbol{w}$ is orthogonal to the hyperplane it defines, which means it is orthogonal to any vector that is parallel to the hyperplane. If $\boldsymbol{v}$ is a vector that lies parallel to the hyperplane, we can write $\boldsymbol{v} = \boldsymbol{x} - \boldsymbol{x}'$ for two points $\boldsymbol{x}$, $\boldsymbol{x}'$ on the hyperplane. Thus

$$\boldsymbol{w}^T\boldsymbol{v} = \boldsymbol{w}^T(\boldsymbol{x} - \boldsymbol{x}') = -b - (-b) = 0. \tag{2}$$

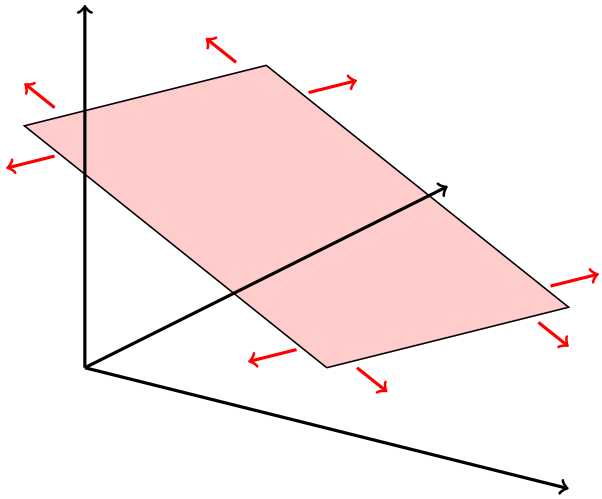

Figure 2: Hyperplane in $d = 3$

We say that $\boldsymbol{w}$ is a *normal*[1] vector. See Figure 3 for an illustration.

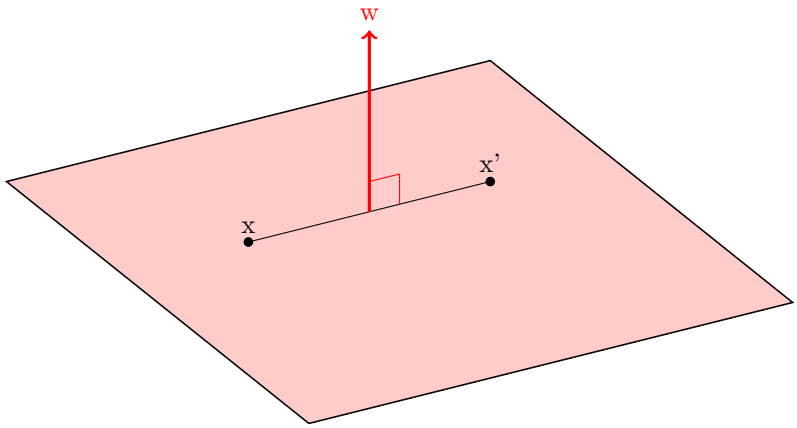

Figure 3: Normal vector in $d = 3$

Given a hyperplane $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$ and a point $\boldsymbol{z} \notin \mathcal{H}$, what is the distance of $\boldsymbol{z}$ to $\mathcal{H}$? We can write $\boldsymbol{z}$ as

$$\boldsymbol{z} = \boldsymbol{z}_0 + r \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|} \tag{3}$$

for unique $\boldsymbol{z}_0 \in \mathcal{H}$ and $r \in \mathbb{R}$ (note r may be negative). See Figure 4.

Then

$$\begin{aligned} \boldsymbol{w}^T\boldsymbol{z} + b &= \boldsymbol{w}^T\boldsymbol{z}_0 + \boldsymbol{w}^T\left(r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}\right) + b \\ &= r\|\boldsymbol{w}\| \end{aligned} \tag{4}$$

[1]The term *normal vector* is also used to describe a vector with unit length.

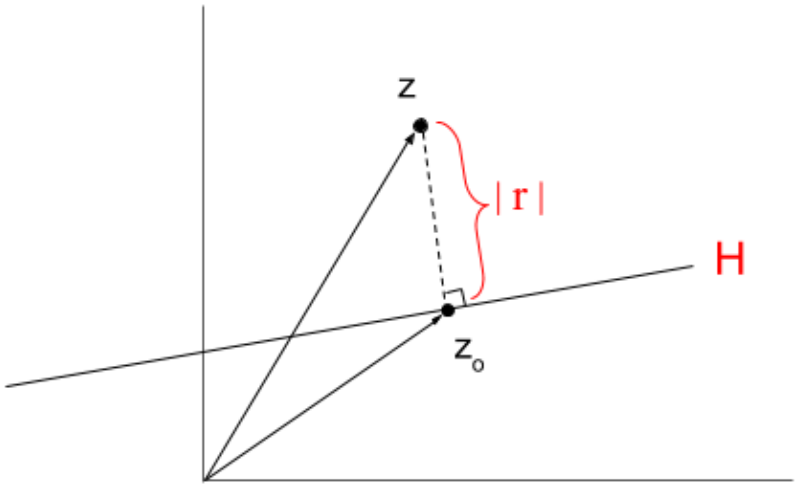

Figure 4: Distance from $\boldsymbol{z}$ to $\mathcal{H}$ in $d = 2$ dimennsions.

because $\boldsymbol{w}^T\boldsymbol{z}_0 + b = 0$ and $\boldsymbol{w}^T\boldsymbol{w} = \|\boldsymbol{w}\|^2$. Hence

$$|r| = \frac{|\boldsymbol{w}^T\boldsymbol{z} + b|}{\|\boldsymbol{w}\|} \tag{5}$$

is the distance from $\boldsymbol{z}$ to $\mathcal{H}$.

3 Seperating Hyperplanes

Consider binary classification with labels $y \in \{-1, 1\}$. Recall that a linear classifier has the form

$$f(\boldsymbol{x}) = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$$

for some $\boldsymbol{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. A simple observation that will be important below is that the classifier does not change if we scale $\boldsymbol{w}$ and b by the same positive scalar. That is, for all $\alpha > 0$ and $\boldsymbol{x} \in \mathbb{R}^d$,

$$\text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b) = \text{sign}((\alpha\boldsymbol{w})^T\boldsymbol{x} + (\alpha b)).$$

Let $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)$ be training data for a binary classification problem. We say the training data are *linearly separable* if there exist $\boldsymbol{w} \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0 \qquad \forall i = 1, \ldots, n.$$

In this case we refer to

$$\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$$

as a *separating hyperplane*. An example of linearly separable data is shown in Figure 5.

Some separating hyperplanes make for better classifiers than others. In the next section we will look at one approach to finding a good separating hyperplane.

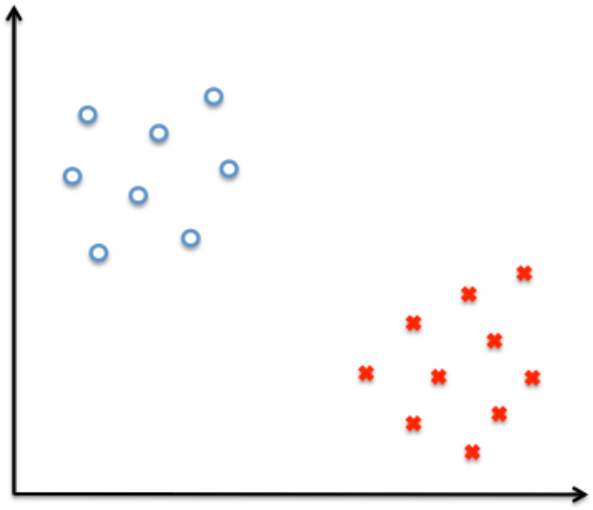

Figure 5: An example data distribution. Blue circles and red crosses represent data points from two different classes. Are all separating hyperlanes equally good in this case?

4 The Maximum Margin Hyperplane

In this section let's assume the training data are linearly separable. The *margin* ρ of a separating hyperplane is the distance from the hyperplane to the nearest training point:

$$\rho(\boldsymbol{w}, b) \coloneqq \min_{i=1,\ldots,n} \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|}. \tag{6}$$

The *maximum margin* or *optimal* separating hyperplane is the solution of

$$\begin{aligned} &\max_{\boldsymbol{w},b} \ \rho(\boldsymbol{w}, b) \\ &\text{s.t. } y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) > 0 \quad \forall i. \end{aligned} \tag{7}$$

The objective of this maximization problem is the margin, and the constraint ensures that the hyperplane is a separating hyperplane. An example of a maximum margin hyperplane is shown in Figure 6. The basic intuition of the maximum margin criterion is that the training data points are noisy, and the maximum margin criterion tries to insulate the separating hyperplane from that noise as much as possible. A key feature of the maximum margin hyperplane is that it is the same distance from the nearest point in each class. If this were not the case, the hyperplane could be shifted to increase the margin.

We would like to be able to compute the optimal separating hyperplane efficiently. To accomplish this, we will reformulate the initial formulation into one called a *quadratic program*, which is a standard type of optimization problem for which efficient solvers can be readily found. This quadratric problem's solution will be the optimal separating hyperplane.

A separating hyperplane is said to be in *canonical form* if $\boldsymbol{w}$ and b are such that

$$\begin{aligned} &y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1 \quad \forall i, \\ &y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 \quad \text{for some } i. \end{aligned}$$

Every separating hyperplane can be expressed in canonical form. To see this, suppose $\mathcal{H} = \{\boldsymbol{x} : \boldsymbol{w}_1^T \boldsymbol{x} + b_1 = 0\}$ is a separating hyperplane (not necessarily in canonical form). Then we may simply rescale $\boldsymbol{w}_1$ and b_1 by

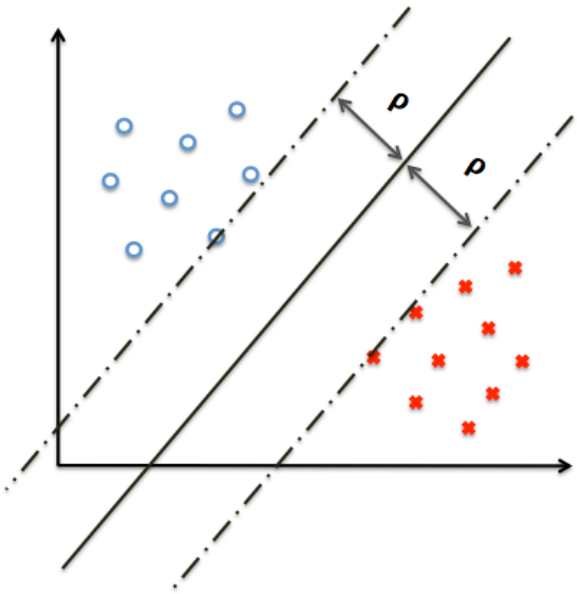

Figure 6: An example of a Maximum Margin Hyperplane. Blue circles and red crosses are two data classes, solid line in the middle is the Maximum Margin Hyperplane and ρ is the margin.

a positive number (which does not change the classifier), producing new parameters $\boldsymbol{w}_2$ and b_2, such that the minimum value of $|\boldsymbol{w}_2^T \boldsymbol{x}_i + b_2|$ is 1. Canonical form provides a unique representation of every separating hyperplane.

This allows us to write the max-margin separating hyperplane as the solution of

$$\begin{aligned} \max_{\boldsymbol{w},b} \ & \min_{i=1,\ldots,n} \frac{|\boldsymbol{w}^T \boldsymbol{x}_i + b|}{\|\boldsymbol{w}\|} \\ \text{s.t. } & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, \quad \forall i, \\ & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1, \quad \exists i. \end{aligned} \tag{8}$$

From the constraints we know that for each i, $|\boldsymbol{w}^T \boldsymbol{x}_i + b| = y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1$, and we also know that for some i, $|\boldsymbol{w}^T \boldsymbol{x}_i + b| = y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$. Therefore, the inner minimization is simply $1/\|\boldsymbol{w}\|$, and we can simplify the above problem to

$$\begin{aligned} \max_{\boldsymbol{w},b} \ & \frac{1}{\|\boldsymbol{w}\|} \\ \text{s.t. } & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \geq 1, \quad \forall i, \\ & y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1, \quad \exists i. \end{aligned} \tag{9}$$

This problem can be further simplified. In particular, we can drop the last constraint because it will automatically be satisfied by the optimizer. To see this, suppose $(\boldsymbol{w}^*, b^*)$ solves (9) and $y_i(\boldsymbol{w}^{*T} \boldsymbol{x}_i + b^*) > 1$, $\forall i$. Then we can shrink the value of $\|\boldsymbol{w}\|$ without violating the constrains, which contradicts the assumed optimality of $\boldsymbol{w}^*, b^*$.

Finally we have

$$\begin{aligned} &\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 \\ &\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, \quad \forall i = 1, \ldots, n. \end{aligned} \tag{10}$$

This is an example of a *constrained optimization problem*, and in particular it is called a *quadratic program*, which is a constrained optimization problem with quadratic objective and linear constraints.

5 Optimal Soft-Margin Linear Classifier

To accommodate nonseparable data, we modify the above QP (quadratic program) by introducing *slack variables* $\xi_1, \ldots, \xi_n \geq 0$. Denoting $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_n)^T$, *optimal soft-margin linear classifier* is the classifier $\text{sign}((\boldsymbol{w}^*)^T\boldsymbol{x} + b^*)$ where $\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*$ solves

$$\begin{aligned} &\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ &\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, \quad \forall i \\ &\qquad \xi_i \geq 0, \quad \forall i. \end{aligned} \tag{11}$$

Notice that this optimization problem differs from the previous one only through the slack variables. The point of the slack variables is that they allow some data points to violate the margin-of-separation requirement, as reflected by the first constraint. The second term in the objective reflects the fact that we want as few data points to violate the margin separation as possible.

This optimization program is another QP. The parameter $C > 0$ is an example of a *tuning parameter*, which is a parameter that is not specified by a learning algorithm, and must be set by the user or some other procedure (such as cross validation) that is external to the learning algorithm. C controls the penalty on data points violating the margin of separation. If C is larger, the algorithm tries harder to correctly classifier as many data points as possible.

6 Looking Ahead

In the future we will show that the optimal soft margin hyperplane can be rederived from a statistical perspective known as empirical risk minimization. This perspective will explain how C can be thought of as the inverse of a regularization parameter. We will also see that the optimal soft-margin hyperplane is a special case of a nonlinear classifier called the *support vector machine*.

Exercise

1. (★) Let $\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^d \,|\, \boldsymbol{w}^T\boldsymbol{x} + b = 0\}$. Argue that scaling both $\boldsymbol{w}$ and b by the same nonzero constant doesn't change $\mathcal{H}$.

2. (★) Let $\mathcal{H} = \{\boldsymbol{x} \in \mathbb{R}^3 \,|\, x_1 - 2x_2 + 3x_3 - 4 = 0\}$. What is the distance from $\mathcal{H}$ to $\boldsymbol{z} = [1\ 1\ 1]^T$?

3. (☆☆) Consider a maximum margin separating hyperplane in 2 dimensions. What is the number of data points whose distance to the hyperplane is exactly the margin $\rho(w, b)$? Describe various cases depending on the data, as needed.

4. (☆☆) Let $\mathcal{H}_1 = \{\boldsymbol{x} \,|\, \boldsymbol{w}_1^T\boldsymbol{x} + b_1 = 0\}$ and $\mathcal{H}_2 = \{\boldsymbol{x} \,|\, \boldsymbol{w}_2^T\boldsymbol{x} + b_2 = 0\}$. Determine the angle between the $\mathcal{H}_1$ and $\mathcal{H}_2$ as a function of the hyperplane parameters.

5. (★) What is the impact of C in the optimal soft-margin hyperplane? Consider the case where outliers are present in the training data.

6. (★) Argue that if $\boldsymbol{x}_i$ is misclassified by the OSM hyperplane, then $\xi_i \geq 1$.

7. (★★) Show that if $\xi_i^* > 0$, then ξ_i^* is proportional to the distance from $\boldsymbol{x}_i$ to the margin hyperplane associated with class y_i (that is, the set $\{\boldsymbol{x} : (\boldsymbol{w}^*)^T\boldsymbol{x} + b^* = y_i\}$), and give the constant of proportionality.

8. (☆☆) Consider the optimization problem

$$\begin{aligned} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \ & \frac{1}{2}\|\boldsymbol{w}\|^2 + \frac{D}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t. } & y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq \kappa - \xi_i && \forall i = 1, \ldots, n \\ & \xi_i \geq 0 && \forall i = 1, \ldots, n \end{aligned}$$

   where $D > 0$, $\kappa > 0$. Show that the above quadratic program yields the same classifier as the optimal soft-margin classifier for a certain parameter C, and express this C in terms of D and κ.

9. (☆) Look up other criteria (besides maximizing the margin) that can be used to define a linear classifier (for separable or nonseparable data) using geometric principles.

10. (★) How else could the max-margin separating hyperplane classifier be extended to non-separable data?

11. (★★) Can the max-margin hyperplane classifier be obtained as a special case of the optimal soft-margin linear classifier by setting C to be a specific value?