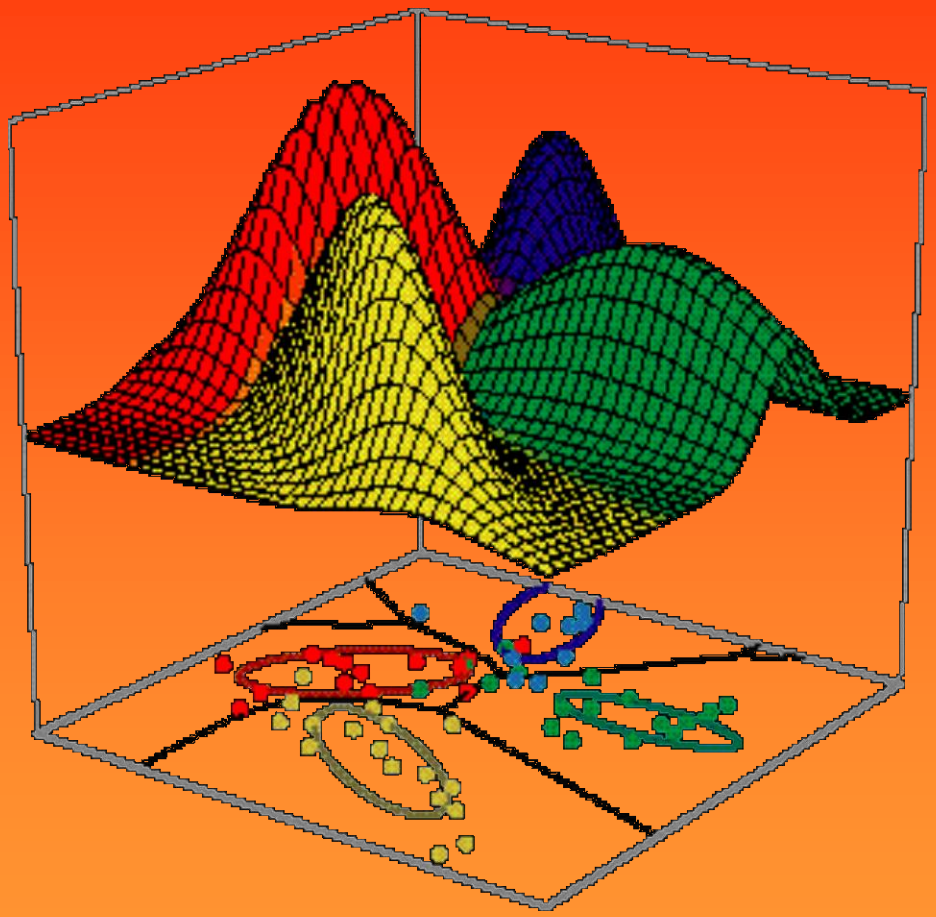

Pattern Classification

All materials in these slides were taken from *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2000*
with the permission of the authors and the publisher

Capitulo 5:
Funciones Discriminantes Lineales (Sections 5.1-5-3)

Introduction

Linear Discriminant Functions and Decisions Surfaces

Generalized Linear Discriminant Functions

Introduccion

- En el capitulo 3, Las densidades de probabilidad subyacentes ern conocidas
- La muestra de entrenamiento fue usada para estimar los parametros de esas distribuciones de probabilidad (estimadores ML, MAP)
- En este capitulo, solo sabemos las formas de las funciones discriminantes : similar a las tecnicas no-parametricas
- Pueden no ser optimas pero son muy faciles de usar
- Nos proveen de clasificadores lineales

Funciones discriminantes lineales y superficies de decision

- Definicion

  Es una funcion que es una combinacion lineal de componentes de x

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

  donde w es el vector de pesos y w_0 el de pesos

- Un clasificador con dos clases con una funcion discriminante de la forma (1) usa la siguiente regla:

  Decide por ω_1 *si* $g(x) > 0$ y ω_2 si $g(x) < 0$

  $\Leftrightarrow$ Decide por ω_1 *si* $w^t x > -w_0$ y ω_2 en otro caso

  si $g(x) = 0 \Rightarrow x$ se asigna a cualquiera de las clases

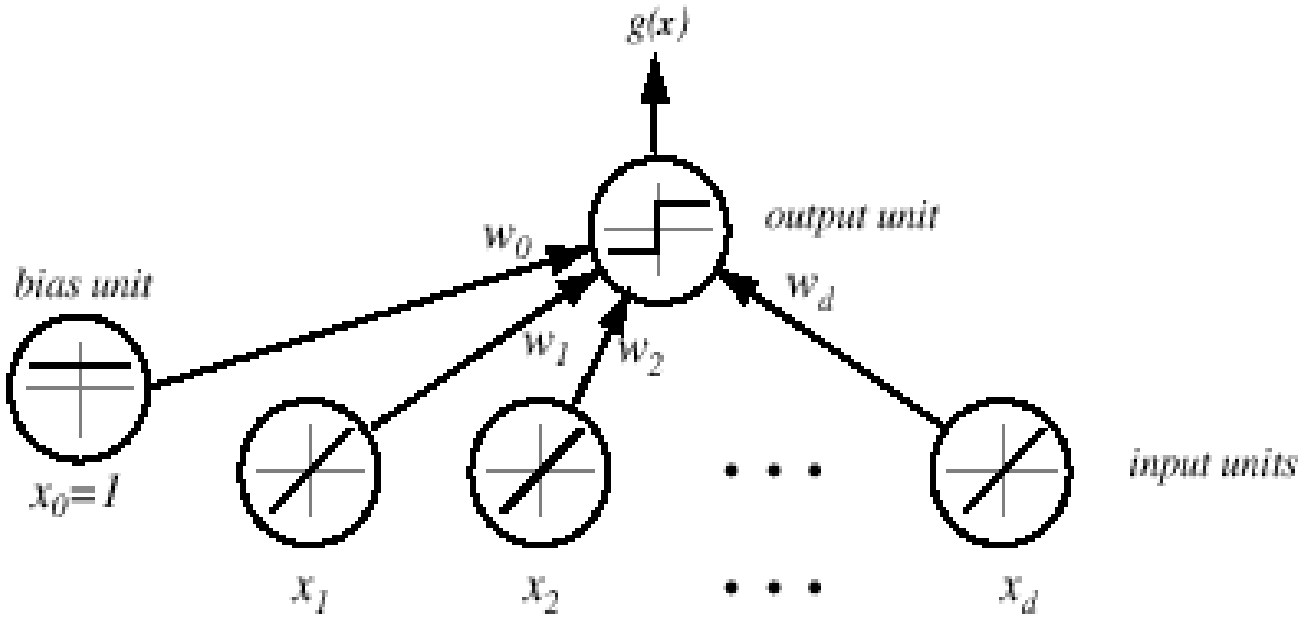

FIGURE 5.1. A simple linear classifier having d input units, each corresponding to the values of the components of an input vector. Each input feature value x_i is multiplied by its corresponding weight w_i; the effective input at the output unit is the sum all these products, $\sum w_i x_i$. We show 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.

- La ecuacion *g(x) = 0* define la superficie de decision que separa puntos asignados a la categoria ω_1 de puntos asignados a la categoria ω_2

- Cuando *g(x)* es lineal, la superficie de decision es un hiperplano
- r es la distancia algebraica de el punto x al hiperplano H, (la minima distancia de un punto fijo a cualquier punto de un plano)

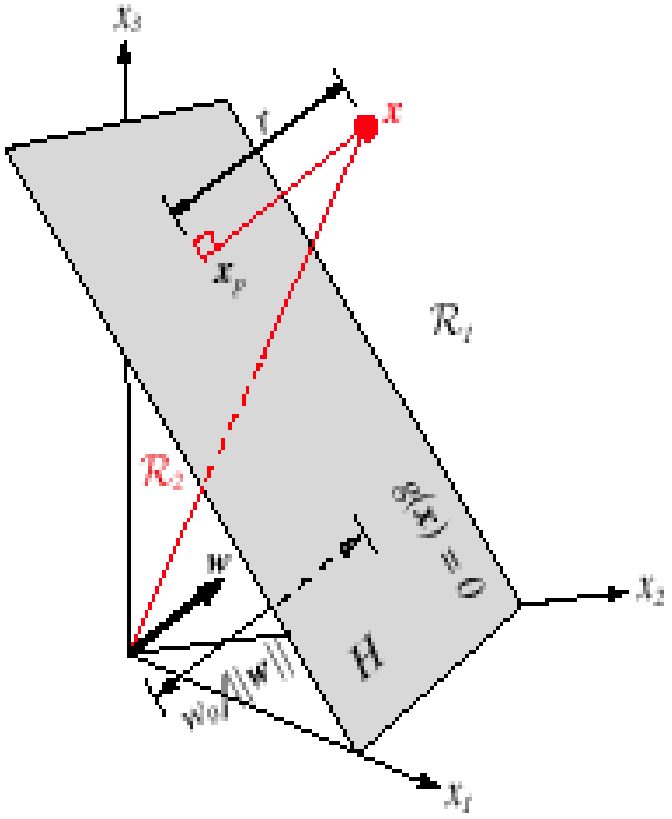

FIGURE 5.2. The linear decision boundary H, where $g(\mathbf{x}) = \mathbf{w}^t\mathbf{x}+w_0 = 0$, separates the feature space into two half-spaces $\mathcal{R}_1$ (where $g(\mathbf{x}) > 0$) and $\mathcal{R}_2$ (where $g(\mathbf{x}) < 0$). From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern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

$$x = x_p + \frac{r.w}{\|w\|}$$

$$g(x) = w^t x + w_0 = w^t x_p + w^t \frac{r.w}{\|w\|} + w_0 = g(x_p) + r\|w\|$$

$$= r\|w\|$$

pues $g(x_p) = 0$ y $w^t.w = \|w\|^2$

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

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

- Como conclusion, una funcion discriminante lineal divide el espacio de caracteristicas por una supercifie de decision (hiperplano)

- La orientacion de la superficie esta determinada por el vector normal a la superficie w y la posicion de la superficie esta determinada por el desvio w_0

Hay mas de una forma de extender clasificadores lineales a mas de dos categorias.
En la figura mostramos el mas simple y le mas complejo

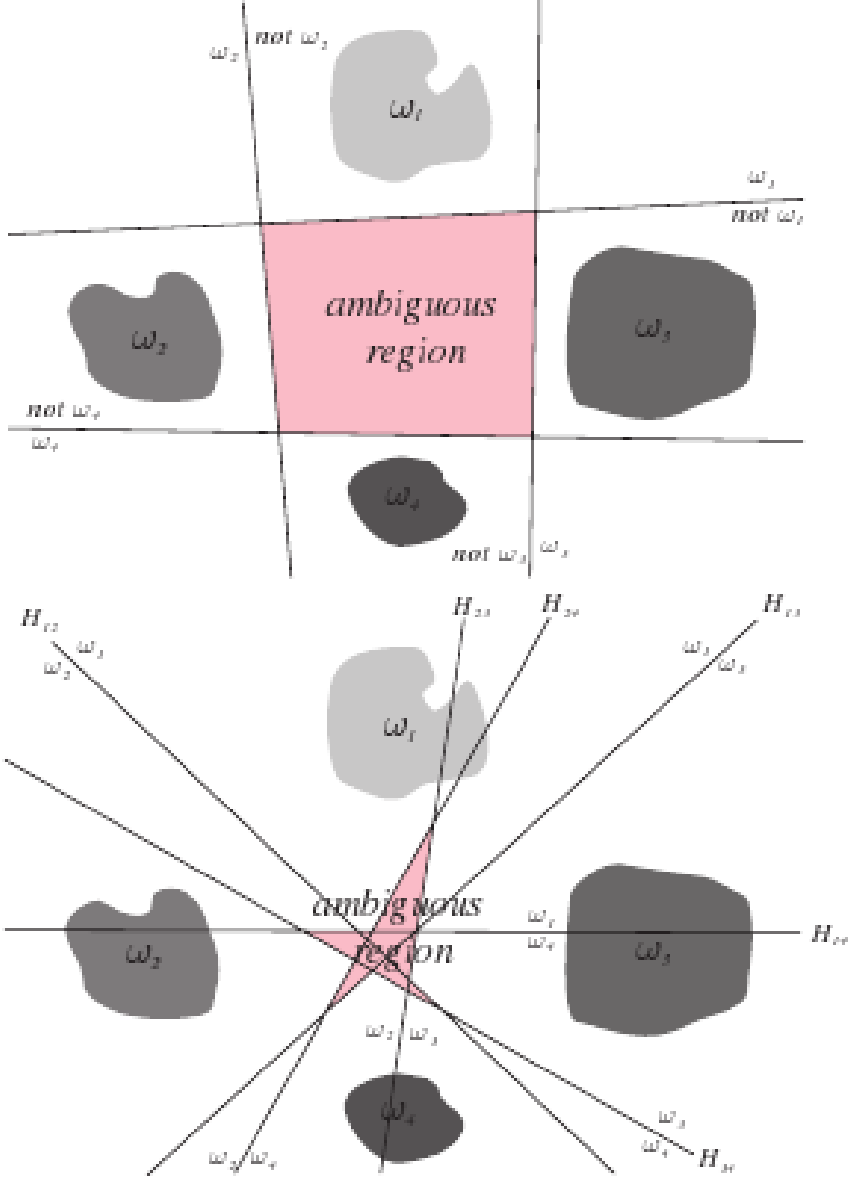

FIGURE 5.3. Linear decision boundaries for a four-class problem. The top figure shows ω_i/not ω_i dichotomies while the bottom figure shows ω_i/ω_j dichotomies and the corresponding decision boundaries H_{ij}. The pink regions have ambiguous category assignments. From: Richard O. Duda, Peter E. Hart,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

- Se definen c funciones discriminantes lineales

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

y se asigna x a ω_i si $g_i(x) > g_j(x)\ \forall j \neq i$; en caso de empates, la clasificacion queda indefinida

- Un clasificador asi se llama “linear machine”
- Una maquina lineal divide el espacio de caracteristicas en c regiones de decision, con $g_i(x)$ el discriminante ma grande si x esta en la region R_i
- dos regiones contiguas R_i and R_j; tienen como frontera que las separa a una porcion de un hiperplano H_{ij} definido por

$$g_i(x) = g_j(x) \text{ sii } (w_i - w_j)^t x + (w_{i0} - w_{j0}) = 0$$

- $w_i - w_j$ es normal a H_{ij}

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

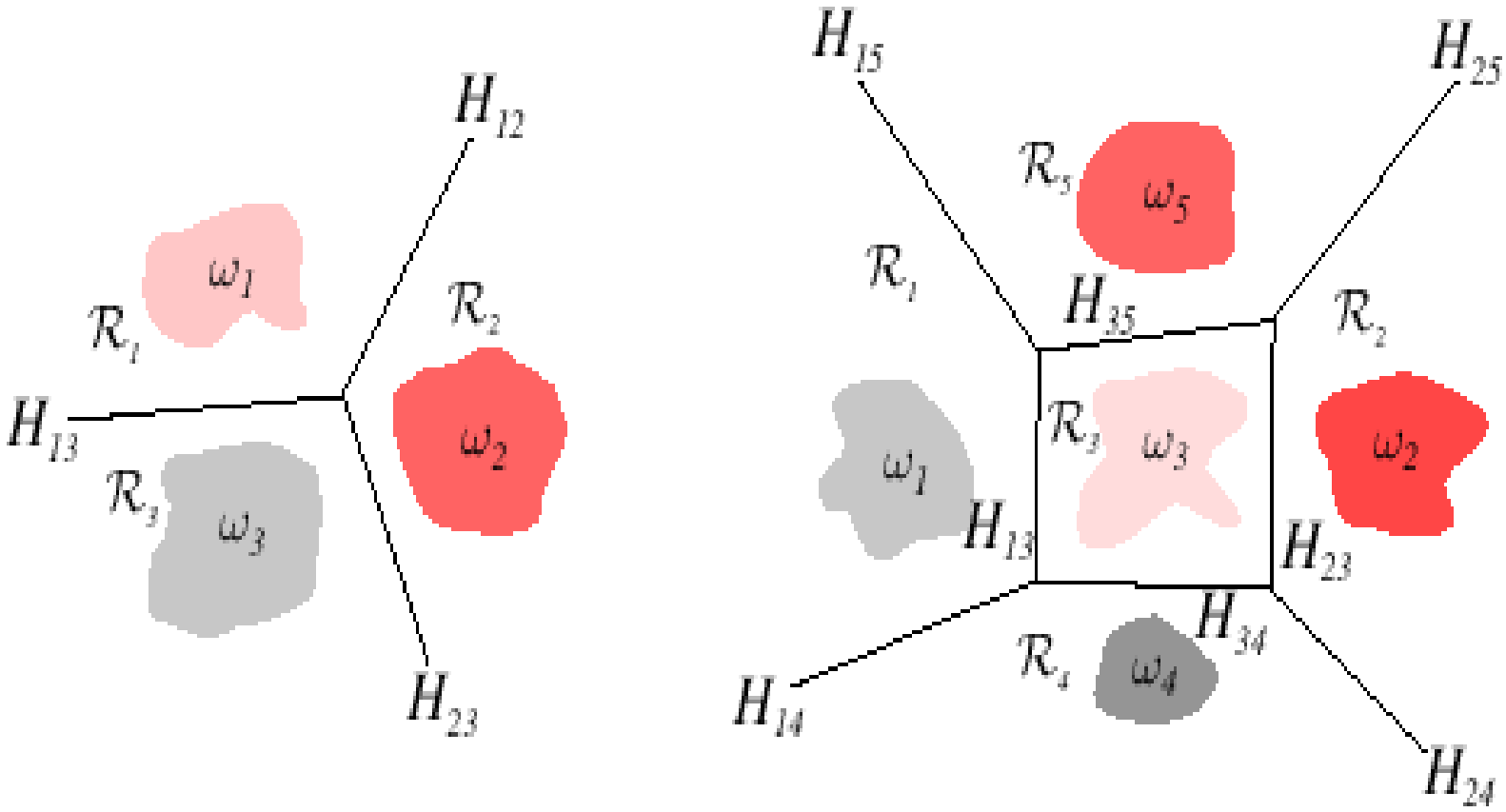

FIGURE 5.4. Decision boundaries produced by a linear machine for a three-class problem and a five-class problem. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- It facil mostrar que las regiones de decision de una maquina lineal son convexas, esta restriccion limita la flexibilidad y la precision del clasificador

Funciones discriminantes lineales generalizadas

- Las fronteras de decision que separan las clases no deberian ser siempre lineales
- La complejidad de las fronteras algunas veces piden el uso de superficies altamente no lineales
- Una forma muy popular de generalizar el concepto de funciones de decision lineal es considerar una funcion de decision generalizada como :

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

donde $f_i(x)$, $1 \leq i \leq N$ son funciones escalares del patron x, $x \in R^n$ *(espacio euclideo)*

- Introduciendo $f_{n+1}(x) = 1$ se tiene:

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

$$donde\ \mathrm{w} = (\mathrm{w}_1, w_2, ..., w_N, w_{N+1})^T \ \mathrm{y}\ \dot{\mathrm{x}} = (\mathrm{f}_1(x), f_2(x), ..., f_N(x), f_{N+1}(x))^T$$

- esta ultima representacion de $g(x)$ implica que toda funcion de decision definida por la ecuacion (1) puede ser tratada como lineal en el espacio $(N + 1)$ dimensional $(N + 1 > n)$

- $g(x)$ mantienen sus caracteristicas no lineales en R^n

- La funcion de decision generalizada mas comun es *g(x)* donde $f_i(x)$ $(1 \leq i \leq N)$ son polinomios

$$g(x) = (\dot{w})^T x$$ T: es la transpuesta

donde $\dot{w}$ es un nuevo vector de pesos, el cual puede ser calculado del vector original w y de las funciones originales $f_i(x)$, $1 \leq i \leq N$

- funciones de decision cuadraticas para un espacio de dos caracteristicas

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

$$aqui: \mathrm{w} = (\mathrm{w}_1, w_2, ..., w_6)^T \text{ y } \dot{\mathrm{x}} = (\mathrm{x}_1^2, x_1 x_2, x_2^2, x_1, x_2, 1)^T$$

- Para patrones en $x \in R^n$, la funcion de decision cuadratica mas usada es:

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

el numero de terminos en el lado derecho es :

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

Este es el numero de pesos que son los parametros libres del problema

- Si por ejemplo n = 3, el vector $\dot{\mathbf{x}}$ es 10-dimensional
- si por ejemplo n = 10, el vector $\dot{\mathbf{x}}$ es 65-dimensional

- En el caso de funciones de decision polinomiales de orden m, una $f_i(x)$ tipica es :

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

$$\text{donde } 1 \le i_1, i_2, \ldots, i_m \le n \text{ y } e_i, 1 \le i \le m \text{ es } 0 \text{ o } 1.$$

  - Si f es polinomial con grado entre 0 y m para no tener repeticiones se pide $i_1 \le i_2 \le \ldots \le i_m$

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

(donde $g^0(x) = w_{n+1}$) es la funcion de decision polynomial de orden m mas general

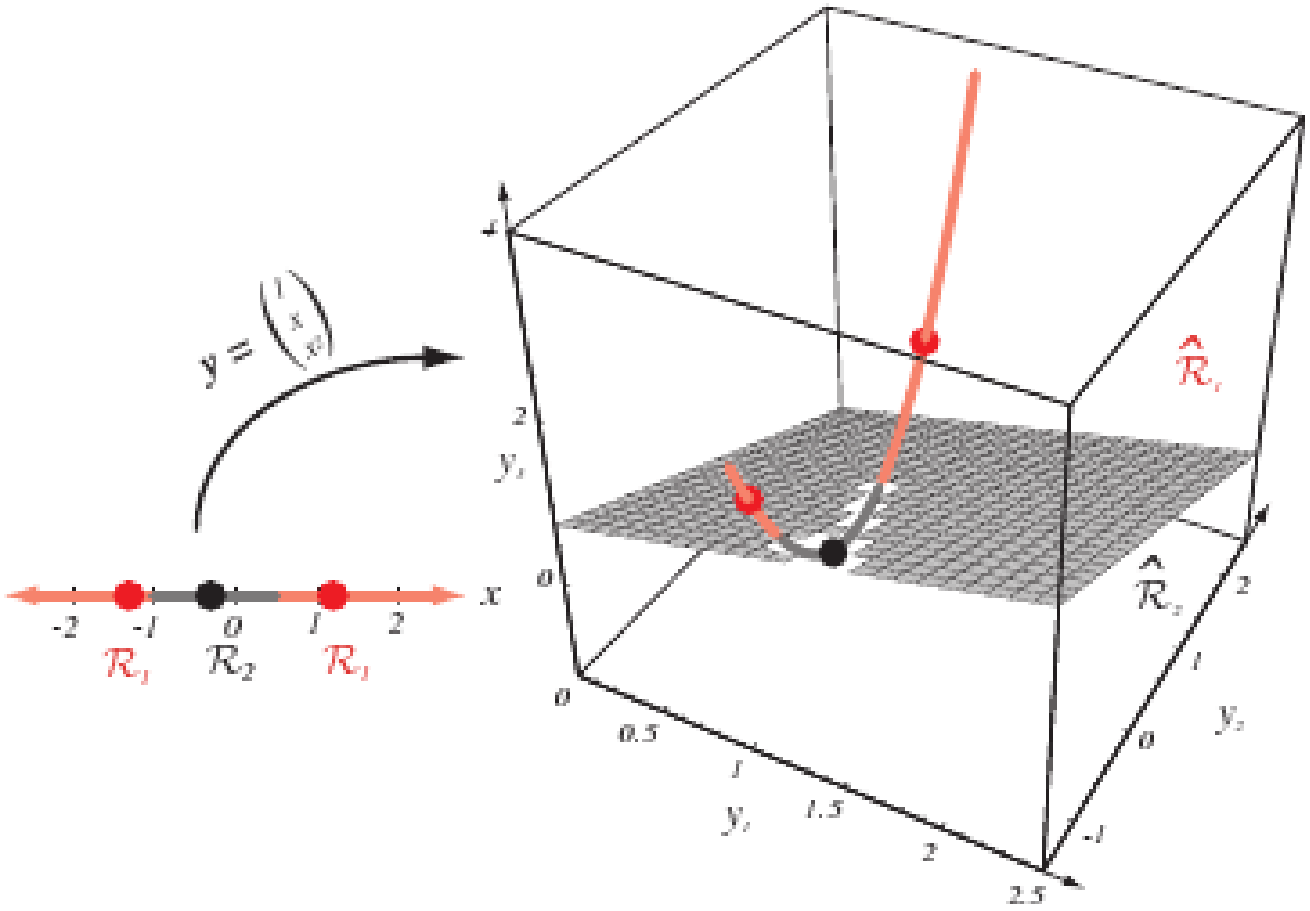

FIGURE 5.5. The mapping $\mathbf{y} = (1, x, x^2)^t$ takes a line and transforms it to a parabola in three dimensions. A plane splits the resulting $\mathbf{y}$-space into regions corresponding to two categories, and this in turn gives a nonsimply connected decision region in the one-dimensional x-space. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

Ejemplo 1: sea $n = 3$ and $m = 2$ entonces:

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

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

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

Ejemplo 2: Sea $n = 2$ y $m = 3$ entonces:

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

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

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

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

- La funcion discriminante cuadratica puede ser representada por una superficie cuadratica n- dimensional

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

donde la matriz $A = (a_{ij})$, el vector $b = (b_1, b_2, \ldots, b_n)^T$ y c, dependen de los pesos w_{ii}, w_{ij}, w_i de la ecuacion (2)

- Si A es definida positiva entonces la funcion de decision es un hiperelipsoide con ejes en las direcciones de los autovalores de A

- En particular: si $A = I_n$ (Identidad), la funcion de decision es simpelemente una hiperesfera n-dimensional

- Si A es definida negativa la funcion de decision describe un a hyperboloide

- En conclusion: es solo la matriz A la que determina la forma y las caracteristicas de la funcion de decision

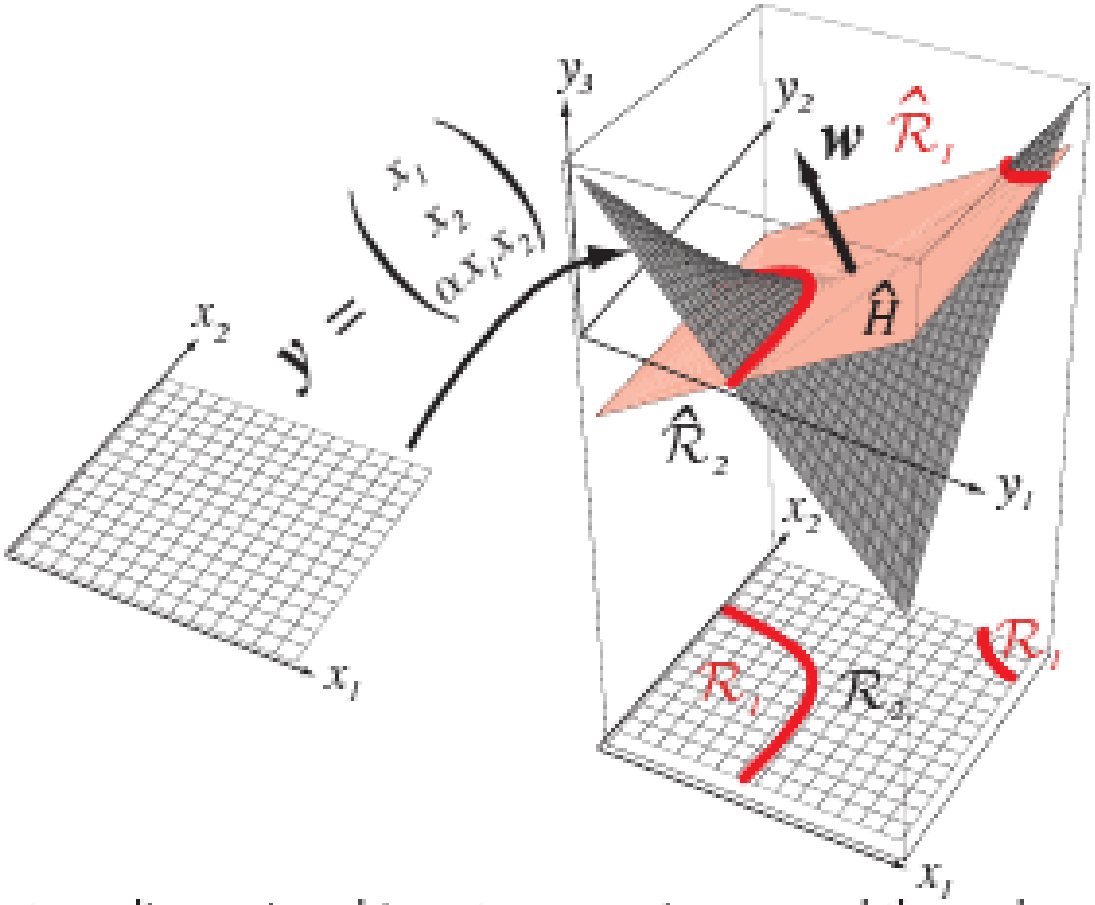

FIGURE 5.6. The two-dimensional input space $\mathbf{x}$ is mapped through a polynomial function f to $\mathbf{y}$. Here the mapping is $y_1 = x_1$, $y_2 = x_2$ and $y_3 \propto x_1 x_2$. A linear discriminant in this transformed space is a hyperplane, which cuts the surface. Points to the positive side 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.

Dos Clases linealmente separables

- Tenemos n muestras de dos clases
- Queremos determinar pesos a en un discriminante lineal

$$g(x)=a^t y$$

- Una muestra y se clasifica correctamente si $a^t y>0$ y esta en la clase 1, o $a^t y<0$ y esta en la clase 2.
- Por lo cual, con una normalizacion, se puede buscar un vector a tal que $a^t y>0$ para todas las muestras
- Un vector asi es llamado vector solucion

solucion lineal

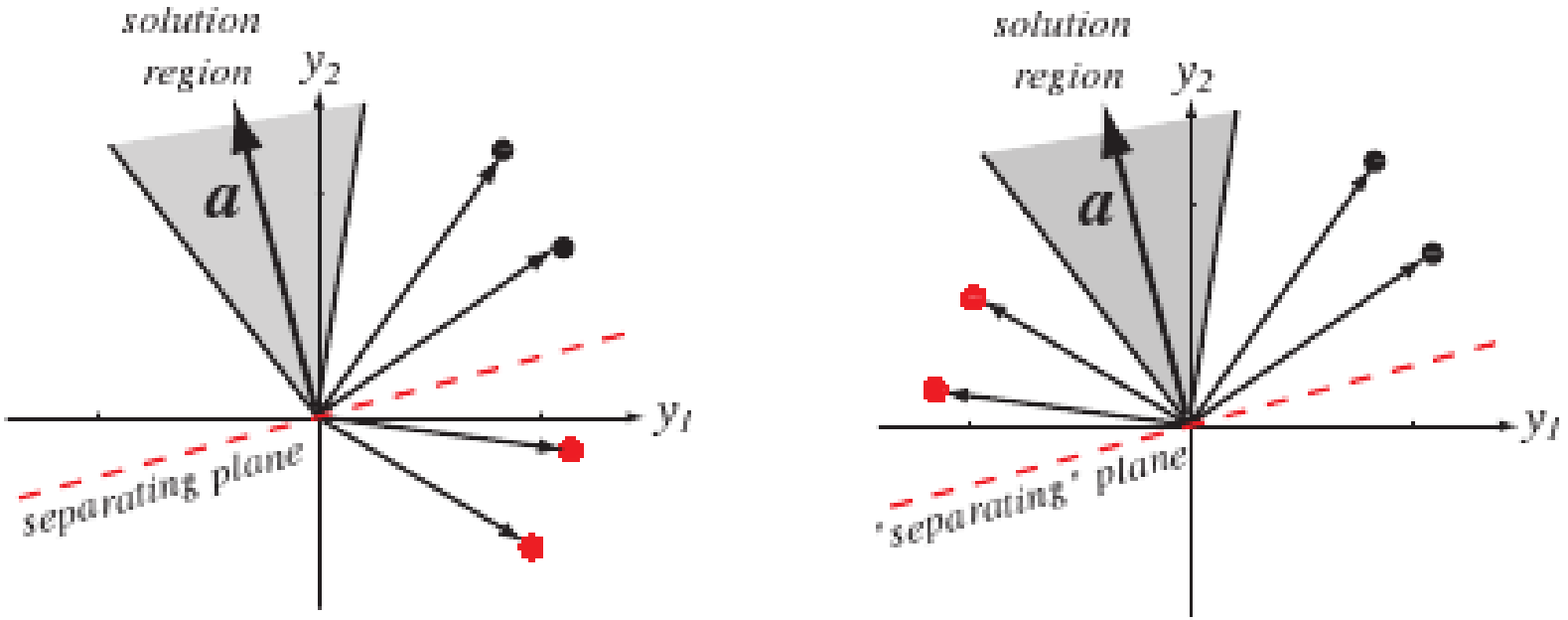

FIGURE 5.8. Four training samples (black for ω_1, red for ω_2) and the solution region in feature space. The figure on the left shows the raw data; the solution vectors leads to a plane that separates the patterns from the two categories. In the figure on the right, the red points have been "normalized"—that is, changed in sign. Now the solution vector leads to a plane that places all "normalized" points on the same side. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

- pesos a son vectores en el espacio de pesos
- cada muestra y es una restriccion a la posicion del vector de pesos
- la ecuacion ay=0 define un hiperplano por el origen del espacio de pesos que tiene a y como vector normal
- por lo cual la solucion tiene que estar en la interseccion de n espacios, y todo vector de ese espacio interseccion es solucion
- Podemos maximizar la minima distancia de las muestras al hiperplano
- Otra posibilidad es buscar el vector a que satisface $a^t y \geq b$ donde b es el margen

con margen

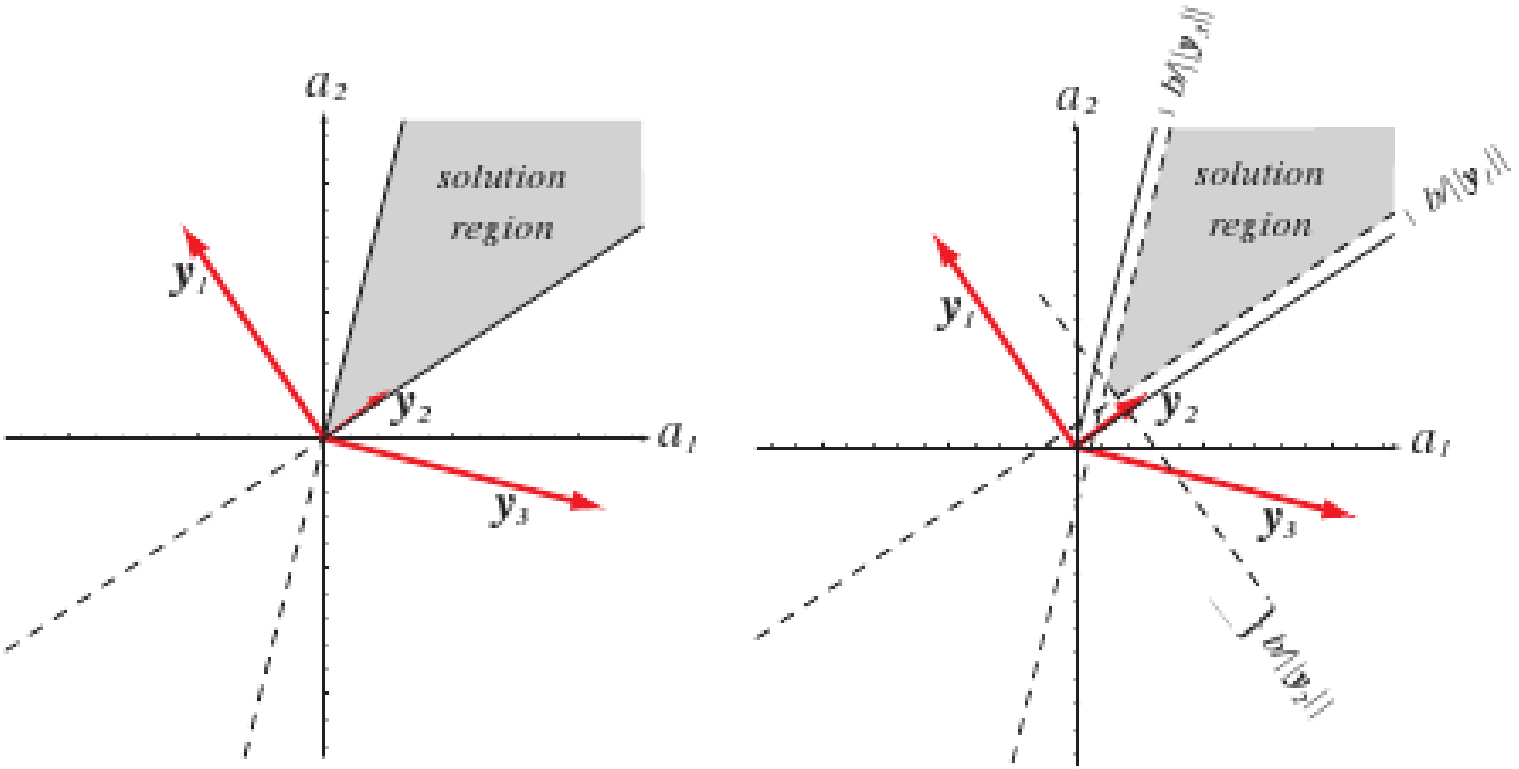

FIGURE 5.9. The effect of the margin on the solution region. At the left is the case of no margin ($b = 0$) equivalent to a case such as shown at the left in Fig. 5.8. At the right is the case $b > 0$, shrinking the solution region by margins $b/\|\mathbf{y}_i\|$. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

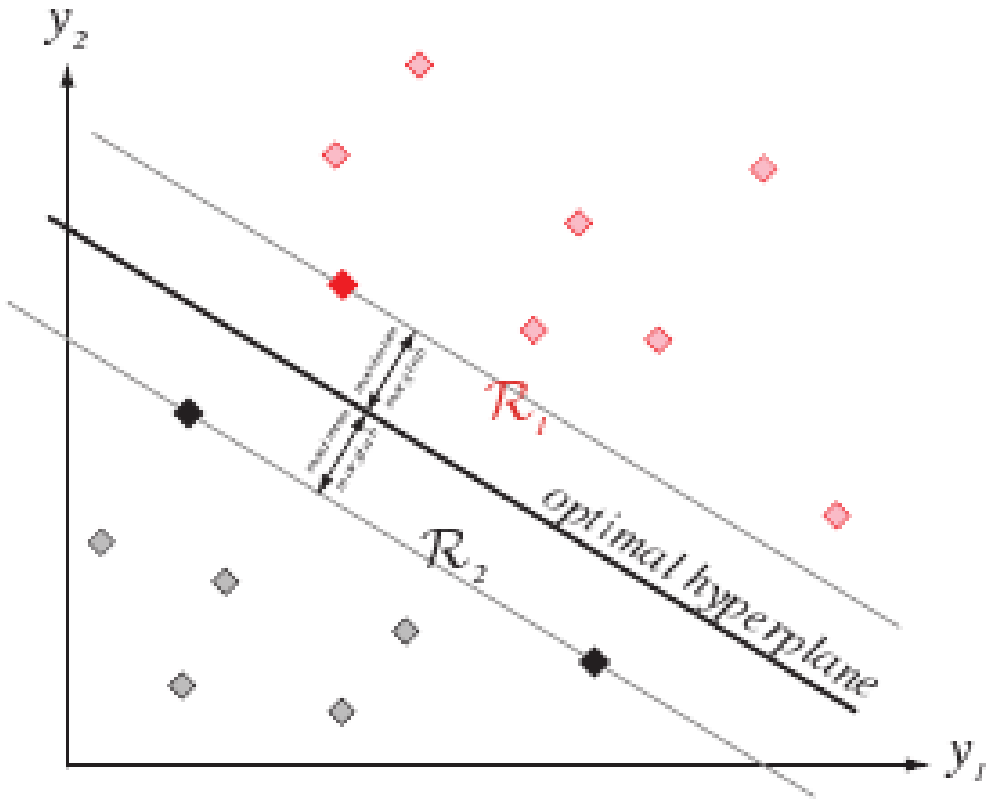

FIGURE 5.19. Training a support vector machine consists of finding the optimal hyperplane, that is, the one with the maximum distance from the nearest training patterns. The support vectors are those (nearest) patterns, a distance b from the hyperplane. The three support vectors are shown as solid dots. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.