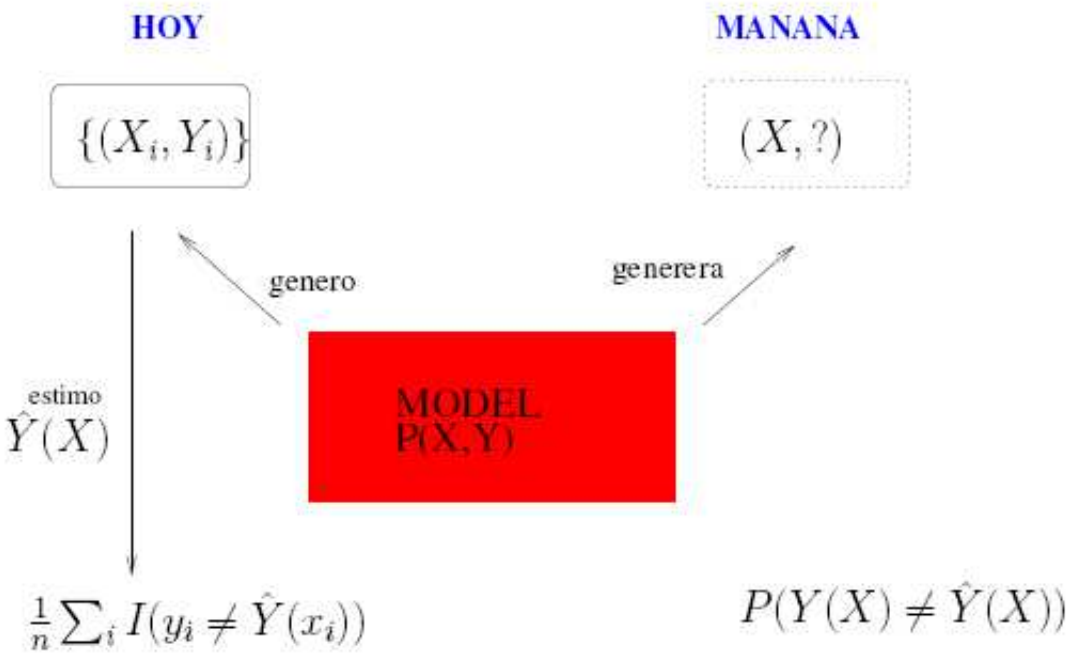
Parte II: Métodos de clasificación

1. Conceptos generales de clasificación
2. Clasificador k-vecino más cercano
3. Clasificador Bayesiano óptimo
4. Análisis discriminante lineal (LDA), enfoque probabilístico vs enfoque geométrico
5. Clasificadores lineales y el Modelo perceptrón
6. Máquinas de soporte vectorial

1. Panorama general



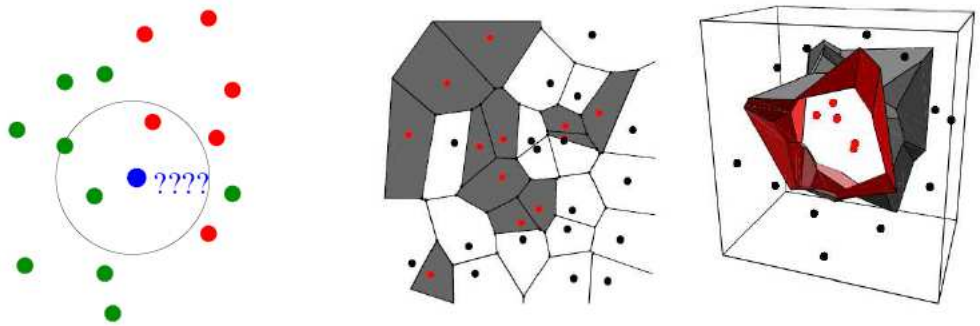
Dos caminos: enfoque geométrico, enfoque probabílistico.

2. Vecino más cercano y k-vecino más cercano

Dado el conjunto $\{(x_i, y_i)\}$

Para una x fija, busca el conjunto C con los k vecinos $\{x_i\}$ más cercanos a x

Regresa la clase que más veces ocurre en C



3. Clasificador Bayesiano óptimo

Dado un clasificador $f : \hat{y}(x) = f(x)$, define una función de perdida $L(y, f(x))$.

Ejemplos: $L(y, f(x)) = I(y \neq f(x)), \quad L(y, f(x)) = (y - f(x))^2, \quad L(y, f(x)) = |y - f(x)|$

Definimos el error como $E(L(Y, f(X))$. El error empírico definimos como $\frac{1}{n}\sum_i L(y_i, f(x_i))$

Observa:

$$E_{X,Y}L(Y, f(X)) = E_X(E_{Y|X}L(Y, f(X))) = \int E_{Y|X=x}L(Y, f(x))dF_X(x)$$

Si minimizamos lo anterior sobre f, es suficiente para cada x minimizar:

$$\arg\min_{f(x)} E_{Y|X=x}L(Y, f(x)), \text{ solución es clasificador Bayesiano óptimo}$$

Si Y toma solamente dos valores:

$$E_{Y|X=x}L(Y, f(x)) = L(0, f(x))P(Y = 0|X = x) + L(1, f(x))P(Y = 1|X = x)$$

$$E_{Y|X=x}L(Y, f(x)) = L(0, f(x))P(Y = 0|X = x) + L(1, f(x))P(Y = 1|X = x)$$

Toma caso binario y el costo de un falso positivo igual a un falso negativo:

si $f(x) = 0 : L(0, f(x)) = 0, \quad L(1, f(x)) = 1$ y el error es $P(Y = 1|X = x)$

si $f(x) = 1 : L(0, f(x)) = 1, \quad L(1, f(x)) = 0$ y el error es $P(Y = 0|X = x)$

Así el clasificador óptimo es

$$f^*(x) = \begin{cases} 0 & \text{si } P(Y = 0|X = x) > P(Y = 1|X = x). \\ 1 & \text{si } P(Y = 1|X = x) \geq P(Y = 0|X = x) \end{cases}$$

En general: asigna x a la categoría más probable según $P(Y|X = x)$.

Observa:

$$P(Y = 0|X = x) > P(Y = 1|X = x) \quad \leftrightarrow \quad P(X = x|Y = 0)P(Y = 0) > P(X = x|Y = 1)P(Y = 1)$$

Si denotamos con L^* el error del clasificador Bayesiano óptimo, y f_n un clasificador basado en $\{(X_i, Y_i)\}_1^n$ y $L(f_n) = E(L(Y, f_n(X))$, se puede demostrar:

Propiedad 1 Si $n \to \infty$ y f_n es el 1-NN:

$$L^* \leq EL(f_n) \leq 2 * L^*$$

Propiedad 2 Si $n \to \infty$ y $k \to \infty$ tal que $k/n \to 0$, si f_n es el k-NN: para cualquier P:

$$EL(f_n) \to L^*$$

Por otro lado:

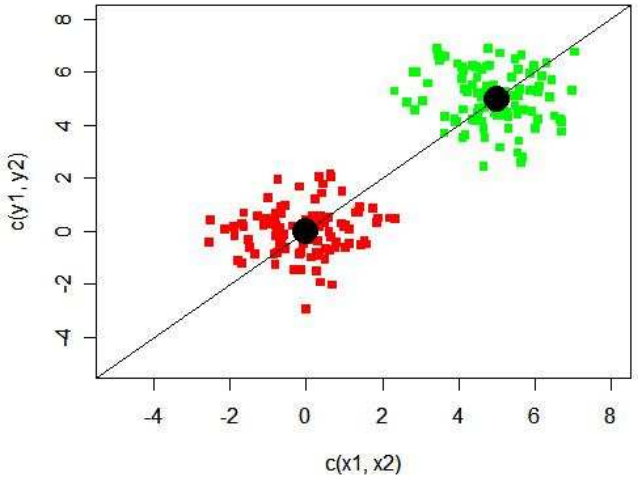
Teorema no free lunch: para n finito, sin ningun supuesto adicional sobre P, ningun clasificador es mejor que otro.

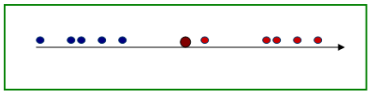
4 Análisis discriminante lineal (LDA)

4.1 enfoque geométrico

Punto de partida:

¿ En qué dirección proyectar los datos para separar los puntos de diferentes clases lo más posible?





Después usamos un clasificador tipo:

$$\hat{y}(x)(= f(x)) \quad = \begin{cases} 0 \text{ si } l^t x < c \\ 1 \text{ en el otro caso} \end{cases}$$

¿ Cómo definir "separar los puntos de diferentes clases lo más posible" ?

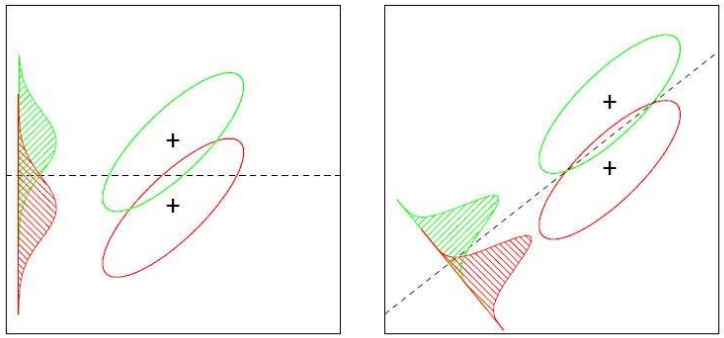
Primera idea:

Considerar los centroides de cada clase como representantes de cada clase.

Separamos estos lo más posible.

Problema:

hay que tomar en cuenta la estructura de la covarianza



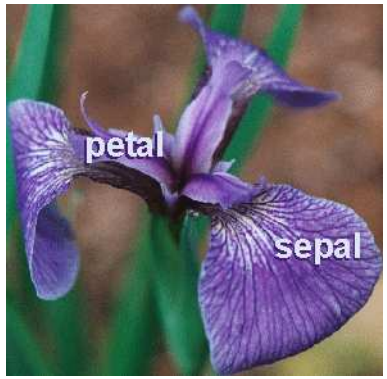
Vamos a suponer que la estructura de covarianza de ambos clases es igual.

Define c_+, c_- como los centroides de cada clase.

Denota con S_W la matriz (muestral) de covarianza (supuesto: no depende de la clase).

Recordando que $Var(l^t X) = l^t Cov(X) l$, buscamos

$$\arg\max_l \frac{(l^t(c_+ - c_-))^2}{l^t S_W l} = \frac{l^t(c_+ - c_-)(c_+ - c_-)^t l}{l^t S_W l} := \frac{l^t S_B l}{l^t S_W l}$$



THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS

By R. A. FISHER, Sc.D., F.R.S.

Table I shows measurements of the flowers of fifty plants each of the two species *Iris setosa* and *I. versicolor*, found growing together in the same colony and measured by Dr E. Anderson, to whom I am indebted for the use of the data. Four flower measurements are given. We shall first consider the question: What linear function of the four measurements $X = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \lambda_4 x_4$ will maximize the ratio of the difference between the specific means to the standard deviations within species? The observed means and their differences are shown in Table II.

Buscamos

$$\arg\max_l = \frac{l^t S_B l}{l^t S_W l}$$

Solución 1

Lo anterior es equivalente a:

$$\max_l l^t S_B l \qquad \text{sujeto a: } l^t S_W l = 1$$

Usar método de lagrange:

$$S_B l = \lambda S_W l,$$

= generalized eigenvalue problem.

Buscamos

$$\arg\max_{l} \frac{l^t S_B l}{l^t S_W l}$$

Solución 2

Usamos la propiedad para una B positiva definida:

$$\arg_{x\neq 0} \frac{(x^t d)^2}{x^t B x} = d^t B^{-1} d,$$

y se alcanza el máximo en $x \sim B^{-1}d$

`incluir demostración`

La dirección que buscamos es dada por

$$\boxed{S_W^{-1}(c_+ - c_-)}$$

Lo que Fischer obtuvo:

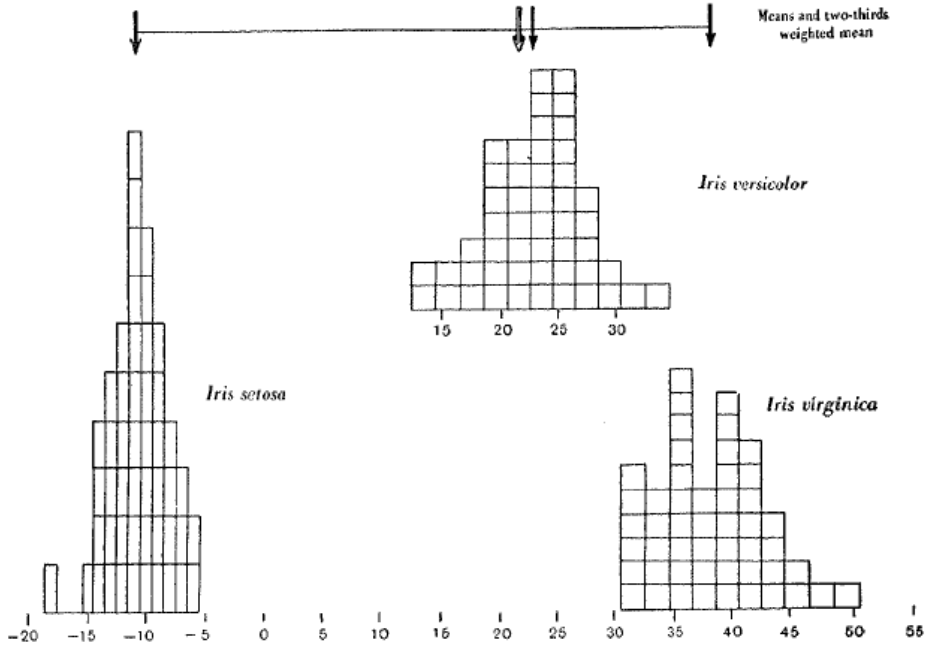


Fig. 1. Frequency histograms of the discriminating linear function, for three species of *Iris*.

También puedes usar R:

```
library(MASS)
 z <- lda(Sp ~ ., Iris, prior =c(1,1)/2)
 p<-predict(z,Iris)
```

4.2 enfoque probabilístico

Retomamos el clasificador Bayesiano óptimo.

Para una función de perdida L, definimos el error $E(L(Y, f(X))$.

Para minimizarla es suficiente para cada x minimizar:

$$\arg\min_{f(x)} E_{Y|X=x} L(Y, f(x)), \text{ solución es clasificador Bayesiano óptimo}$$

En el caso binario y si $L(x, y) = I(x \neq y)$, el clasificador óptimo es

$$f^*(x) = \begin{cases} 0 & \text{si } P(Y = 0|X = x) > P(Y = 1|X = x). \\ 1 & \text{si } P(Y = 1|X = x) \geq P(Y = 0|X = x) \end{cases}$$

Observa:

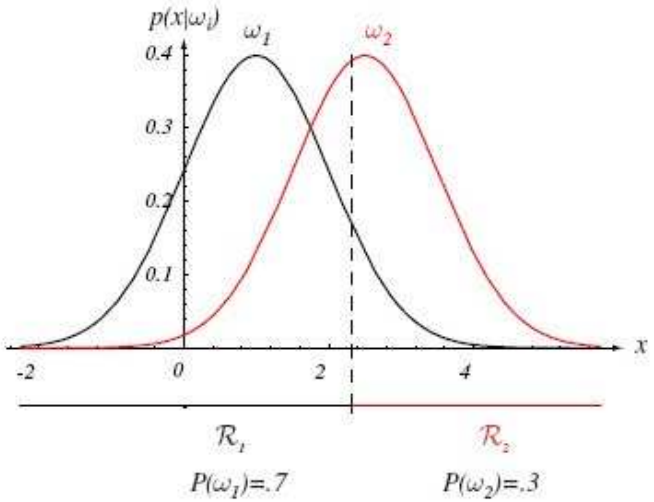
$$P(Y = 0|X = x) > P(Y = 1|X = x) \quad \leftrightarrow \quad P(X = x|Y = 0)P(Y = 0) > P(X = x|Y = 1)P(Y = 1)$$

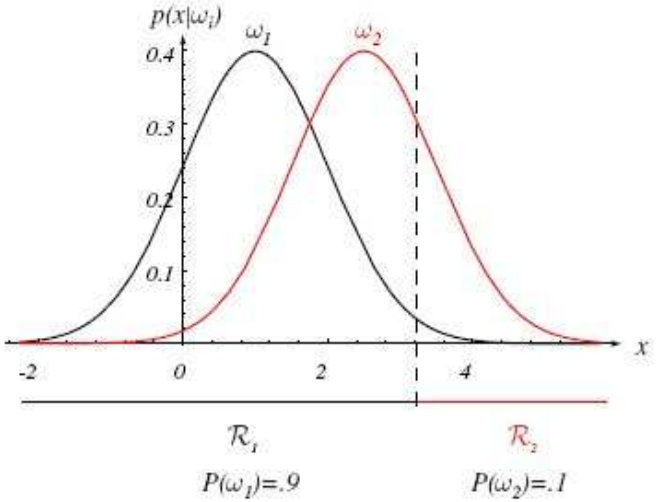
Si $P(Y = 0) = P(Y = 1)$, se asigna x a la categoría más probable según $P(Y|X = x)$.

Obtuvimos:

$$f^*(x) = \begin{cases} 0 & \text{si } P(X = x|Y = 0)P(Y = 0) > P(X = x|Y = 1)P(Y = 1). \\ 1 & \text{si } P(X = x|Y = 1)P(Y = 1) \geq P(X = x|Y = 0)P(Y = 0) \end{cases}$$

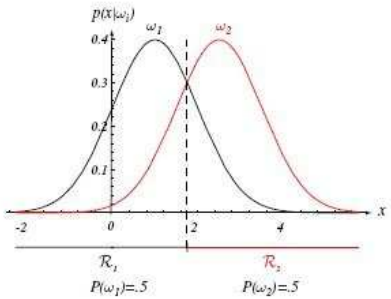
Caso especial: $X|Y = y \sim \mathcal{N}(\mu_y, \sigma^2)$

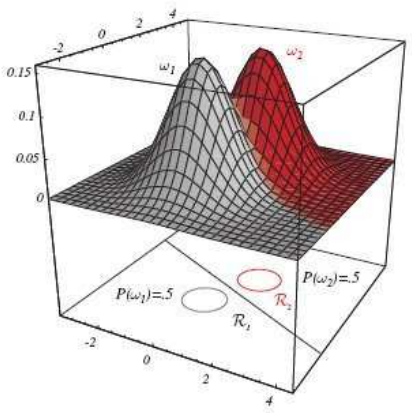


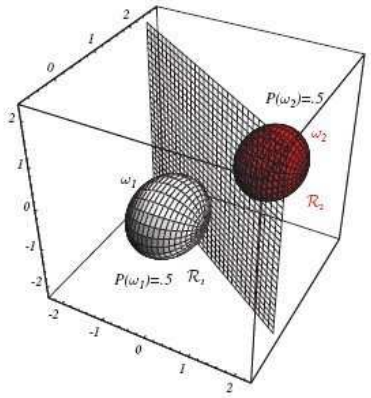


Observa: $f^*(x)$ es de la forma $I(x > c)$.

Caso especial: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma)$

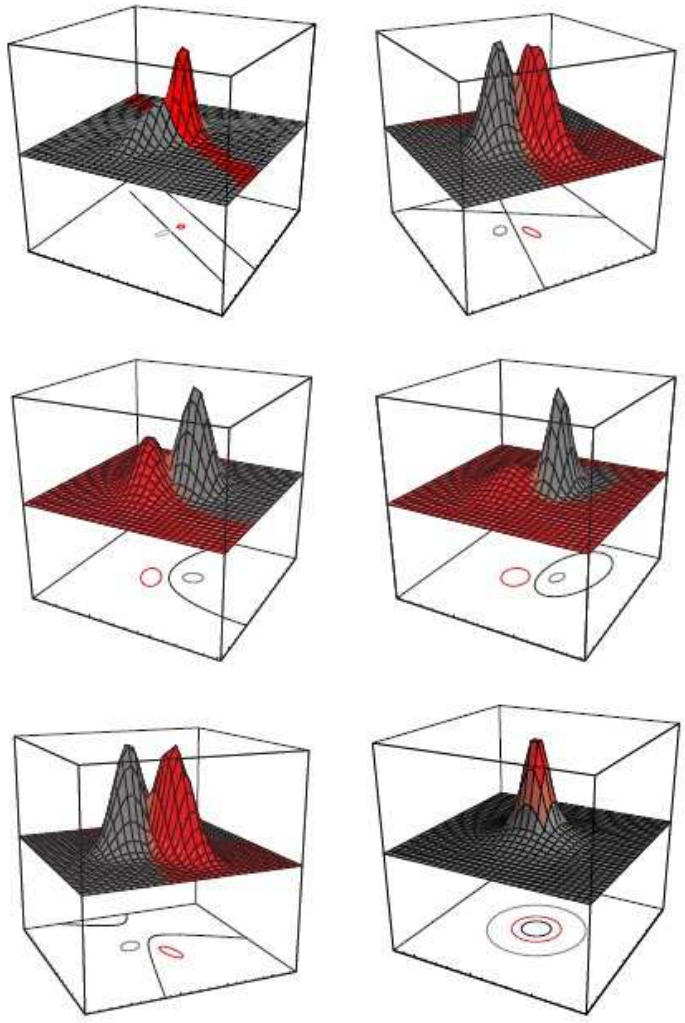






Observa: $f^*(x)$ es de la forma $I(l^t x > c)$ (insertar derivación)

Caso especial: $X|Y = y \sim \mathcal{N}(\mu_y, \Sigma_y)$, covarianza cambia!



Observa: $f^*(x)$ es basada en forma cuadrática (insertar derivación).

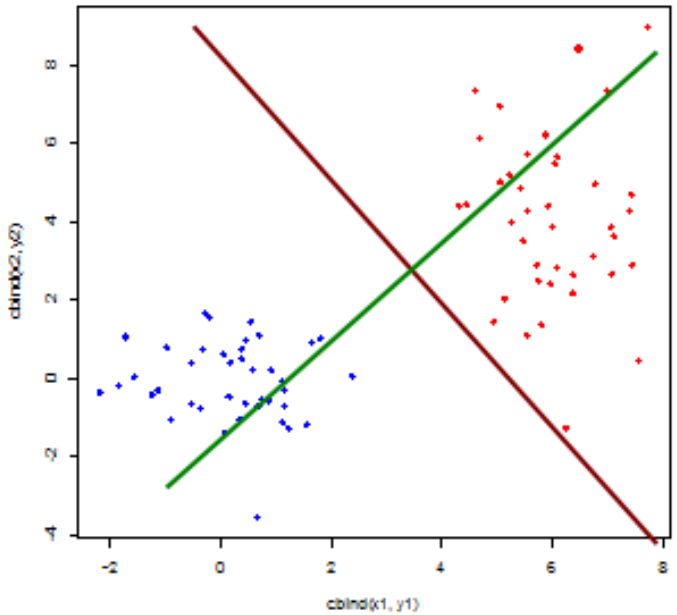
5 Clasificadores lineales y modelo perceptrón

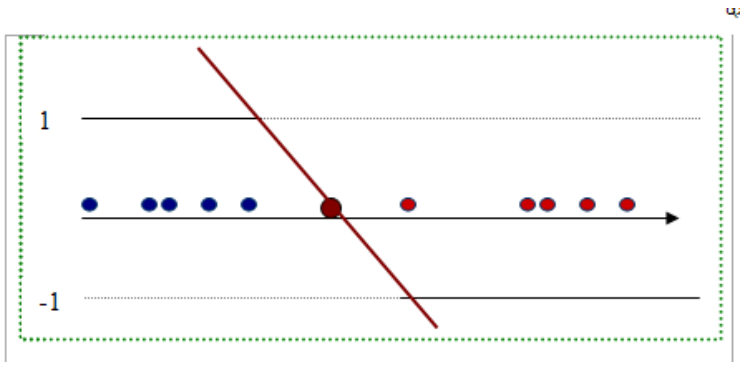
! A partir de ahora: codificamos las categorías como -1 y 1!

Buscamos clasificadores de la forma

$$f(x) = sign(g(x)) = sign(\beta^t x + \alpha).$$

Observa: la frontera entre las dos clases es una linea de contorno de g.



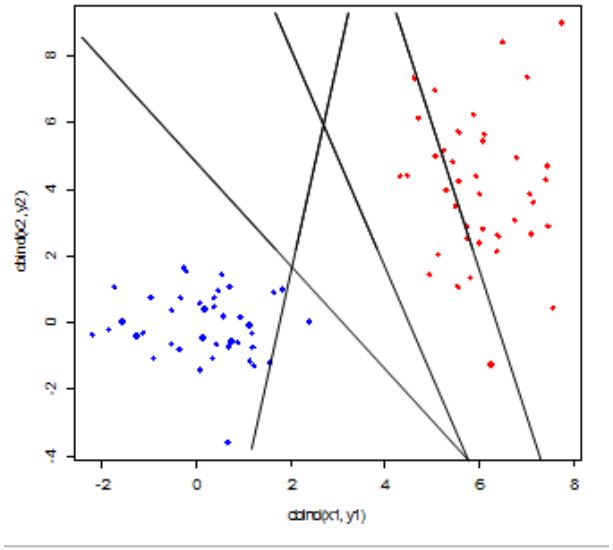


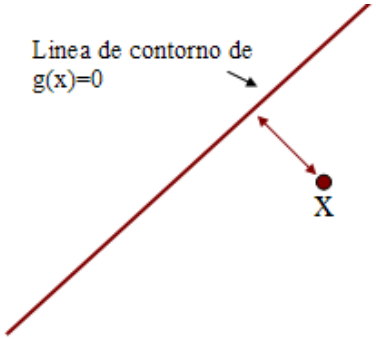
¿ **Cómo determinar** α **y** β**?**: Definir una función de costo: $C(\alpha, \beta)$.

! Supongamos (primero) que los datos son linealmente separables!

Problema: el error empírico $\frac{1}{n}\sum_i I(sign(g_{\alpha,\beta}(x_i)) \neq y_i)$ no es derivable en α, β.

¿ Cómo construir funciones de costo más manejables ?





La distancia entre un punto x y la recta $g_{\alpha,\beta}(x) = 0$ está dada por (insertar demostración)

$$\frac{|g(x)|}{\|\beta\|}$$

Por construcción: x_i está mal clasificado si: $g(x_i)y_i < 0$.

En base de lo anterior, proponemos:

$$C(\alpha, \beta) = -\sum_{i:g(x_i)y_i<0} g(x_i)y_i/\|\beta\|$$

$$-\sum_{i:g(x_i)y_i<0} g(x_i)y_i/\|\beta\| \qquad \text{y}$$

$$\|\beta\|C(\alpha,\beta) = -\sum_{i:g(x_i)y_i<0} g(x_i)y_i$$

tienen el mismo mínimo (si los datos son separables). Preferimos trabajar con

$$C_n(\alpha,\beta) = -\sum_{i:g(x_i)y_i<0} g(x_i)y_i$$

.

Usamos un método tipo gradiente para minimizarlo (insertar derivación):

`elige` α, β

`Repite hasta convergencia`

`para cada dato` (x_i, y_i) `mal clasificado:`

$\beta = \beta + \eta x_i y_i$

$\alpha = \alpha + \eta y_i$

Si los datos son linealmente separables, habrá convergencia en tiempo finito.

Problemas con mínimos locales!!!

Lo anterior es conocido como un ejemplo de un clasificador tipo perceptrón.

Relación con LDA

LDA resuelve un problema de regresión (y también un *optimal scoring problem*):

$$\min_{\alpha,\beta} \sum_i (\theta(y_i) - \alpha - \beta^t x_i)^2,$$

con

$$\theta(y) = \begin{cases} -\frac{n}{n_-} & \text{si } y = -1 \\ \frac{n}{n_+} & \text{si } y = 1 \end{cases},$$

n_+, n_-, número de obervaciones por clase.