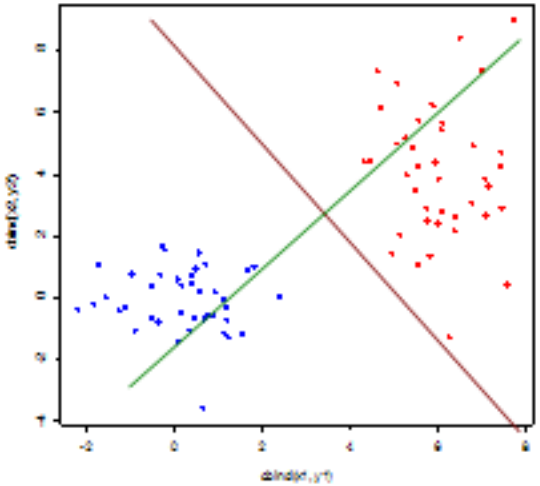
Plan de trabajo para la clase de hoy

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

6 Máquina de soporte vectorial

(seguimos suponiendo que los datos son separables)



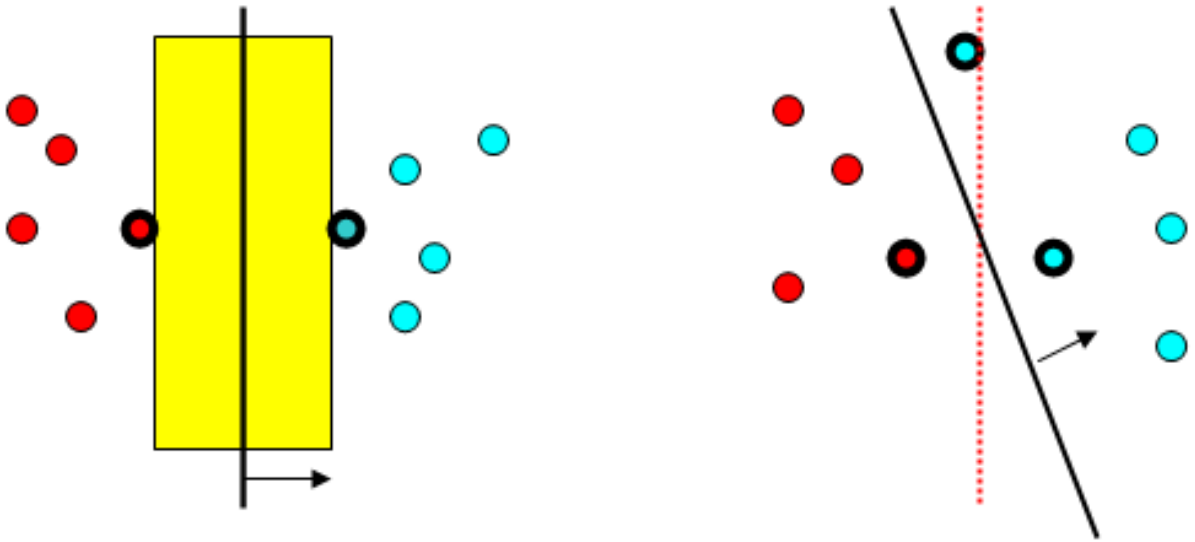
Definimos la **margen** de un clasificador lineal como la distancia más pequena entre una observación y la frontera de decisión.

Idea: buscar clasificador lineal con margen máximo.

$$\max_{\alpha,\beta} C \quad \text{sujeto a } \frac{g(x_i)y_i}{\|\beta\|} > C.$$

Los vectores de soporte son las observaciones que están a una distancia de la frontera de decisión igual al margen.

En 2D si los datos son continuos, hay dos situaciones típicas:



lado izquierdo: 2 vectores de soporte; margen es 0.5*(distancia entre ellos)

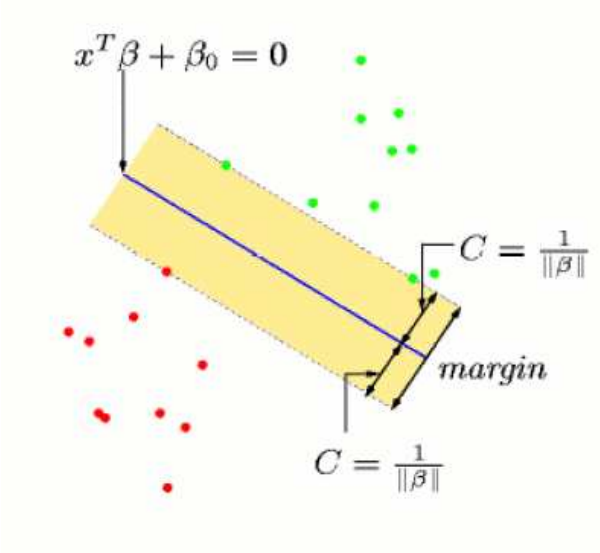
lado derecho: 3 vectores de soporte;

Si hay 4 vectores de soportes o más, habrá colinealidad.

Si cambiamos $g()$ por $cg(), c > 0$, el clasificador f no cambia.

Podemos suponer que g en los puntos más cercanos a la frontera de decisión toma valor 1 o -1.

$\Rightarrow$ margen es $= \frac{1}{\|\beta\|}$.



El problema:

$$\max_{\alpha,\beta} C \quad \text{sujeto a } \frac{g(x_i)y_i}{\|\beta\|} \geq C, \forall i$$

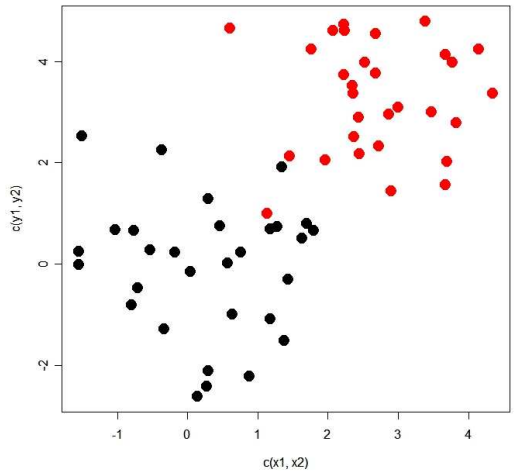
se convierte en:

$$\max_{\alpha,\beta} 1/\|\beta\| \quad \text{sujeto a: } g(x_i)y_i \geq 1, \forall i$$

o equivalente

$$\min_{\alpha,\beta} \|\beta\|^2 \quad \text{sujeto a: } g(x_i)y_i \geq 1, \forall i.$$

Supongamos que los datos no son linealmente separables.



Antes se exigía $\frac{g(x_i)y_i}{\|\beta\|} \geq C, \forall i$

Ahora se introducen variables $\epsilon_i \geq 0$, y se cambia lo anterior en:

$$\frac{g(x_i)y_i}{\|\beta\|} > C(1-\epsilon_i), \forall i$$

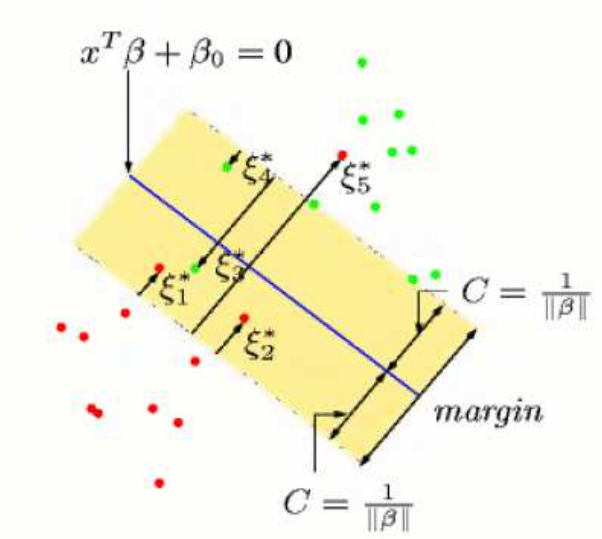
Al mismo tiempo se castigan cada $\epsilon_i \neq 0$, incluyendolos en la función de costo.

$$\min_{\alpha,\beta} \ \|\beta\|^2 + \gamma \sum_i \epsilon_i \quad \text{sujeto a: } g(x_i)y_i \geq 1-\epsilon_i, \quad \text{y} \quad \epsilon_i \geq 0 \ \ \forall i.$$

en lugar de: $\min_{\alpha,\beta} \ \|\beta\|^2$ sujeto a: $g(x_i)y_i > 1, \forall i.$

Interpretación:

$$\min_{\alpha,\beta} \ \|\beta\|^2 + \gamma \sum_i \epsilon_i \quad \text{sujeto a: } g(x_i)y_i \geq 1 - \epsilon_i, \quad \text{y} \quad \epsilon_i \geq 0 \ \ \forall i.$$



ϵ_i es la distancia mínima que hay que *transladar* x_i para que esté al lado correcto de la frontera de decisión y al menos a una distancia de $1/\|\beta\|$.

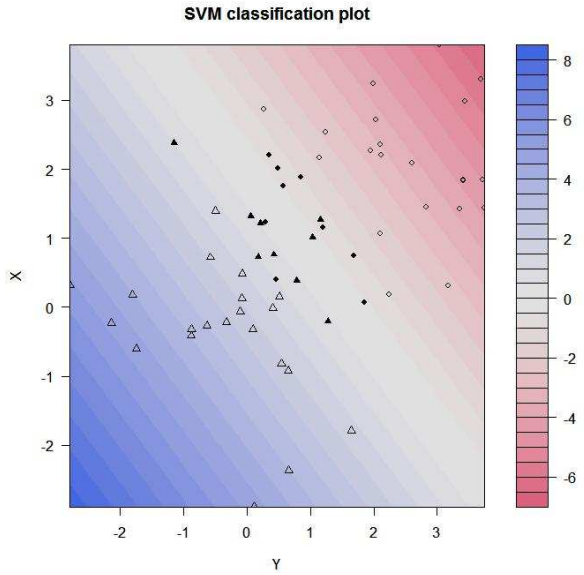
- Si $g(x_i)y_i < 0$ o $g(x_i)y_i \geq 0$ pero $g(x_i)y_i < 1$ podemos tomar ϵ_i tal que $g(x_i)y_i = 1 - \epsilon_i$.
- Si $g(x_i)y_i \geq 1$, podemos tomar $\epsilon_i = 0$.

Función de costo es equivalente a: $\min_{\alpha,\beta} \ \sum_i (1 - g(x_i)y_i)_+ + \gamma\|\beta\|^2$.

Se puede mostrar que la solucíon es de la forma:

$$g(x) = \sum \alpha_i \langle x, x_i \rangle + b.$$

Solamente algunas α_is son diferentes de 0: corresponden a observaciones con $\epsilon_i \neq 0$
(=los llamamos vectores de soporte).




```
library(kernlab)
s<-ksvm(categoria~X+Y,data=d,kernel="p",cost=1,kpar=list(degree=1,offset=0))
```

Alternativa

```
library(e1071); s<-svm(X58~.,kernel="l",data=spam.train)
```