Optimal Margins

The General Case

- Given an $N \times k$ feature matrix X in "tidy" format as usual.
- We also have an $N \times 1$ vector Y whose entries are ± 1. Y distinguishes the two classes in the data.
- The goal is to predict Y given X.

Some geometry: hyperplanes

An (affine) hyperplane in $\mathbb{R}^k$ is given by an equation

$$f(x_1, \dots, x_k) = 0$$

where $f(x_1, \dots, x_k)$ is a degree 1 polynomial

$$f(x_1, \dots, x_k) = w_1 x_1 + w_2 x_2 + \cdots + w_k x_k + b \tag{1}$$

Better: We write eq. 1 by giving a non-zero vector

$$w = (w_1, \dots, w_k) \in \mathbb{R}^k$$

and a constant b so that

$$f(x) = w \cdot x + b$$

for $x \in \mathbb{R}^k$.

Hyperplanes: key facts

Given $w \in \mathbb{R}^k$ and $b \in \mathbb{R}$, let $f(x) = w \cdot x + b$. Then

- The inequalities $f(x) > 0$ and $f(x) < 0$ divide up $\mathbb{R}^k$ into half spaces.

- The vector w is normal to the hyperplane $f(x) = 0$.

- The (perpendicular) distance D from a point $p = (u_1, \ldots, u_k)$ to the hyperplane $f(x) = 0$ is

$$D = \frac{f(p)}{\|w\|}$$

Linear separability

We think of our data as a family of points in $\mathbb{R}^k$; each point has coordinates given by a row of the data matrix X.

Definition: Our data (given as an $N \times k$ data matrix X and a label vector Y) is linearly separable if there is a vector $w \in \mathbb{R}^k$ and a constant $b \in \mathbb{R}$ so that

$$f(x) = w \cdot x + b > 0$$

when x is a row of X corresponding to a y-value of $+1$, and

$$f(x) = w \cdot x + b < 0$$

when x is a row of X corresponding to to a y value of -1. In this case $f(x) = w \cdot x + b = 0$ is called a *separating hyperplane* for the data.

Criteria for linear separability

How can we tell if our data is linearly separable?

Let A^+ be the set of points in $\mathbb{R}^k$ with label $+1$ and A^- the set of points with label -1. Can we find $w \in \mathbb{R}^k$ and $b \in R$ so that

$$w \cdot x + b > 0 \text{ for all } x \in A^+$$

and

$$w \cdot x + b < 0 \text{ for all } x \in A^-?$$

Proposition: A^+ and A^- are linearly separable if there is a $w \in \mathbb{R}^k$ so that

$$\max_{x \in A^-} w \cdot x < \min_{x \in A^+} w \cdot B \tag{2}$$

More on linear separability

Let

$$B^{-}(w) = \max_{x \in A^{-}} w \cdot x$$

and

$$B^{+}(w) = \min_{x \in A^{+}} w \cdot B$$

So our sets $A^{\pm}$ are linearly separable if $B^{-} < B^{+}$ and in this case any $-b$ between this two values gives a separating hyperplane $f(x) = w \cdot x + b = 0$.

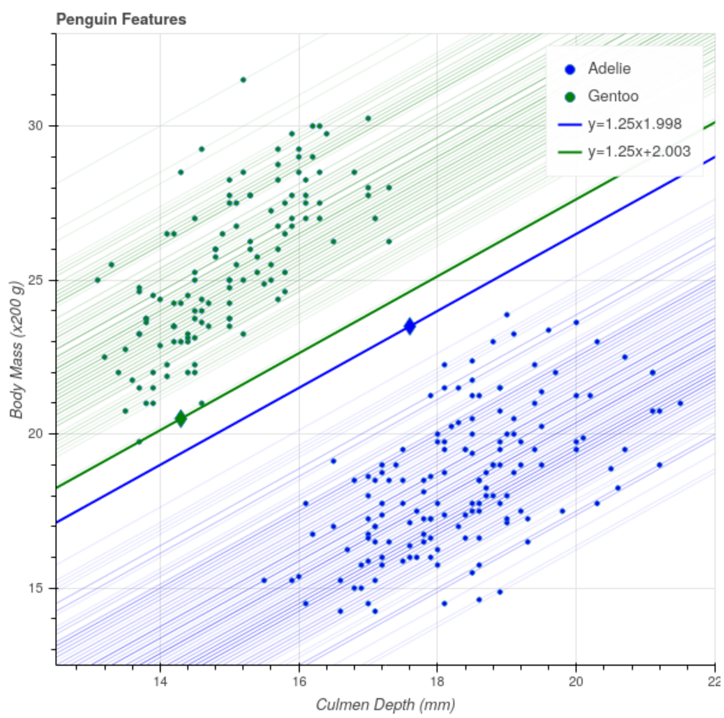
Supporting hyperplanes and geometric margins

A different point of view:

- Let $f_i^+(x) = w \cdot x - w \cdot x_i$ for $x_i \in A^+$.

- Let $f_i^-(x) = w \cdot x - w \cdot x_i$ for $x_i \in A^-$.

Then $f_i^{\pm}(x) = 0$ is a family of hyperplanes parallel to w through the points in $A^{\pm}$.



Margin

Choose points:

- $x^+ \in A^+$ so that $B^+ = \min_{x \in A^+} w \cdot x = w \cdot x^+$.
- $x^- \in A^-$ so that $B^- = \max_{x \in A^-} w \cdot x = w \cdot x^-$.

Supporting hyperplanes

Definition: Let A be a set of points in $\mathbb{R}^k$. A hyperplane $f(x) = w \cdot x + b = 0$ is a *supporting hyperplane* for A if:

- $f(x) \geq 0$ for all $x \in A$ and there exists at least one $x \in A$ with $f(x) = 0$, or
- $f(x) \leq 0$ for all $x \in A$ and there exists at least one $x \in A$ with $f(x) = 0$.

Geometric margin

Definition: The perpendicular distance between $f^+(x) = 0$ and $f^-(x) = 0$ is called the *geometric margin* between $A^\pm$ in the direction perpendicular to w.

$$\tau_w = \frac{w \cdot (x^+ - x^-)}{\|w\|} = \frac{B^+(w) - B^-(w)}{\|w\|}$$

The best separating hyperplane in the w direction runs halfway between the two supporting hyperplanes:

$$f(x) = w \cdot x - \frac{B^+(w) + B^-(w)}{2}$$

The optimal margin problem

Definition: The *optimal margin* $\tau(A^+, A^-)$ between A^+ and A^- is the largest value of τ_w as w varies over vectors in $\mathbb{R}^k$ such that $B^-(w) < B^+(w)$:

$$\tau(A^+, A^-) = \max_w \tau_w(A^+, A^-) = \max_w \frac{B^+(w) + B^-(w)}{\|w\|}$$

If w gives this maximum value, then the *optimal margin classifying hyperplane* is the hyperplane

$$f(x) = w \cdot x - \frac{B^+(w) - B^-(w)}{2}$$

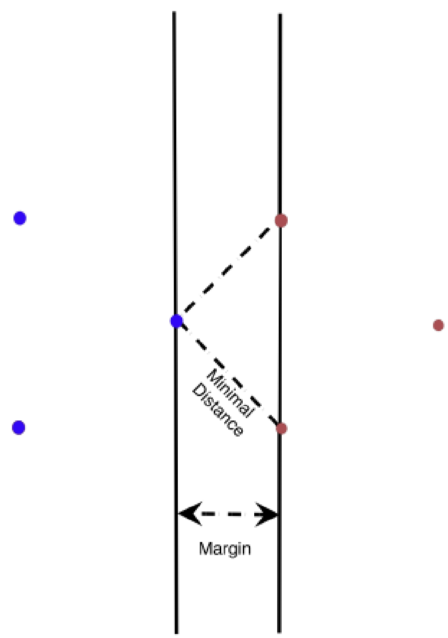
that runs "down the middle" between the two supporting hyperplanes $f^-(x) = w \cdot x - B^-(w) = 0$ and $f^+(x) = w \cdot x - B^+(w) = 0$.

Closest points and optimal margin

One might think that the optimal margin is the closest distance between the sets A^+ and A^-, but that isn't true.

Proposition: The closest distance between points in A^+ and A^- is greater than or equal to the optimal margin:

$$\min_{p \in A^+, q \in A^-} \|p - q\| \geq \tau(A^+, A^-).$$



Counterexample