SEPARATING HYPERPLANES

LDA and logistic regression are "plug-in" methods for linear classification. They make assumptions about the distribution of the data, and reduce classification to

Ⓐ ________ / ________ estimation.

In these notes we'll discuss an approach to linear classification that

1. makes no distributional assumptions
2. does not require solving an intermediate (and potentially more difficult) problem.

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be training data,

$x_i \in \mathbb{R}^d$, $y_i \in \{-1, +1\}$

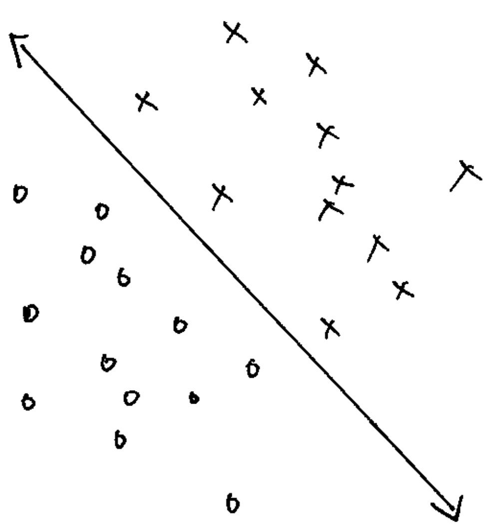
<u>Definition</u> We say the data are <u>linearly separable</u> if there exists $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ such that

$$y_i = \text{sign}\{ w^T x_i + b \}$$

for $i = 1, \ldots, n$. We refer to

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

B as a ______________ ______________ .



Assume for now that the data are linearly separable. How can we find a separating hyperplane?

Geometry

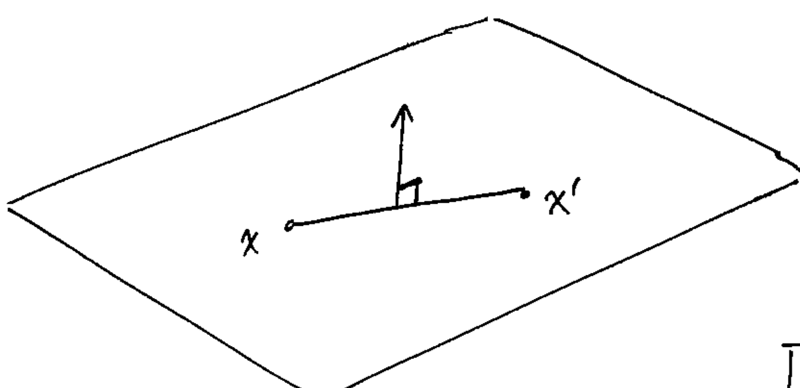
Let w, b define a hyperplane.
If x, x' are points on the hyperplane, then

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

$$=$$

©

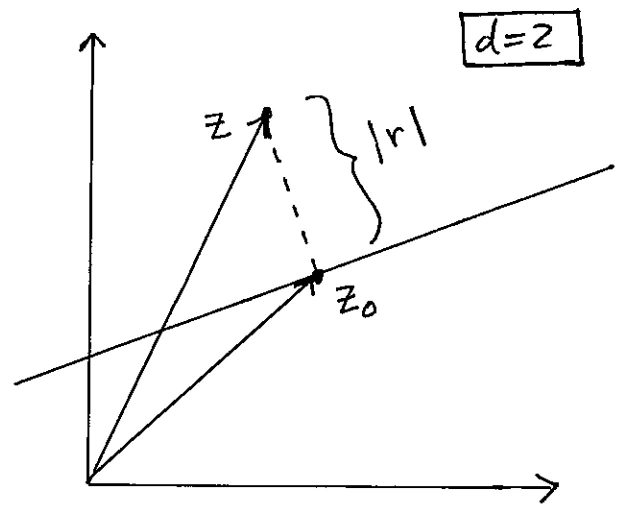
Hence w is ____________ to all vectors that are ____________ to the hyperplane



$\boxed{d=3}$

(D) We call $\frac{w}{\|w\|}$ the ________ vector to the hyperplane. It is unique up to its ______.

Question | Let $z \in \mathbb{R}^d$. How far is z from $\{x \in \mathbb{R}^d : w^T x + b = 0\}$?



Answer | Write

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

where $w^T z_0 + b = 0$

and r may be negative.

Then

(E)
$$w^T z + b =$$

$$=$$

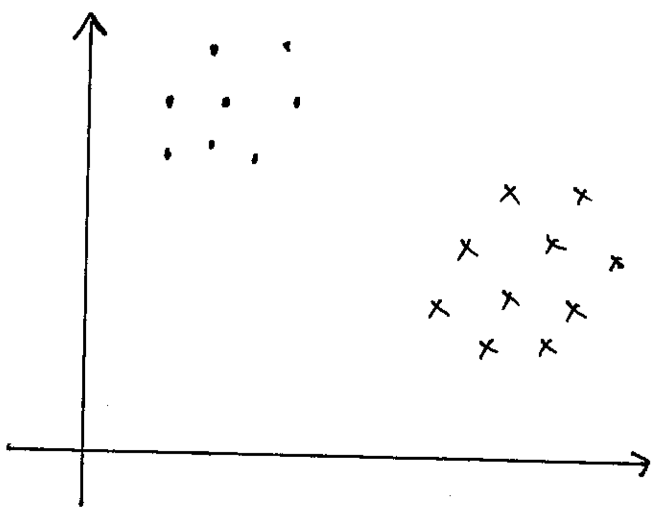
$$=$$

$$\Longrightarrow |r| =$$

We refer to r as the "signed distance" to the hyperplane

The Maximum Margin Hyperplane

Are all separating hyperplanes equally good?



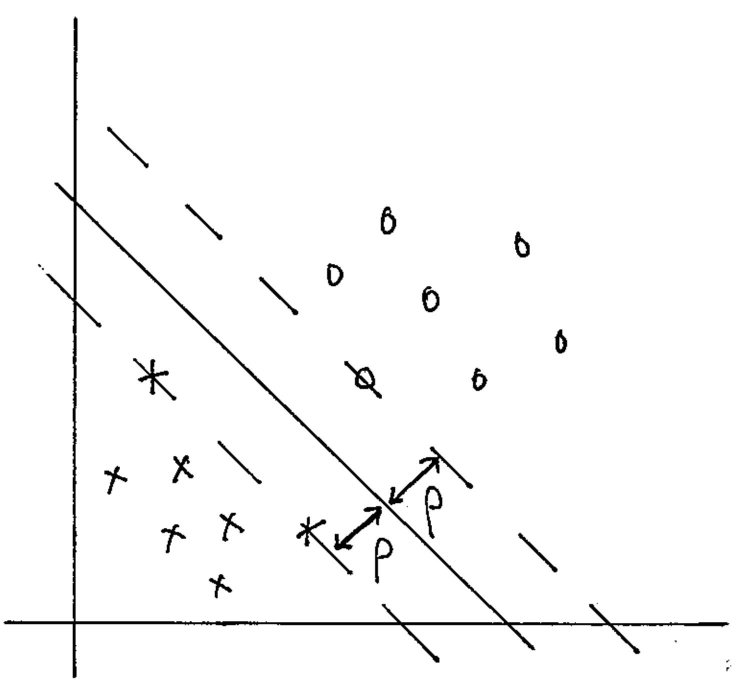
Definitions

1. The <u>margin</u> ρ of a separating hyperplane is the distance from the hyperplane to the closest x_i

(F) $$\rho(w, b) :=$$

2. The <u>maximum margin</u> or <u>optimal</u> separating hyperplane is the solution of

$$(w^*, b^*) = \arg\max_{w,b} \rho(w, b)$$



larger margin
$\Longrightarrow$ better generalization

Canonical Form

We may rescale any separating hyperplane

(G) so that it is in ____________ _______:

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

Exercise | Express the margin of a hyperplane in canonical form as a function of w and b. Express w^*, b^* as the solution of a constrained optimization problem.

Solution

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

The optimal separating hyperplane is therefore the solution of

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

☆

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

Terminology

H

- ☆ is an example of a __________ __________.
- Those x_i such that $y_i(w^T x_i + b) = 1$ are called __________ __________.

Optimal Soft-Margin Hyperplane

Real data is often not linearly separable.

To accomodate nonseparable data, we modify the QP by introducing ________

I ________ $\xi_1, \dots, \xi_n \geq 0$

This results in the optimal soft-margin hyperplane:

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

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

$$\xi_i \geq 0, \quad i = 1, \dots, n.$$

Remarks

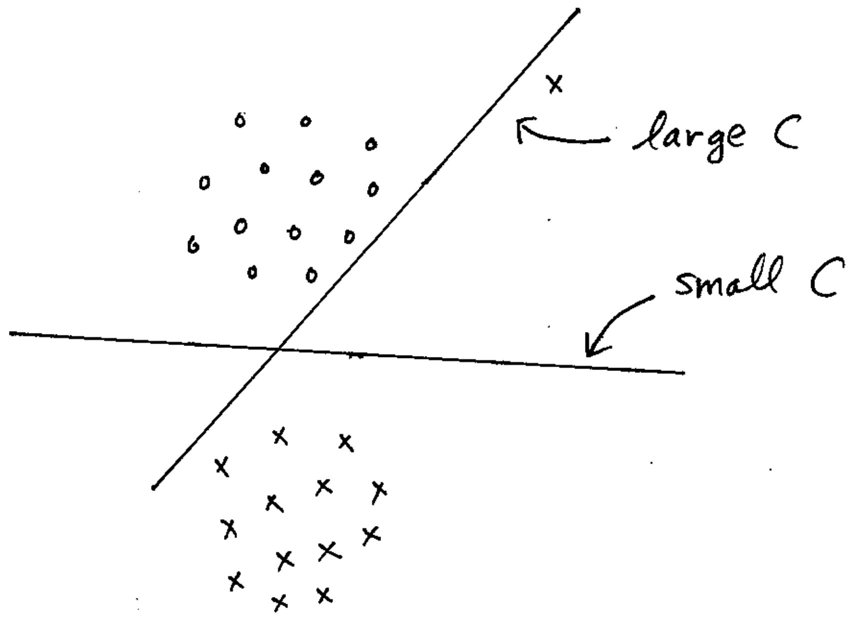
- This is another QP
- If x_i is misclassified, then

J Therefore

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq$$

- C is a cost-complexity tradeoff parameter. It should be set using error estimation.

(K) • C also controls the influence of ________ .



Other Linear Classifiers

Several other criteria have been proposed for learning linear classifiers. These include

- the perceptron
- single-layer neural net
- Fisher's linear discriminant
- least squares
- linear programming approaches
- perceptron with margin

See Duda, Hart, & Stork for more.

$\boxed{\text{Key}}$ A. density / function B. separating hyperplane

C. $w^T(x - x')$, orthogonal, parallel

D. normal, sign

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

$$= \underbrace{w^T z_0 + b}_{0} + r\frac{w^T w}{\|w\|}$$

$$= r\,\|w\|$$

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

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

G. canonical form

H. quadratic program, support vectors

I. slack variables

J. x_i misclassified $\Longrightarrow \xi_i > 1$

$$\frac{1}{n}\sum_{i=1}^{n} \xi_i \geq \text{training error}$$

K. outliers