SEPARATING HYPERPLANES

Vapnik's Maxim

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

"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

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

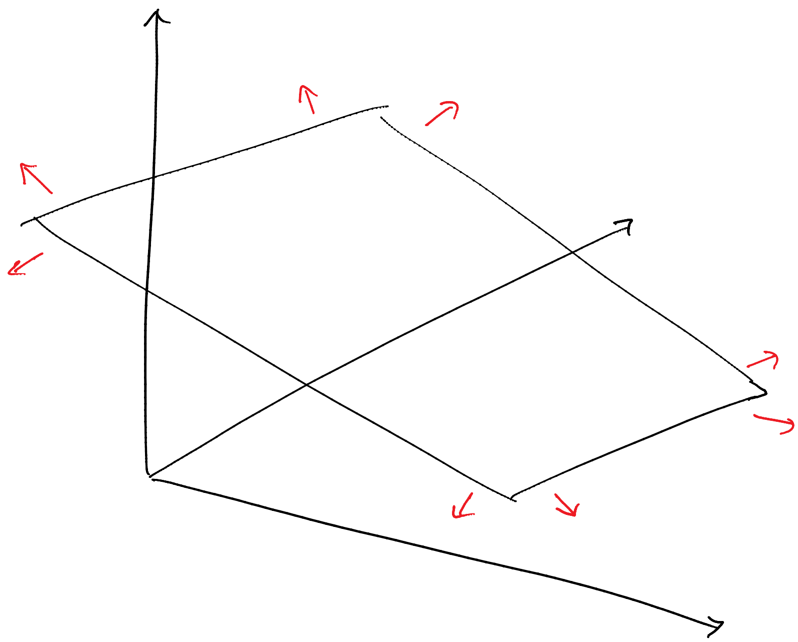
In these notes we'll look at a method for linear classification that estimates the classifier more directly.

Hyperplanes

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

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

for some $w \in \mathbb{R}^d$, $b \in \mathbb{R}$. When $d=3$, we have this picture:

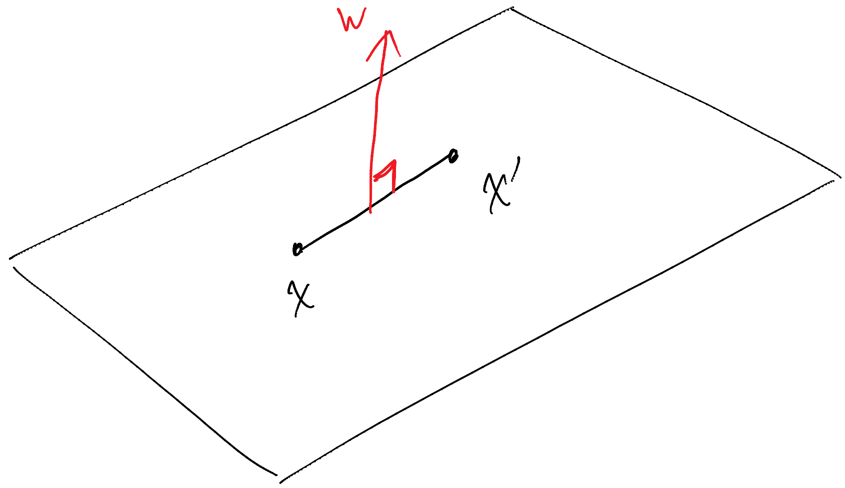


In general, a hyperplane is an affine subspace of dimension $d-1$

The vector w is orthogonal to the hyperplane. If v is a vector that lies parallel to the hyperplane, we can write $v = x - x'$ for two points x, x' on the hyperplane. Thus

$$w^T v = w^T(x - x') = -b - (-b) = 0.$$

We call w a <u>normal</u> vector.

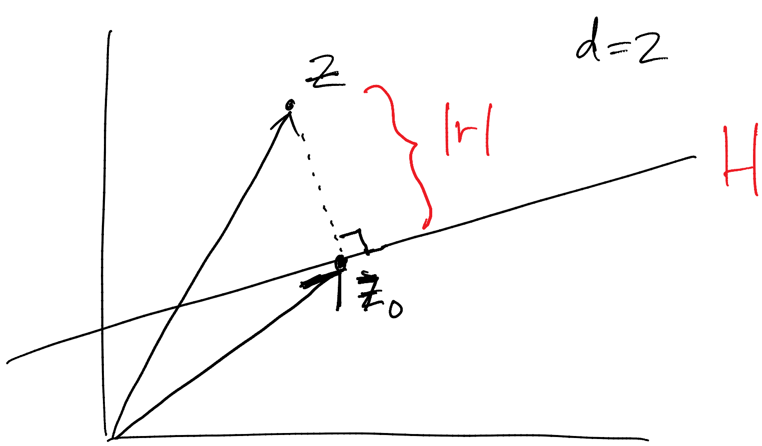


Given a hyperplane $H = \{ x : w^T x + b \}$ and a point $z \notin H$, what is the distance of z to H?

We can write z as

$$z = z_0 + r \cdot \frac{w}{\|w\|}$$

for unique $z_0 \in H$ and $r \in \mathbb{R}$ (note r may be negative).



Then

$$w^T z + b = w^T z_0 + w^T\left(r \frac{w}{\|w\|}\right) + b$$

$$= r \cdot \|w\| \qquad \left[w^T z_0 + b = 0 \right]$$

Hence

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

$$\boxed{\text{Separating Hyperplanes}}$$

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be training data for a binary classification problem, and assume $y_i \in \{-1, 1\}$.

We say the training data are <u>linearly separable</u> if there exist $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

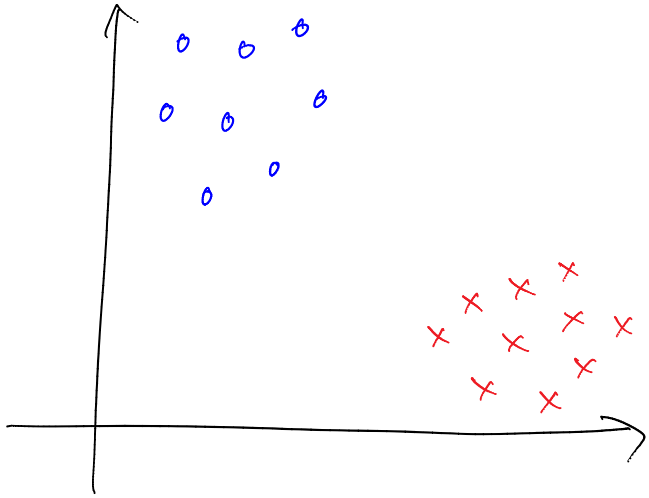
$$y_i (w^T x_i + b) > 0 \qquad \forall i = 1, \ldots, n.$$

In this case we refer to

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

as a <u>separating hyperplane</u>.

Are all separating hyperplanes equally good?



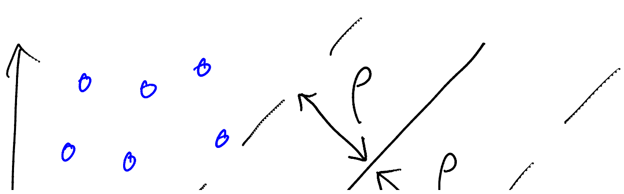
The Maximum Margin Hyperplane

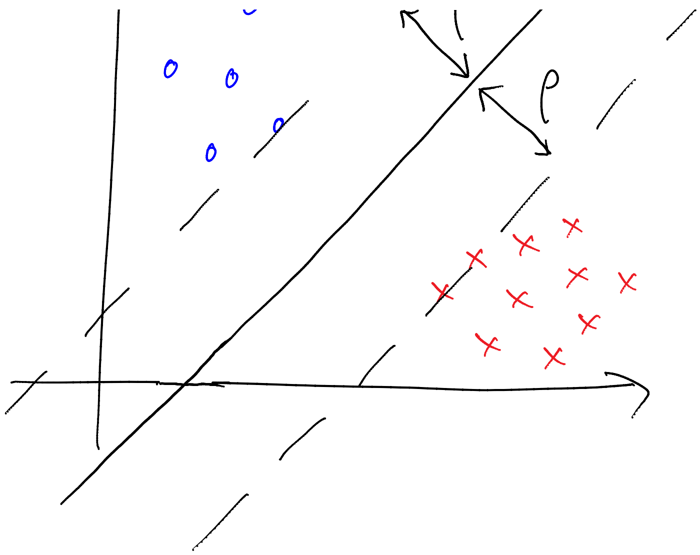
The margin ρ of a separating hyperplane is the distance from the hyperplane to the nearest training point:

$$\rho(w, b) := \min_{i=1,\dots,n} \frac{|w^T z + b|}{\|w\|}$$

The maximum margin or optimal separating hyperplane is the solution of

$$\max_{w,b} \ \rho(w,b)$$
$$\text{s.t.} \quad y_i(w^T x_i + b) > 0 \quad \forall i$$





A separating hyperplane is said to be in <u>canonical form</u> if w and b are such that

$$y_i(w^T x_i + b) \geq 1 \qquad \forall i$$

$$y_i(w^T x_i + b) = 1 \qquad \text{for some } i$$

Every separating hyperplane can be expressed in canonical form. Suppose $H = \{x : w_1^T x + b_1 = 0\}$ is a separating hyperplane (not necessarily in canonical form). Let

$$m := \min_{i=1,\dots,n} |w_1^T x_i + b_1|$$

and define

$$w_2 = \frac{w_1}{m}, \quad b_2 = \frac{b_1}{m}.$$

$$w_2 = \frac{w_1}{m}, \quad b_2 = \frac{\quad}{m}.$$

Then w_2, b_2 express H in canonical form.

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

$$\max_{w,b} \frac{|w^T x_i + b|}{\|w\|}$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 \quad \forall i$$

$$y_i(w^T x_i + b) = 1 \quad \text{for some } i.$$

How can this be simplified?

$$\max_{w,b} \frac{1}{\|w\|}$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 \quad \forall i$$

$$y_i(w^T x_i + b) = 1 \quad \text{for some } i.$$

Can we simplify further? Yes, we can drop the second constraint because it will automatically be satisfied by the optimizer. So we finally have

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 \quad \forall\, i = 1, \ldots, n$$

This is an example of a constrained optimization problem, and in particular it is called a quadratic program (quadratic objective, linear constraints)

Optimal Soft-Margin Hyperplane

To accomodate nonseparable data, we modify the above QP by introducing slack variables $\xi_1, \ldots, \xi_n \geq 0$. This leads to the so-called optimal soft-margin hyperplane, where $\xi = (\xi_1, \ldots, \xi_n)^T$:

$$\min_{w,b,\xi} \quad \frac{1}{2}\|w\|^2 + \frac{C}{n}\sum_{i=1}^{n} \xi_i$$

$$\text{s.t.} \quad y_i(w^T x_i + b) \geq 1 - \xi_i, \quad \forall i$$

$$\xi_i \geq 0, \quad \forall i$$

$\xi_i \geq 0, \quad \forall i$

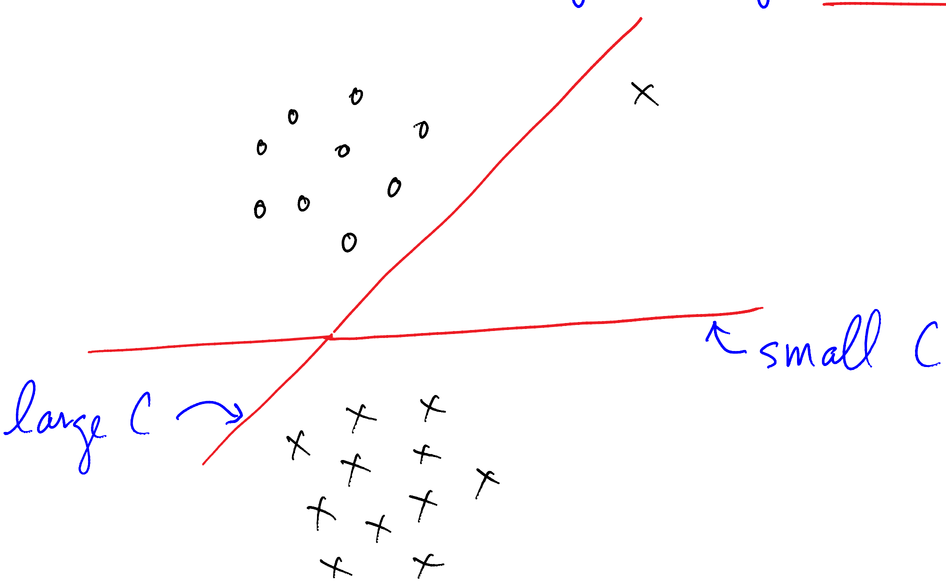
Remarks

- This is another QP
- $C > 0$ is a user-specified tuning parameter. Later we'll discuss how to select it.
- If x_i is misclassified, then $\xi_i \geq 1$. Therefore

$$\frac{1}{n}\sum \xi_i \geq \frac{1}{n}\sum 1_{\{y_i \neq \text{sign}(w^T x_i + b)\}}$$

$$= \underline{\text{training error}}$$

- C controls the influence of outliers



- The optimal soft-margin hyperplane is a special case of the more general support vector machine which we'll study later.

- We'll discuss algorithms to solve the QP in the future.