Chapter 5: Linear Discriminant Functions (Sections 5.1-5.3)

- Introduction
- Linear Discriminant Functions and Decision Surfaces
- Generalized Linear Discriminant Functions

Introduction

- In chapter 3, the underlying probability densities were known (or given)
- The training sample was used to estimate the parameters of these probability densities (ML, MAP estimations)
- In this chapter, we only know the proper forms for the discriminant functions: similar to non-parametric techniques
- They may not be optimal, but they are very simple to use
- They provide us with linear classifiers

Linear discriminant functions and decision surfaces

- Definition

It is a function that is a linear combination of the components of x

$$g(x) = w^t x + w_0 \qquad (1)$$

where w is the weight vector and w_0 is the bias

- A two-category classifier with a discriminant function of the form (1) uses the following rule:

Decide w_1 *if* $g(x) > 0$ and w_2 if $g(x) < 0$

or Decide w_1 *if* $w^t x > -w_0$ and w_2 otherwise

If $g(x) = 0$ Þ x is assigned to either class

Linear discriminant functions (LDF): discriminant functions are linear with respect to the input feature vector (usually denoted by X).

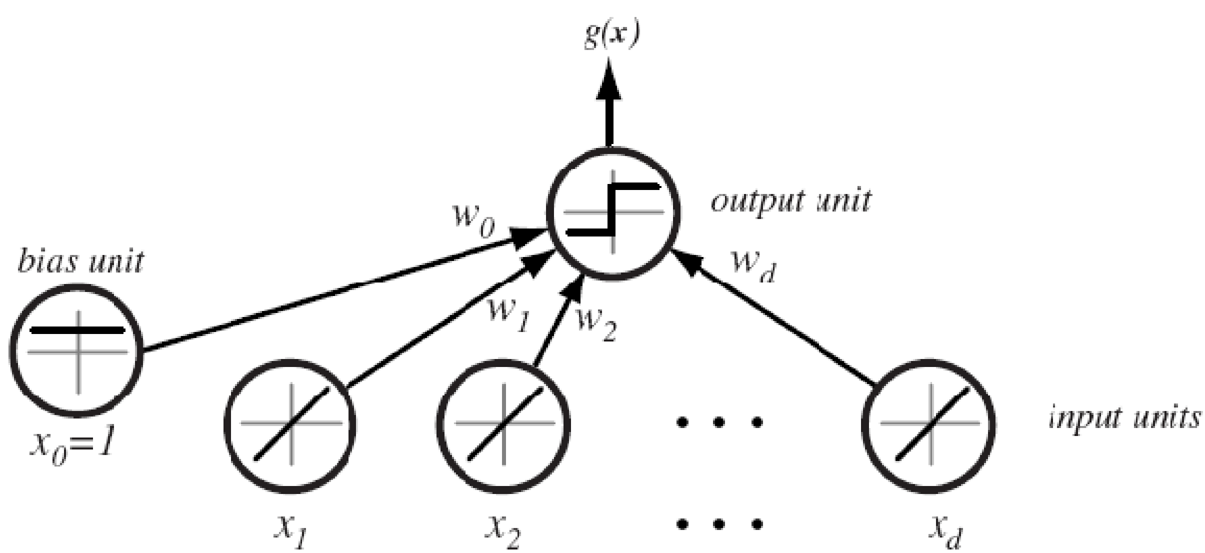
$$g(x) = w^T X + w_0 = w_0 + \Sigma w_k x_k$$

where

$X = [x_1, x_2, \ldots, x_d]$ is the pattern **feature vector**,
$w = [w_1, w_2, \ldots, w_d]$ is the **weight vector**,
w_0 is the **threshold weight** (also called **bias**), and
d is the number of features extracted from a pattern.

A simplified model is obtained by defining $x_0 = 1$:

$$g(x) = w_0 x_0 + w_1 x_1 + w_2 x_2 + \ldots + w_d x_d = \Sigma w_k x_k \text{ for } k = 0, 1, \ldots, d$$



- The equation $g(x) = 0$ defines the decision surface that separates points assigned to the category w_1 from points assigned to the category w_2

- When $g(x)$ is linear, the decision surface is a hyperplane

- Algebraic measure of the distance from x to the hyperplane (interesting result!)

TWO-CATEGORY CASE

- Let's consider a two-category classifier implementing the following decision rule: decide class ω_1 if $g(x) > 0$, and class ω_2 if $g(x) < 0$. Then, the **decision boundary surface** is the hyper-plane H defined by $g(x) = 0$ and dividing the feature space into two half-spaces: decision region R_1 for class ω_1 and decision region R_2 for class ω_2.

  For any two points X_1 and X_2 on the decision surface:

  $$w^T X_1 + w_0 = w^T X_2 + w_0 \quad \Rightarrow \quad w^T(X_1 - X_2) = 0$$

  In conclusion, w is normal to any vector lying on H.

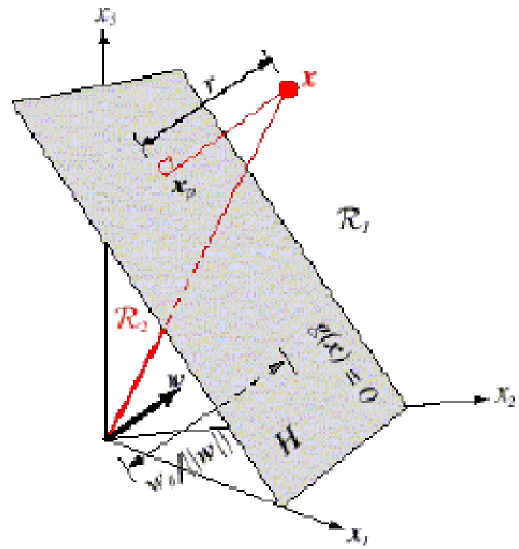
  Subsequently, if x_p is the projection of x on H, and r is the distance from x to H, then $x = x_p + r\,(w / ||w||)$.

  Since $g(x_p) = 0$ and $g(x) = w^T x + w_0$, then:

  $g(x) = w^T x_p + w_0 + r\,(w^T w / ||w||) = r(||w||^2 / ||w||) = r\,||w||$

  Finally:

  $$r = g(x) / ||w||$$



- **Conclusion:** the orientation of the decision surface is determined by the normal vector w, while its location is determined by the bias w_0.

Multicategory case

Using one linear decision boundary for each class (c boundaries)

Using one linear decision boundary for every pair of classes (c(c-1)/2 boundaries)

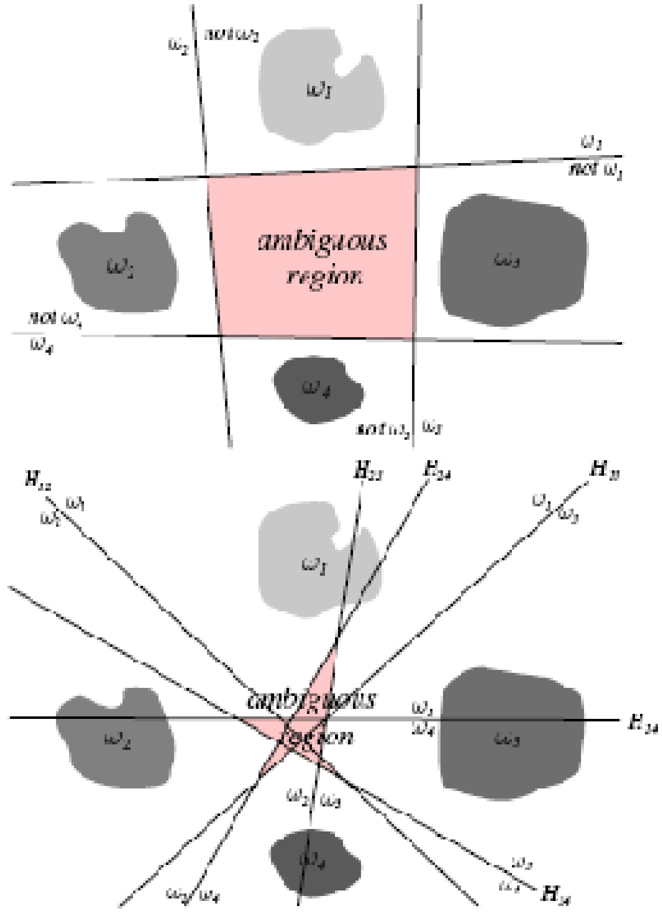


FIGURE 5.3. Linear decision boundaries for a four-class problem. The top figure shows ω_i/not ω_i dichotomies while the bottom figure shows ω_i/ω_j dichotomies and the corresponding decision boundaries H_{ij}. The pink regions have ambiguous category assignments. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

q The multi-category case

- q We define c linear discriminant functions

$$g_i(x) = w_i^t x + w_{i0} \qquad i = 1,\ldots,c$$

and assign x to w_i if $g_i(x) > g_j(x)$ " j ¹ i; in case of ties, the classification is undefined

- q In this case, the classifier is a “linear machine”
- q A linear machine divides the feature space into c decision regions, with $g_i(x)$ being the largest discriminant if x is in the region R_i
- q For a two contiguous regions R_i and R_j; the boundary that separates them is a portion of hyperplane H_{ij} defined by:

$$g_i(x) = g_j(x)$$

$$(w_i - w_j)^t x + (w_{i0} - w_{j0}) = 0$$

- q $w_i - w_j$ is normal to H_{ij} and

$$d(x, H_{ij}) = r = \frac{g_i - g_j}{\|w_i - w_j\|}$$

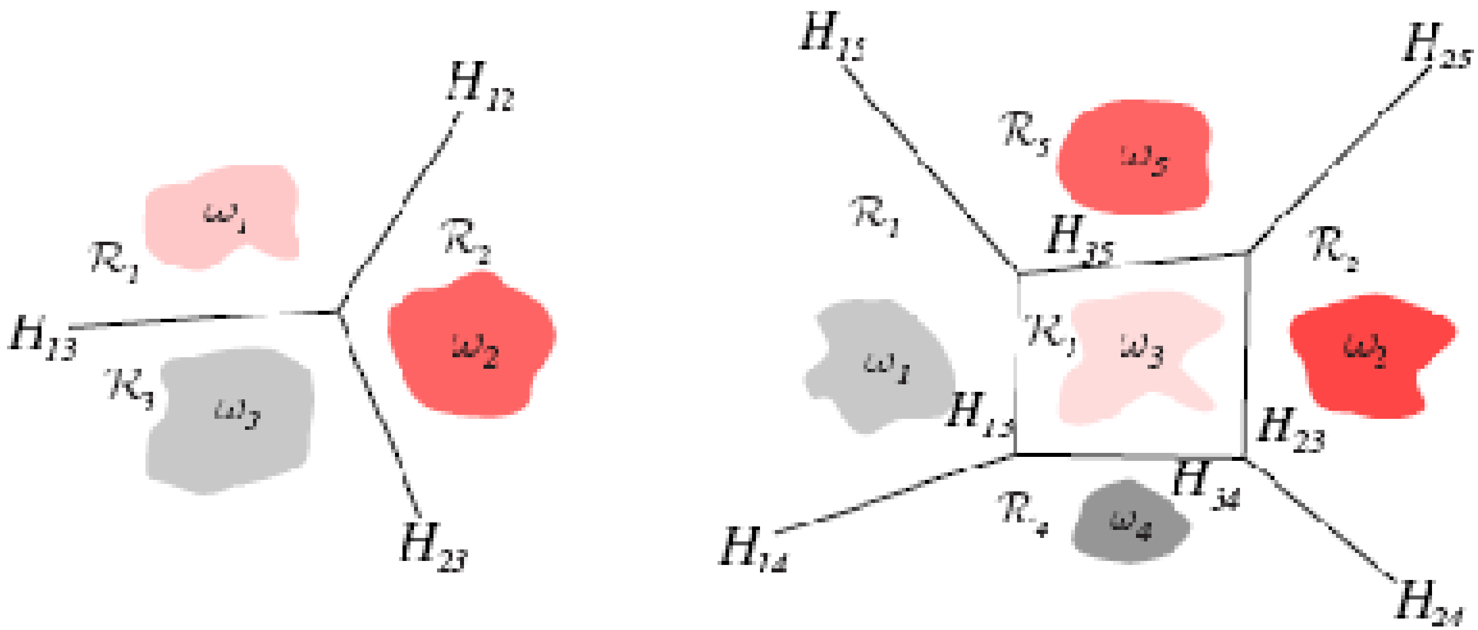
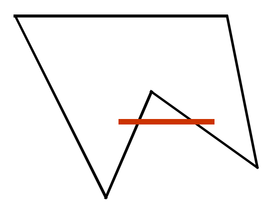


FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- It is easy to show that the decision regions for a linear machine are convex, this restriction limits the flexibility and accuracy of the classifier



GENERAL CASES

- **Linear case:** linear discriminant function.

$$g(x) = w_0 + \sum_{k=1}^{d} w_k x_k$$

  - Decision boundary surface: hyper-plane.

- **Quadratic case:** quadratic discriminant function.

$$g(x) = w_0 + \sum_{k=1}^{d} w_k x_k + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j$$

  - Decision boundary: hyper-quadratic surface (such as hyper-spheres, hyper-ellipsoids, hyper-hyperboloids).

- **Polynomial case:** polynomial discriminant function.

$$g(x) = w_0 + \sum_{k=1}^{d} w_k x_k + \sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} x_i x_j + \sum_{i=1}^{d} \sum_{j=1}^{d} \sum_{k=1}^{d} w_{ijk} x_i x_j x_k$$

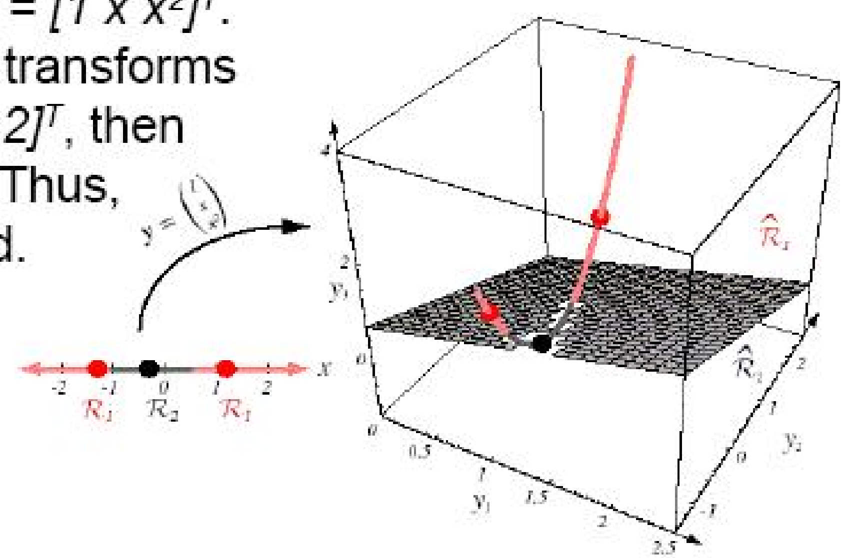
GENERALIZED LDF

- **General case:**
  - Polynomial discriminant function can be reduced to a linear form by introducing some arbitrary functions $y = f(x)$ (and making the discriminant linear on y instead of x).

$$y_k = \begin{bmatrix} 1 \\ x_k \\ x_k^2 \\ x_k^3 \end{bmatrix}$$

$$g(x) = w_0 + \sum_{k=1}^{d} a_k y_k(x)$$

- **Example:**
  - If $g(x) = a_0 + a_1 x + a_2 x^2$, then $y = [1\ x\ x^2]^T$.
  - This mapping takes a line and transforms it to a 3D parabola. If $a = [-1\ 1\ 2]^T$, then $g(x) > 0$ for $x < -1$ and $x > 0.5$. Thus, region R_1 is multiple connected.



$$g(x) = a_1 x_1 + a_2 x_2 + a_3 x_1 x_2$$

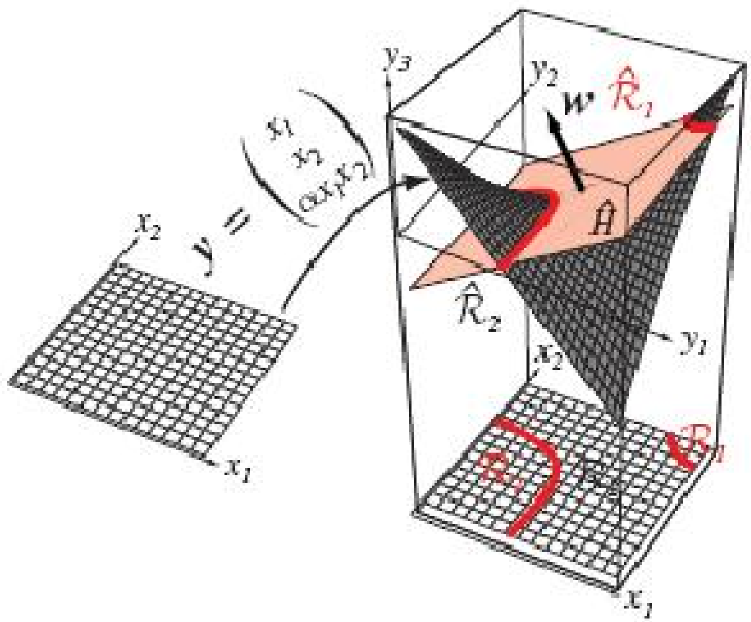


FIGURE 5.6. The two-dimensional input space $\mathbf{x}$ is mapped through a polynomial function f to $\mathbf{y}$. Here the mapping is $y_1 = x_1$, $y_2 = x_2$ and $y_3 \propto x_1 x_2$. A linear discriminant in this transformed space is a hyperplane, which cuts the surface. Points to the positive side of the hyperplane $\hat{H}$ correspond to category ω_1, and those beneath it correspond to category ω_2. Here, in terms of the $\mathbf{x}$ space, $\mathcal{R}_1$ is a not simply connected. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Bias Elimination

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

$g(\mathbf{y}) = \mathbf{a}^t\mathbf{y}$

- $\mathbf{a}^t = [w_0 , \mathbf{w}^t]$
- $\mathbf{y}^t = [1 , \mathbf{x}]$

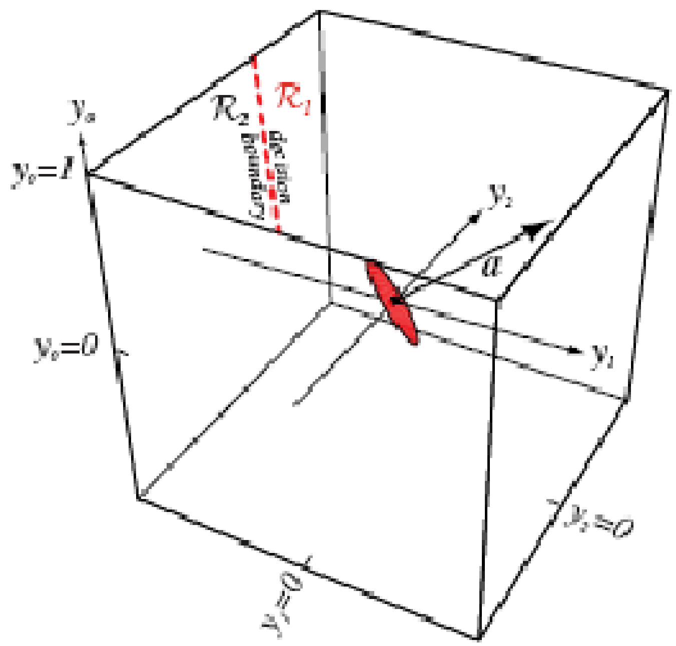


FIGURE 5.7. A three-dimensional augmented feature space $\mathbf{y}$ and augmented weight vector $\mathbf{a}$ (at the origin). The set of points for which $\mathbf{a}^t\mathbf{y} = 0$ is a plane (or more generally, a hyperplane) perpendicular to $\mathbf{a}$ and passing through the origin of $\mathbf{y}$-space, as indicated by the red disk. Such a plane need not pass through the origin of the two-dimensional feature space of the problem, as illustrated by the dashed decision boundary shown at the top of the box. Thus there exists an augmented weight vector $\mathbf{a}$ that will lead to any straight decision line in $\mathbf{x}$-space. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Generalized Linear Discriminant Functions

q Decision boundaries which separate between classes may not always be linear

q The complexity of the boundaries may sometimes request the use of highly non-linear surfaces

q A popular approach to generalize the concept of linear decision functions is to consider a generalized decision function as:

$$g(x) = w_1 f_1(x) + w_2 f_2(x) + \dots + w_N f_N(x) + w_{N+1} \qquad (1)$$

where $f_i(x)$, 1 £ i £ N are scalar functions of the pattern x

- Introducing $f_{N+1}(x) = 1$ we get:

$$g(x) = \sum_{i=1}^{N+1} w_i f_i(x) = w^T . \hat{x}$$

$$\text{where } w = (w_1, w_2, \ldots, w_N, w_{N+1})^T \text{ and } \hat{x} = (f_1(x), f_2(x), \ldots, f_N(x), f_{N+1}(x))^T$$

- This latter representation of *g(x)* implies that any decision function defined by equation (1) can be treated as linear in the *(N + 1)* dimensional space
- *g(x)* maintains its non-linearity characteristics related to x

- The most commonly used generalized decision function is *g(x)* for which *f_i(x) (1 £ i £N)* are polynomials

$$g(x) = (\vec{w})^T \vec{x}$$ T: is the vector transpose form

Where $\vec{w}$ is a new weight vector, which can be calculated from the original w and the original linear *f_i(x), 1 £ i £N*

- Quadratic decision functions for a 2-dimensional feature space

$$g(x) = w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_2^2 + w_4 x_1 + w_5 x_2 + w_6$$

$$here: \vec{w} = (w_1, w_2, ..., w_6)^T \text{ and } \vec{x} = (x_1^2, x_1 x_2, x_2^2, x_1, x_2, 1)^T$$

- For patterns $x \in R^n$, the most general quadratic decision function is given by:

$$g(x) = \sum_{i=1}^{n} w_{ii} x_i^2 + \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} w_{ij} x_i x_j + \sum_{i=1}^{n} w_i x_i + w_{n+1} \qquad (2)$$

The number of terms at the right-hand side is:

$$l = N + 1 = n + \frac{n(n-1)}{2} + n + 1 = \frac{(n+1)(n+2)}{2}$$

This is the total number of weights which are the free parameters of the problem

- If for example n = 3, the vector $\vec{x}$ is 10-dimensional
- If for example n = 10, the vector $\vec{x}$ is 66-dimensional

- In the case of polynomial decision functions of order m, a typical $f_i(x)$ is given by:

$$f_i(x) = x_{i_1}^{e_1} x_{i_2}^{e_2} \ldots x_{i_m}^{e_m}$$

$$\textit{where } 1 \leq i_1, i_2, \ldots, i_m \leq n \textit{ and } e_i, 1 \leq i \leq m \textit{ is } 0 \textit{ or } 1.$$

  - It is a polynomial with a degree between 0 and m. To avoid repetitions, we request $i_1 \leq i_2 \leq \ldots \leq i_m$

$$g^m(x) = \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \cdots \sum_{i_m=i_{m-1}}^{n} w_{i_1 i_2 \ldots i_m} x_{i_1} x_{i_2} \ldots x_{i_m} + g^{m-1}(x)$$

(where $g^0(x) = w_{n+1}$) is the most general polynomial decision function of order m

Example 1: Let $n = 3$ and $m = 2$ then:

$$g^2(x) = \sum_{i_1=1}^{3} \sum_{i_2=i_1}^{3} w_{i_1 i_2} x_{i_1} x_{i_2} + w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4$$

$$= w_{11} x_1^2 + w_{12} x_1 x_2 + w_{13} x_1 x_3 + w_{22} x_2^2 + w_{23} x_2 x_3 + w_{33} x_3^2$$

$$+ w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4$$

Example 2: Let $n = 2$ and $m = 3$ then:

$$g^3(x) = \sum_{i_1=1}^{2} \sum_{i_2=i_1}^{2} \sum_{i_3=i_2}^{2} w_{i_1 i_2 i_3} x_{i_1} x_{i_2} x_{i_3} + g^2(x)$$

$$= w_{111} x_1^3 + w_{112} x_1^2 x_2 + w_{122} x_1 x_2^2 + w_{222} x_2^3 + g^2(x)$$

$$\text{where } g^2(x) = \sum_{i_1=1}^{2} \sum_{i_2=i_1}^{2} w_{i_1 i_2} x_{i_1} x_{i_2} + g^1(x)$$

$$= w_{11} x_1^2 + w_{12} x_1 x_2 + w_{22} x_2^2 + w_1 x_1 + w_2 x_2 + w_3$$

- The commonly used quadratic decision function can be represented as the general n- dimensional quadratic surface:

$$g(x) = x^T A x + x^T b + c$$

where the matrix $A = (a_{ij})$, the vector $b = (b_1, b_2, \ldots, b_n)^T$ and c, depends on w_{n+1} of equation (2)

- If A is positive definite then the decision function is a hyperellipsoid with axes in the directions of the eigenvectors of A
  - In particular: if $A = I_n$ (Identity), the decision function is simply the n-dimensional hypersphere

- If A is negative definite, the decision function describes a hyperhyperboloid

- In conclusion: it is only the matrix A which determines the shape and characteristics of the decision function

Problem: Consider a 3 dimensional space and a quadratic decision function

Let R^3 be the original pattern space and let the decision function associated with the pattern classes w_1 and w_2 be:

$$g(x) = 2x_1^2 + x_3^2 + x_2 x_3 + 4x_1 - 2x_2 + 1$$

for which $g(x) > 0$ *if* x Î w_1 and $g(x) < 0$ *if* x Î w_2

a) Rewrite $g(x)$ as $g(x) = x^T A x + x^T b + c$

b) Determine the class of each of the following pattern vectors:

(1,1,1), (1,10,0), (0,1/2,0)

q Characteristics of a Matrix

1. A square matrix A is *positive definite if* $x^T Ax > 0$ for all nonzero column vectors x.

2. It is negative definite if $x^T Ax < 0$ for all nonzero x.

3. It is *positive semi-definite* if $x^T Ax$ ³ 0.

4. And *negative semi-definite* if $x^T Ax$ £ 0 for all x.

These definitions are hard to check directly and you might as well forget them for all practical purposes.

More useful in practice are the following properties, which hold when the matrix A is symmetric and which are easier to check.

The *ith principal minor* of A is the matrix A_i formed by the first i rows and columns of A. So, the first principal minor of A is the matrix $A_1 = (a_{11})$, the second principal minor is the matrix:

$$A_2 = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \textit{ and so on.}$$

q The matrix A (nxn) is positive definite if all its principal minors $A_1, A_2, \ldots, A_n$ have strictly positive determinants

q If these determinants are non-zero and alternate in signs, starting with $det(A_1)<0$, then the matrix A is negative definite

q If the determinants are all non-negative, then the matrix is positive semi-definite

q If the determinant alternate in signs, starting with $det(A_1)$£0, then the matrix is negative semi-definite

To fix ideas, consider a 2x2 symmetric matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}.$$

- It is positive definite if:
  - a) $det(A_1) = a_{11} > 0$
  - b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} > 0$

- It is negative definite if:
  - a) $det(A_1) = a_{11} < 0$
  - b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} > 0$

- It is positive semi-definite if:
  - a) $det(A_1) = a_{11} \geq 0$
  - b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} \geq 0$

- And it is negative semi-definite if:
  - a) $det(A_1) = a_{11} \leq 0$
  - b) $det(A_2) = a_{11}a_{22} - a_{12}a_{12} \geq 0.$

Exercise 1: Check whether the following matrices are positive definite, negative definite, positive semi-definite, negative semi-definite or none of the above.

$$(a)\ A = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$$

$$(b)\ A = \begin{pmatrix} -2 & 4 \\ 4 & -8 \end{pmatrix}$$

$$(c)\ A = \begin{pmatrix} -2 & 2 \\ 2 & -4 \end{pmatrix}$$

$$(d)\ A = \begin{pmatrix} 2 & 4 \\ 4 & 3 \end{pmatrix}$$

Solutions of Exercise 1:

- $A_1 = 2 > 0$
  $A_2 = 8 - 1 = 7 > 0$ $\Rightarrow$ A is positive definite

- $A_1 = -2$
  $A_2 = (-2 \times -8) - 16 = 0$ $\Rightarrow$ A is negative semi-definite

- $A_1 = -2$
  $A_2 = 8 - 4 = 4 > 0$ $\Rightarrow$ A is negative definite

- $A_1 = 2 > 0$
  $A_2 = 6 - 16 = -10 < 0$ $\Rightarrow$ A is none of the above

Exercise 2:

Let $A = \begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}$

1. Compute the decision boundary assigned to the matrix A $(g(x) = x^T A x + x^T b + c)$ in the case where $b^T = (1 , 2)$ and $c = -3$

2. Solve $det(A - \lambda I) = 0$ and find the shape and the characteristics of the decision boundary separating two classes w_1 and w_2

3. Classify the following points:
   - $x^T = (0 , -1)$
   - $x^T = (1 , 1)$

Solution of Exercise 2:

1.

$$g(x) = (x_1, x_2)\begin{pmatrix} 2 & 1 \\ 1 & 4 \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + (x_1, x_2)\begin{pmatrix} 1 \\ 2 \end{pmatrix} - 3$$

$$= (2x_1 + x_2, x_1 + 4x_2)\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} + x_1 + 2x_2 - 3$$

$$= 2x_1^2 + x_1x_2 + x_1x_2 + 4x_2^2 + x_1 + 2x_2 - 3$$

$$= 2x_1^2 + 4x_2^2 + 2x_1x_2 + x_1 + 2x_2 - 3$$

2.

$$\text{For } \lambda_1 = 3 + \sqrt{2} \text{ using } \begin{pmatrix} 2-\lambda & 1 \\ 1 & 4-\lambda \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0, \text{ we obtain :}$$

$$\begin{cases} (-1-\sqrt{2})x_1 + x_2 = 0 \\ x_1 + (1-\sqrt{2})x_2 = 0 \end{cases} \Leftrightarrow (-1-\sqrt{2})x_1 + x_2 = 0$$

This latter equation is a straight line colinear to the vector:

$$\vec{V}_1 = (1, 1+\sqrt{2})^T$$

$$For\ l_2 = 3 - \sqrt{2}\ \text{using} \begin{pmatrix} 2 - l & 1 \\ 1 & 4 - l \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0,\ \text{we obtain}:$$

$$\begin{cases} (\sqrt{2} - 1)x_1 + x_2 = 0 \\ x_1 + (1 + \sqrt{2})x_2 = 0 \end{cases} \Leftrightarrow (\sqrt{2} - 1)x_1 + x_2 = 0$$

This latter equation is a straight line colinear to the vector:

$$\vec{V}_2 = (1, 1 - \sqrt{2})^T$$

The ellipsis decision boundary has two axes, which are respectively colinear to the vectors V_1 and V_2

3. X = (0 , -1) ^{T} Þ **g(0 , -1) = -1 < 0** Þ **x** Î w_2

X = (1 , 1) ^{T} Þ **g(1 , 1) = 8 > 0** Þ **x** Î w_1