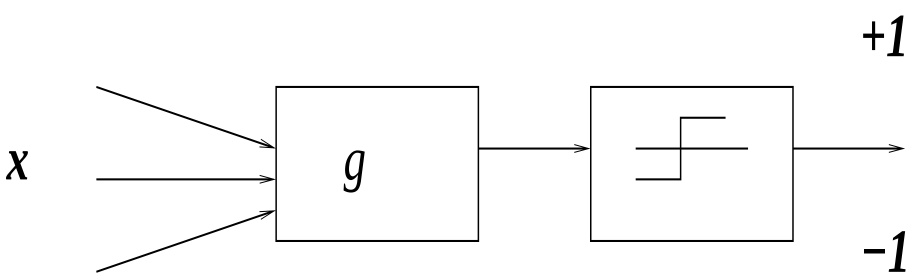
Lecture 5

Linear discriminant analysis

Linear discriminant function

There are many different ways to represent a two class pattern classifier. One way is in terms of a discriminant function $g(\mathbf{x})$.



For a new sample $\mathbf{x}$ and a given discriminant function, we can decide on $\mathbf{x}$ belongs to Class 1 if $g(\mathbf{x}) > 0$, otherwise it's Class 2.

A discriminant function that is a linear combination of the components of $\mathbf{x}$ can be written as

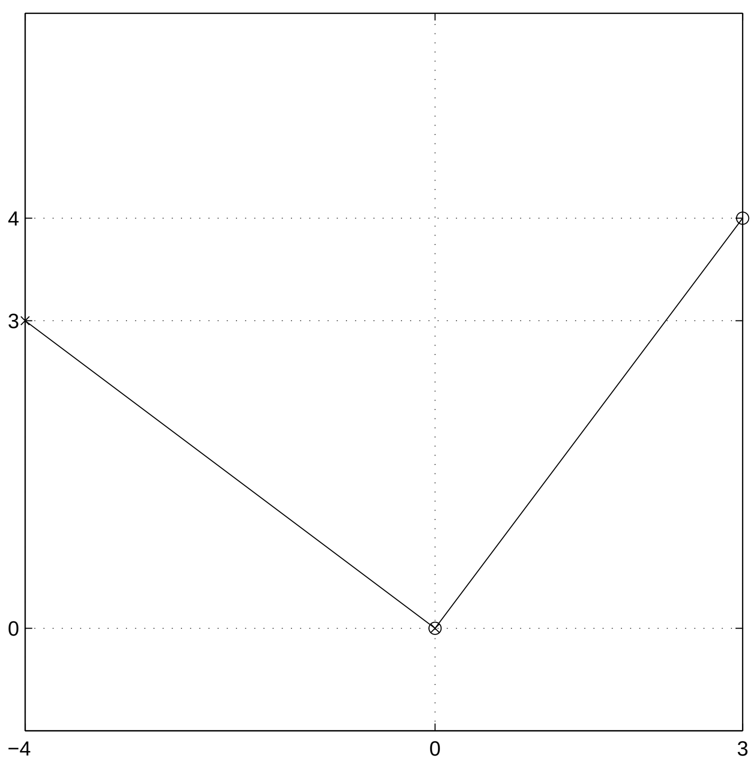
$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + w_0$$

where $\mathbf{w}$ is called the weight vector and w_0 the threshold weight.

The equation $g(\mathbf{x}) = 0$ defines the decision surface that separates data samples assigned to Class 1 from data samples assigned to Class 2. This is a hyperplane when $g(\mathbf{x})$ is linear.

Two vectors $\mathbf{a}$ and $\mathbf{b}$ are normal to each other if $\mathbf{a}^T\mathbf{b} = 0$. In Figure below we see $[3, 4]$ and $[-4, 3]$ are normal to each other, in algebraic terms,

$$[3, 4][-4, 3]^T = 3 \times (-4) + 4 \times 3 = 0$$



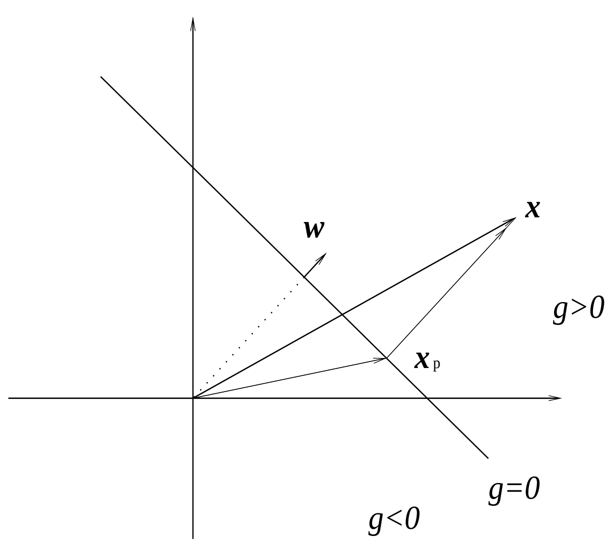
If two points $\mathbf{x}_1$, $\mathbf{x}_2$ are both on the decision surface, then

$$g(\mathbf{x}_1) = g(\mathbf{x}_2) = 0$$

$$\mathbf{w}^T\mathbf{x}_1 + w_0 = \mathbf{w}^T\mathbf{x}_2 + w_0 = 0$$

$$\mathbf{w}^T(\mathbf{x}_1 - \mathbf{x}_2) = 0$$

This means that $\mathbf{w}$ is normal to any vector lying in the hyperplane ($(\mathbf{x}_1 - \mathbf{x}_2)$ is a vector lying on the the decision surface as it starts from $\mathbf{x}_2$, ends at $\mathbf{x}_1$).



Write

$$\mathbf{x} = \mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}$$

where $\mathbf{x}_p$ is the projection of $\mathbf{x}$ on the hyperplane. r is the distance from $\mathbf{x}$ to the hyperplane.

$$\begin{aligned} g(\mathbf{x}) &= \mathbf{w}^T\mathbf{x} + w_0 \\ &= \mathbf{w}^T[\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}] + w_0 \\ &= \mathbf{w}^T\mathbf{x}_p + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} + w_0 \\ &= \underbrace{\mathbf{w}^T\mathbf{x}_p + w_0}_{0} + r\|\mathbf{w}\| \\ &= r\|\mathbf{w}\| \end{aligned}$$

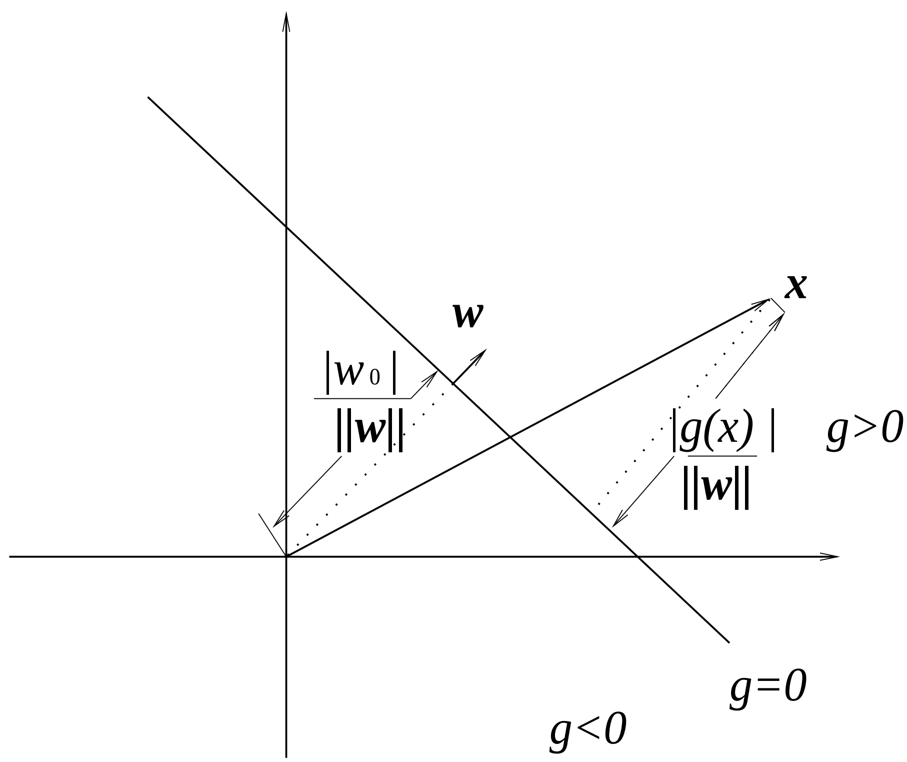
Hence the distance of any data point to the hyperplane is given by

$$r = \frac{g(\mathbf{x})}{\|\mathbf{w}\|}$$

In particular, when $\mathbf{x} = [0, 0]^T$.

$$r = \frac{w_0}{||\mathbf{w}||}$$

A linear discriminant function divides the feature space by a hyperplane, of which the orientation is determined by the normal vector $\mathbf{w}$, and the location is determined by w_0. If $w_0 = 0$, then the hyperplane passes origin. If $w_0 > 0$, the origin is on the positive side of the hyperplane.



Example 1: In order to select the best candidates, an over-subscribed secondary school sets an entrance exam on two subjects of English and Mathematics. The marks of 5 applicants as listed in the Table below and the decision for acceptance is passing an average mark of 75.

(i) Show that the decision rule is equivalent of the method of linear discriminant function.

(ii) Plot the decision hyperplane, indicating the half planes of both Accept and Reject, and location of the 5 applicants.

Candidate No.	English	Math	Decision
1	80	85	Accept
2	70	60	Reject
3	50	70	Reject
4	90	70	Accept
5	85	75	Accept

Solution: (i) Denote marks of English and Math as x_1 and x_2, respectively. The decision rule is if $\frac{x_1+x_2}{2} > 75$, accept, otherwise reject. This is equivalent to using a linear discriminant function

$$g(\mathbf{x}) = x_1 + x_2 - 150$$

with decision rule: if $g(\mathbf{x}) > 0$, accept, otherwise reject.

(ii) To plot $g(\mathbf{x}) = 0$, the easiest way is to set $x_1 = 0$, find the value of x_2 so that $g(\mathbf{x}) = 0$. i.e. $0 = 0 + x_2 - 150$, so $x_2 = 150$.

$[0, 150]^T$ is on the hyperplane.

Likewise we can also set $x_2 = 0$, find the value of x_1 so that $g(\mathbf{x}) = 0$. i.e. $0 = x_1 + 0 - 150$, so $x_1 = 150$.

$[150, 0]^T$ is on the hyperplane.

Plot a straight line linking $[0, 150]^T$ and $[150, 0]^T$.

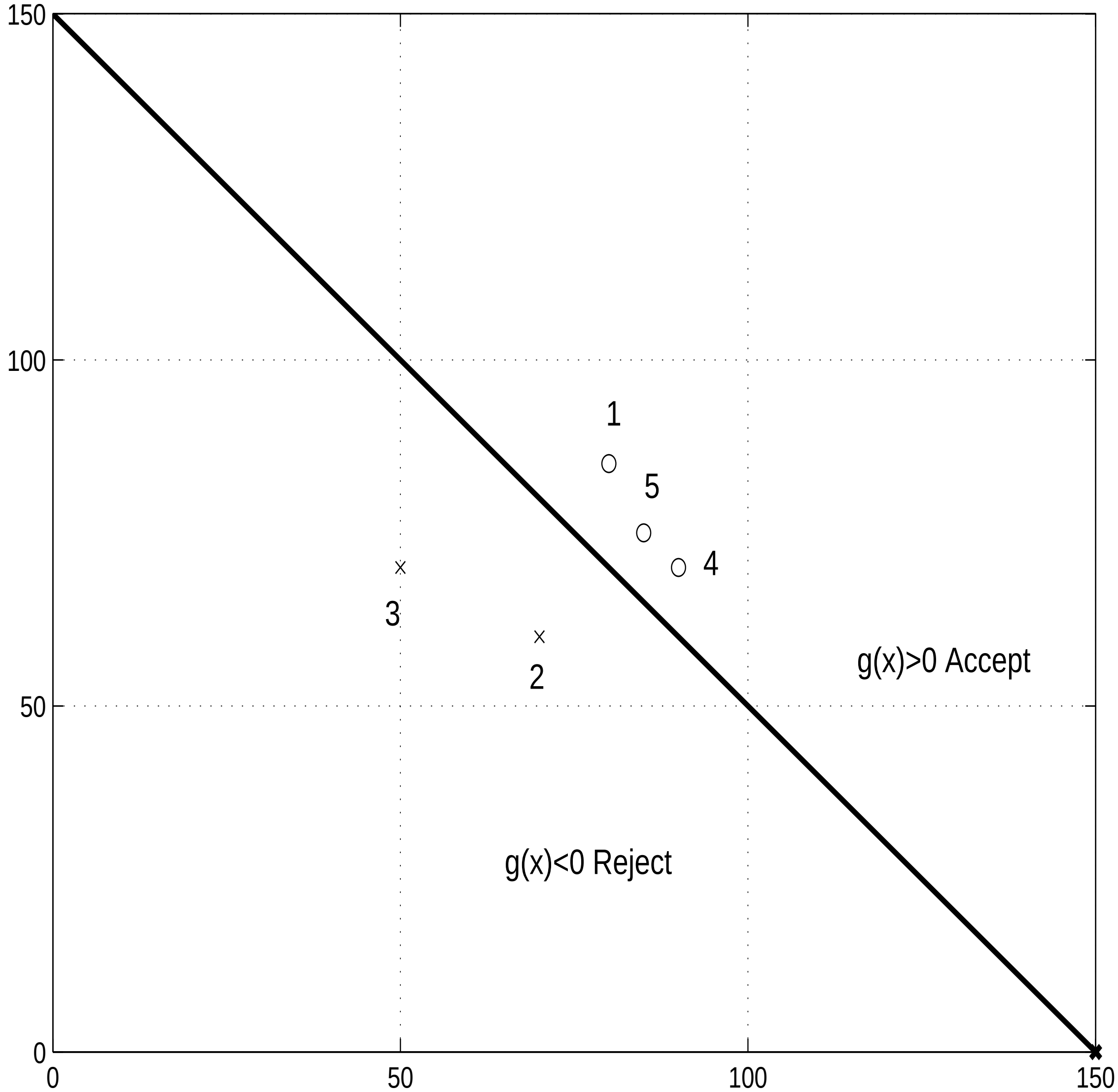


Figure: The solution to the example 1 (ii).

There are many ways of determining the linear discriminant function $g(\mathbf{x})$ given a set of training data samples. One way is to set the labelled data samples some target values. e.g. +1 for one class and -1 for another class, then the weights of the linear discriminant function are adjusted.

Using the same example, a set of linear equations can be constructed based on values in the previous Table.

$$\begin{cases} 80w_1 + 85w_2 + w_0 = 1 \\ 70w_1 + 60w_2 + w_0 = -1 \\ 50w_1 + 75w_2 + w_0 = -1 \\ 90w_1 + 70w_2 + w_0 = 1 \\ 85w_1 + 75w_2 + w_0 = 1 \end{cases}$$

There are 5 equations to solve 3 unknown parameters. There is no exact solution. Instead, the weights are determined by minimizing the overall errors between both sides.

The solution to this problem is often based on the least squares estimate, given by

$$\begin{bmatrix} w_1 \\ w_2 \\ w_0 \end{bmatrix} =$$

$$\{\begin{pmatrix} 80 & 70 & 50 & 90 & 85 \\ 85 & 60 & 75 & 70 & 75 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{pmatrix} 80 & 85 & 1 \\ 70 & 60 & 1 \\ 50 & 75 & 1 \\ 90 & 70 & 1 \\ 85 & 75 & 1 \end{pmatrix}\}^{-1}$$

$$\times \begin{pmatrix} 80 & 70 & 50 & 90 & 85 \\ 85 & 60 & 75 & 70 & 75 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix} \begin{bmatrix} 1 \\ -1 \\ -1 \\ 1 \\ 1 \end{bmatrix}$$

$$= [0.0571, 0.0580, -8.3176]^T$$

So $g(\mathbf{x}) = 0.0571x_1 + 0.0580x_2 - 8.3176$. Note that this is the same hyperplane defined by $g(\mathbf{x}) = x_1 + 1.0106x_2 - 145.6684$, close to the hyperplane used in Example 1.