Chapter 5: Linear Discriminant Functions

- Introduction
- Linear Discriminant Functions and Decisions 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 decisions 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 the bias
  - **A two-category classifier** with a discriminant function of the form (1) uses the following rule:
    Decide ω_1 if $g(x) > 0$ and ω_2 if $g(x) < 0$
    $\Leftrightarrow$ Decide ω_1 if $w^t x > -w_0$ and ω_2 otherwise
    If $g(x) = 0 \Rightarrow$ x is assigned to either class

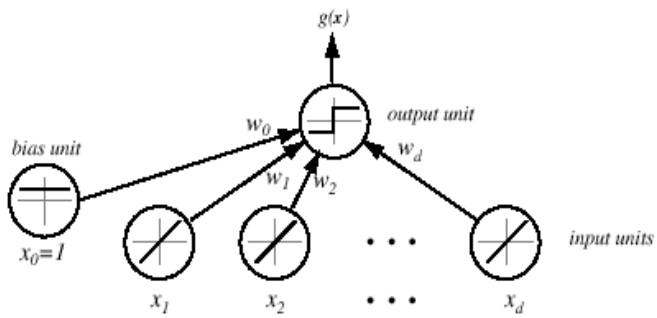

FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show in each unit its effective input-output function. Thus each of the d input units is linear, emitting exactly the value of its corresponding feature value. The single bias unit unit always emits the constant value 1.0. The single output unit emits a $+1$ if $\mathbf{w}^t\mathbf{x} + w_0 > 0$ or a -1 otherwise. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- The equation g(x) = 0 defines the decision surface that separates points assigned to the category ω_1 from points assigned to the category ω_2
- When g(x) is linear, the decision surface is a hyperplane
- Algebraic measure of the distance from x to the hyperplane (interesting result!)

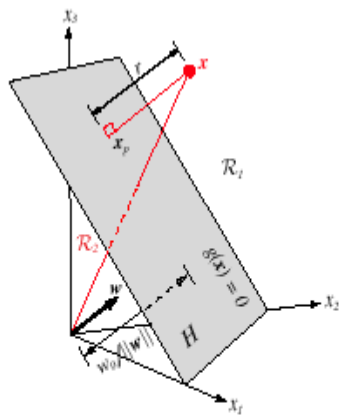

FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}+w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

$$x = x_p + \frac{r.w}{\|w\|} \quad \text{(since w is colinear with } x - x_p \text{ and } \frac{w}{\|w\|} = 1)$$

$$\text{since } g(x) = 0 \text{ and } w^t.w = \|w\|^2$$

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

$$\text{in particular } d(0,H) = \frac{w_0}{\|w\|}$$

- In conclusion, a linear discriminant function divides the feature space by a hyperplane decision surface
- The orientation of the surface is determined by the normal vector w and the location of the surface is determined by the bias

– The multi-category case

- We define c linear discriminant functions
$$g_i(x) = w_i^t x + w_{i0} \qquad i = 1,\ldots,c$$
and assign x to ω_i if $g_i(x) > g_j(x)\ \forall\ j \neq i$; in case of ties, the classification is undefined
- In this case, the classifier is a "linear machine"
- 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
- For a two contiguous regions R_i and R_j; the boundary that separates them is a portion of hyperplane H_{ij} defined by:
$$g_i(x) = g_j(x)$$
$$\Leftrightarrow (w_i - w_j)^t x + (w_{i0} - w_{j0}) = 0$$
- $w_i - w_j$ is normal to H_{ij} and
$$d(x,H_{ij}) = \frac{g_i - g_j}{\|w_i - w_j\|}$$

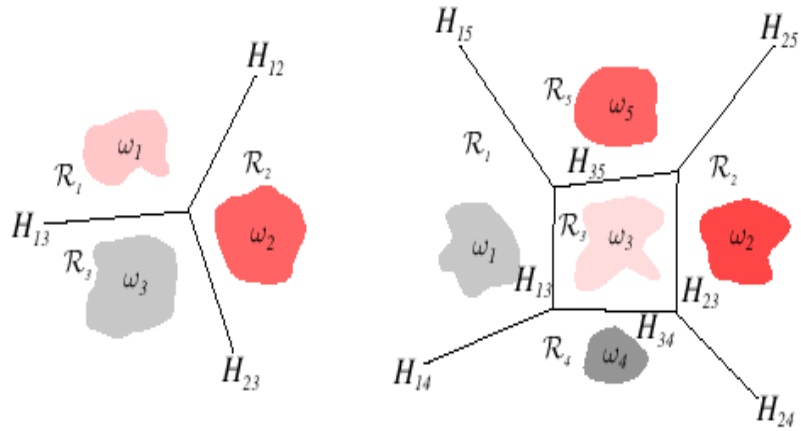

FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- 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

Class Exercises

- Ex. 13 p.159
- Ex. 3 p.201
- Write a C/C++/Java program that uses a k-nearest neighbor method to classify input patterns. Use the table on p.209 as your training sample.

  Experiment the program with the following data:
  - k = 3 $x_1 = (0.33, 0.58, -4.8)$
    $x_2 = (0.27, 1.0, -2.68)$
    $x_3 = (-0.44, 2.8, 6.20)$
  - Do the same thing with k = 11
  - Compare the classification results between k = 3 and k = 11

  (use the most dominant class voting scheme amongst the k classes)

Generalized Linear Discriminant Functions

- Decision boundaries which separate between classes may not always be linear
- The complexity of the boundaries may sometimes request the use of highly non-linear surfaces
- A popular approach to generalize the concept of linear decision functions is to consider a generalized decision function as:

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

where $f_i(x)$, $1 \le i \le N$ are scalar functions of the pattern x, $x \in IR^n$

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

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

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

- 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 (N + 1 > n)

- g(x) maintains its non-linearity characteristics in IR^n

– The most commonly used generalized decision function is g(x) for which $f_i(x)$ $(1 \leq i \leq N)$ are polynomials

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

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

– Quadratic decision functions for a 2-dimensional feature space

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

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

- For patterns $x \in IR^n$, 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 $\dot{x}$ is 10-dimensional
- If for example n = 10, the vector $\dot{x}$ is 65-dimensional

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

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

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

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

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

(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 the weights w_{ii}, w_{ij}, w_i 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 cubic polynomial decision functions

1. How many terms are needed to represent a decision function if only cubic and linear functions are assumed

2. Present the general 4^{th} order polynomial decision function for a 2 dimensional pattern space

3. Let IR^3 be the original pattern space and let the decision function associated with the pattern classes ω_1 and ω_2 be:

$$g(x) = 2x_1^2 + x_3^2 + x_2x_3 + 4x_1 - 2x_2 + 1$$

for which $g(x) > 0$ if $x \in \omega_1$ and $g(x) < 0$ if $x \in \omega_2$

a) Rewrite g(x) as $g(x) = x^TAx + x^Tb + c$

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

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

- Positive Definite Matrices

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 $x^T Ax \geq 0$.
4. And *negative semi-definite* if $x^T Ax \leq 0$ 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 $A_i = (a_{11})$, the second principal minor is the matrix:

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

- The matrix A is positive definite if all its principal minors $A_1, A_2, \ldots, A_n$ have strictly positive determinants
- If these determinants are non-zero and alternate in signs, starting with $\det(A_1)<0$, then the matrix A is negative definite
- If the determinants are all non-negative, then the matrix is positive semi-definite
- If the determinant alternate in signs, starting with $\det(A_1)\leq 0$, then the matrix is negative semi-definite

To fix ideas, consider a 2x2 symmetric matrix:

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

- 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.

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

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

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

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

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-positive
- $A_1 = -2$
  $A_2 = 8 - 4 = 4 > 0$ $\Rightarrow$ A is negative definite
- $A_1 = 2 > 0$
  $A_2 = 6 - 16 = -10 < 0$ $\Rightarrow$ A is none of the above

Exercise 2:

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

1. Compute the decision boundary assigned to the matrix A ($g(x) = x^T Ax + x^T b + c$) 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 ω_1 and ω_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. For $\lambda_1 = 3 + \sqrt{2}$ using $\begin{pmatrix} 2 - \lambda & 1 \\ 1 & 4 - \lambda \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$, 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 $\lambda_2 = 3 + \sqrt{2}$ using $\begin{pmatrix} 2-\lambda & 1 \\ 1 & 4-\lambda \end{pmatrix}\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = 0$, 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. $X = (0 , -1)^T \Rightarrow g(0 , -1) = -1 < 0 \Rightarrow x \in \omega_2$

   $X = (1 , 1)^T \Rightarrow g(1 , 1) = 8 > 0 \Rightarrow x \in \omega_1$