

2023-2024

Aprendizaje Automático

5. Máquinas de vectores soporte



Francisco Casacuberta Nolla
(fcn@dsic.upv.es)

Alfons Juan Císcar
(ajuan@dsic.upv.es)

(Con material de Enrique Vidal Ruiz)

Departament de Sistemes Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 Núcleos ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 Aplicaciones ▷ 49
- 6 Notación ▷ 52

Index

- 1 *Funciones discriminantes lineales* ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 Núcleos ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 Aplicaciones ▷ 49
- 6 Notación ▷ 52

Clasificación en dos clases con funciones discriminantes lineales

FUNCIÓN DISCRIMINANTE LINEAL (FDL)

$$\phi : \mathbb{R}^d \rightarrow \mathbb{R} : \phi(\mathbf{x}; \Theta) = \boldsymbol{\theta}^t \mathbf{x} + \theta_0 = \sum_{i=1}^d \theta_i x_i + \theta_0$$

$\Theta = (\boldsymbol{\theta}, \theta_0)$: $\boldsymbol{\theta} \in \mathbb{R}^d$ es un *vector de pesos* y $\theta_0 \in \mathbb{R}$ se denomina *umbral*.

El número de parámetros de Θ es pues $D = d + 1$.

REGLA DE CLASIFICACIÓN (2 CLASES)

Asumiendo que las etiquetas de clase, c , son $+1$ y -1 :

$$f(\mathbf{x}) = \begin{cases} +1 & \text{si } \phi(\mathbf{x}; \Theta) \geq 0 \\ -1 & \text{si } \phi(\mathbf{x}; \Theta) < 0 \end{cases}$$

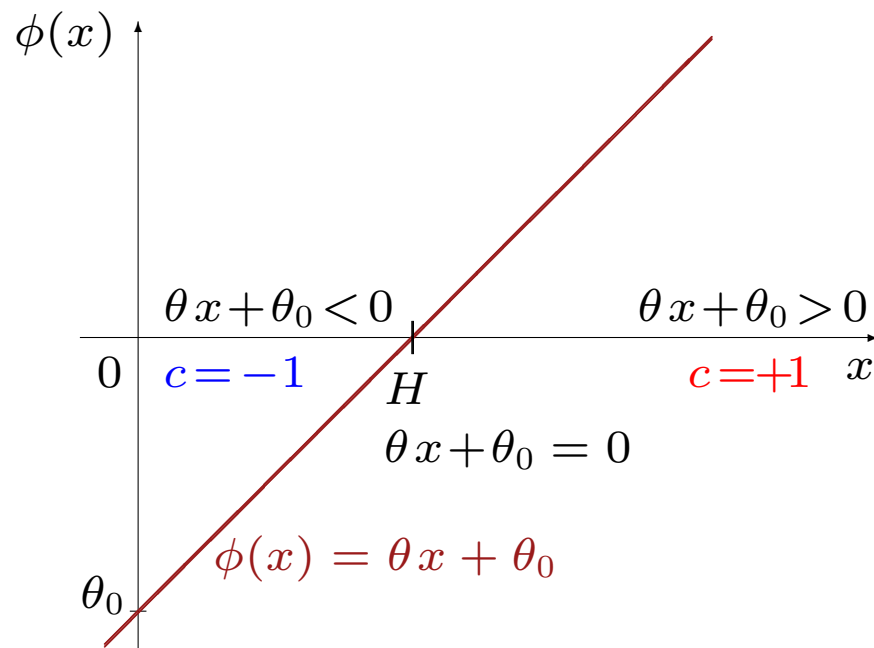
Propiedades de las funciones discriminantes lineales

1. Una FDL ϕ define el hiperplano de decisión $H = \{x \mid \phi(x; \Theta) = 0\}$.
2. H divide a \mathbb{R}^d en dos semiespacios: $\phi(x; \Theta) \geq 0$, $c = +1$ y $\phi(x; \Theta) < 0$, $c = -1$.
3. Si $\gamma \in \mathbb{R}^+$, entonces $\gamma \phi(x; \Theta)$ y $\phi(x; \Theta)$ representan al mismo H .

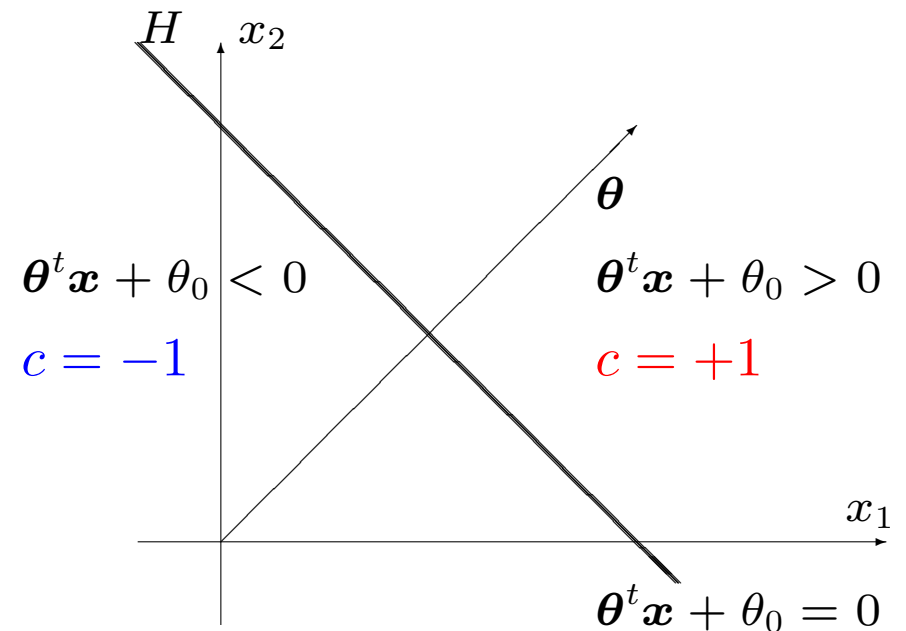
Propiedades de las funciones discriminantes lineales

1. Una FDL ϕ define el hiperplano de decisión $H = \{x \mid \phi(x; \Theta) = 0\}$.
2. H divide a \mathbb{R}^d en dos semiespacios: $\phi(x; \Theta) \geq 0$, $c = +1$ y $\phi(x; \Theta) < 0$, $c = -1$.
3. Si $\gamma \in \mathbb{R}^+$, entonces $\gamma \phi(x; \Theta)$ y $\phi(x; \Theta)$ representan al mismo H .

Ejemplo con $d = 1$



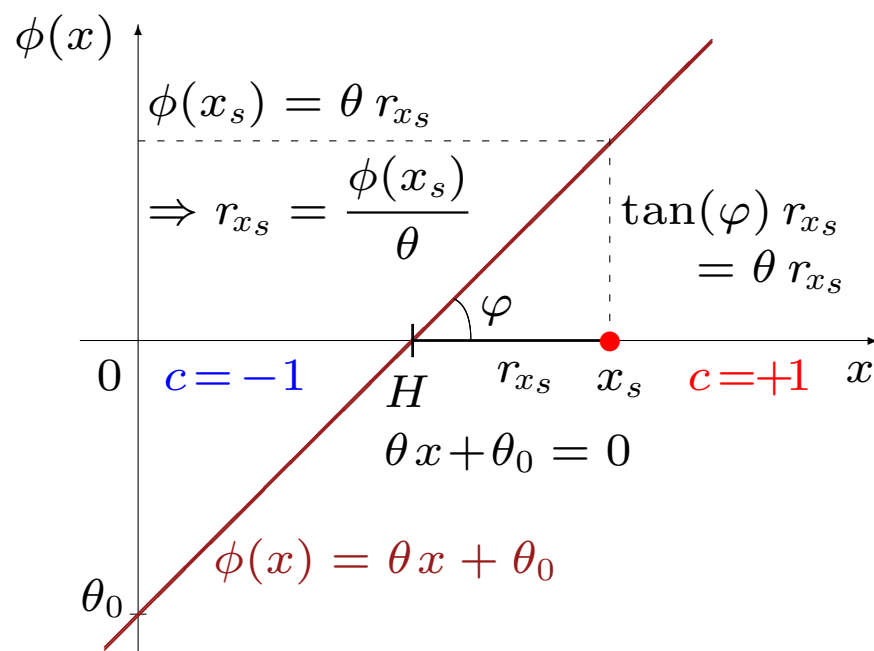
Ejemplo con $d = 2$



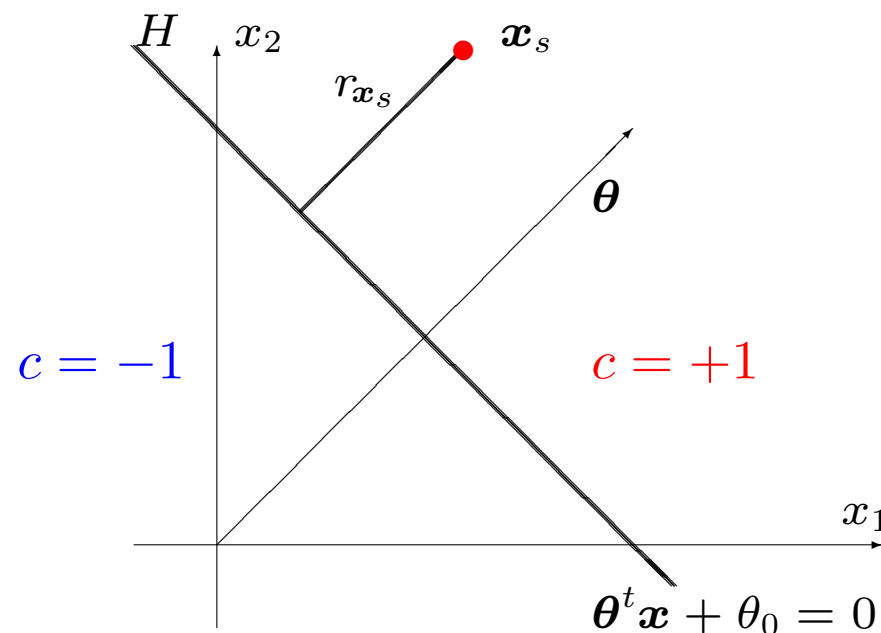
Propiedades de las funciones discriminantes lineales

1. Una FDL ϕ define el hiperplano de decisión $H = \{x \mid \phi(x; \Theta) = 0\}$.
2. H divide a \mathbb{R}^d en dos semiespacios: $\phi(x; \Theta) \geq 0$, $c = +1$ y $\phi(x; \Theta) < 0$, $c = -1$.
3. Si $\gamma \in \mathbb{R}^+$, entonces $\gamma \phi(x; \Theta)$ y $\phi(x; \Theta)$ representan al mismo H .
4. La distancia de cualquier punto x_s a H es: $r_{x_s} = \frac{|\phi(x_s; \Theta)|}{\|\theta\|} = \frac{|\theta^t x_s + \theta_0|}{\|\theta\|}$

Ejemplo con $d = 1$



Ejemplo con $d = 2$



Aprendizaje de funciones discriminantes lineales

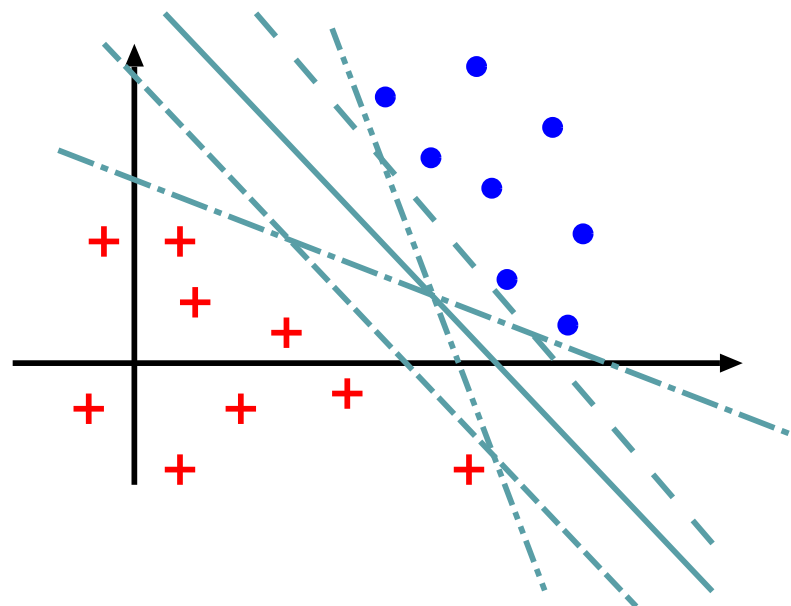
$$S = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}, \quad \mathbf{x}_n \in \mathbb{R}^d, c_n \in \{+1, -1\}, 1 \leq n \leq N.$$

S es *linealmente separable* si $\exists \boldsymbol{\theta} \in \mathbb{R}^d, \theta_0 \in \mathbb{R}$, tales que:

$$c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) > 0, \quad 1 \leq n \leq N$$

Aprendizaje: Dada una muestra linealmente separable S , encontrar $\Theta = (\boldsymbol{\theta}, \theta_0)$ que la separe.

Aproximación usual: Minimizar alguna función objetivo $q_S(\boldsymbol{\theta}, \theta_0)$ utilizando descenso por gradiente. Por ejemplo: el algoritmo Perceptrón, o el algoritmo Adaline.



Problema: probablemente hayan muchas soluciones.

Soluciones con *margen* $b \in \mathbb{R}^{\geq 0}$: $c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) \geq b$

Ejercicio: Escribir el algoritmo Perceptrón con margen

Forma canónica respecto a un conjunto de puntos

Para un hiperplano separador H dado, hay múltiples posibilidades de definirlo mediante diferentes FDLs $\phi(\mathbf{x}; \Theta)$.

La *FDL canónica* de un H dado con respecto a un conjunto S de N puntos se define por $\check{\Theta} \equiv (\check{\boldsymbol{\theta}}, \check{\theta}_0)$, tal que:

$$\min_{1 \leq n \leq N} | \phi(\mathbf{x}_n; \check{\Theta}) | = \min_{1 \leq n \leq N} | \check{\boldsymbol{\theta}}^t \mathbf{x}_n + \check{\theta}_0 | = 1$$

Por tanto, la distancia \check{r} del vector $\check{\mathbf{x}} \in S$ más próximo al hiperplano separador H es:

$$\check{r} = \frac{| \check{\boldsymbol{\theta}}^t \check{\mathbf{x}} + \check{\theta}_0 |}{\| \check{\boldsymbol{\theta}} \|} = \frac{1}{\| \check{\boldsymbol{\theta}} \|}$$

Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 *Clasificadores de margen máximo: SVM* ▷ 7
- 3 Núcleos ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 Aplicaciones ▷ 49
- 6 Notación ▷ 52

Forma canónica y margen de un clasificador respecto a un conjunto de puntos

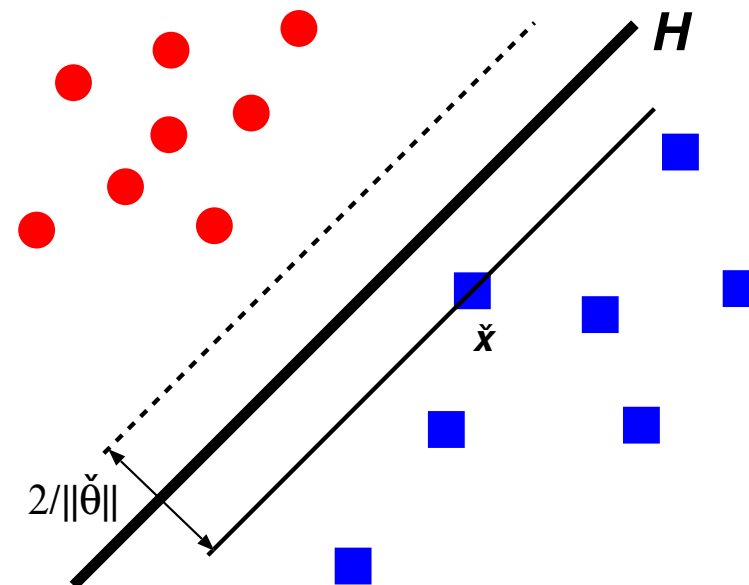
Dado un hiperplano separador H y su *FDL canónica* con respecto a un conjunto S de N puntos $\check{\Theta} \equiv (\check{\theta}, \check{\theta}_0)$.

Hemos visto que la distancia \check{r} del vector $\check{x} \in S$ más próximo al hiperplano separador H es:

$$\check{r} = \frac{1}{\|\check{\theta}\|}$$

Y el *margen* de H con respecto a S se define como:

$$2\check{r} = \frac{2}{\|\check{\theta}\|}$$



En adelante se asume que Θ es siempre canónico respecto a S ; es decir $\check{\Theta} \rightarrow \Theta$

Clasificadores de margen máximo

- *Aprendizaje*: dada una muestra linealmente separable S , encontrar $\theta \in \mathbb{R}^d$ y $\theta_0 \in \mathbb{R}$ que:

- *maximicen*: $\frac{2}{\|\theta\|}$

- *sujetas a*: $c_n (\theta^t x_n + \theta_0) \geq 1, \quad 1 \leq n \leq N$

- Equivalentemente, *buscar* $\theta \in \mathbb{R}^d$ y $\theta_0 \in \mathbb{R}$ *que*:

- *minimicen*: $\frac{1}{2} \theta^t \theta$

- *sujetas a*: $c_n (\theta^t x_n + \theta_0) \geq 1, \quad 1 \leq n \leq N$

Aplicación de la técnica de los multiplicadores de Lagrange

- Función de Lagrange:

$$\Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta} - \sum_{n=1}^N \alpha_n (c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) - 1)$$

donde $\alpha_n \geq 0$, $1 \leq n \leq N$ son los *multiplicadores de Lagrange*.

- Resolver $\nabla_{\boldsymbol{\theta}, \theta_0} \Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha}) = \mathbf{0}$

$$\nabla_{\boldsymbol{\theta}} \Lambda = \mathbf{0} \Rightarrow \boldsymbol{\theta}^* = \sum_{n=1}^N c_n \alpha_n \mathbf{x}_n; \quad \frac{\partial \Lambda}{\partial \theta_0} = 0 \Rightarrow \sum_{n=1}^N \alpha_n c_n = 0$$

- Lagrangiana dual (sustituyendo las anteriores expresiones en $\Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha})$):

$$\Lambda_D(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N c_n c_m \alpha_n \alpha_m \mathbf{x}_n^t \mathbf{x}_m$$

- Maximizar $\Lambda_D(\boldsymbol{\alpha})$ sujeto a: $\sum_{n=1}^N \alpha_n c_n = 0$; $\alpha_n \geq 0$, $1 \leq n \leq N \longrightarrow \boldsymbol{\alpha}^*$

Ejercicio: Desarrollar completamente los pasos anteriores hasta obtener $\Lambda_D(\boldsymbol{\alpha})$.

Maximización del margen: problemas equivalentes

- *Original*: minimizar $\frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta}$, sujeto a: $c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) \geq 1, 1 \leq n \leq N$
- *Primal*: minimizar $\Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha}) = \frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta} - \sum_{n=1}^N \alpha_n (c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) - 1)$
sujeto a $\alpha_n \geq 0, 1 \leq n \leq N \quad \longrightarrow \quad \boldsymbol{\theta}^*(\boldsymbol{\alpha})$

Dual: maximizar $\Lambda_D(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N c_n c_m \alpha_n \alpha_m \mathbf{x}_n^t \mathbf{x}_m$
sujeto a: $\sum_{n=1}^N \alpha_n c_n = 0; \quad \alpha_n \geq 0, 1 \leq n \leq N \quad \longrightarrow \quad \boldsymbol{\alpha}^*$

Los dos son problemas de *optimización cuadrática*, para los que existen técnicas de optimización más o menos costosas; típicamente en $\mathcal{O}(N^3)$.

Ventajas de la formulación dual: Permite soluciones computacionales más eficientes.

Resumen de propiedades de los clasificadores de máximo margen

Las soluciones θ^* , θ_0^* , α^* verifican:

$$1. \theta^* = \sum_{n=1}^N c_n \alpha_n^* \mathbf{x}_n$$

$$2. \sum_{n=1}^N \alpha_n^* c_n = 0$$

$$3. \alpha_n^* \geq 0, \quad 1 \leq n \leq N$$

4. Condición complementaria de Karush-Kuhn-Tucker (KKT):

$$\alpha_n^* \left(c_n (\theta^{*t} \mathbf{x}_n + \theta_0^*) - 1 \right) = 0, \quad 1 \leq n \leq N$$

Esto implica que hay dos posibilidades para cada n :

$$\alpha_n^* = 0, \text{ o bien } \alpha_n^* \neq 0, \quad c_n (\theta^{*t} \mathbf{x}_n + \theta_0^*) = 1$$

Vectores soporte

- **Vectores soporte:** muestras de entrenamiento x_n para las que $\alpha_n^* \neq 0$

$$\mathcal{V} = \left\{ n \in \mathbb{N}, 1 \leq n \leq N \mid (x_n, c_n) \in S, c_n(\theta^{*t} x_n + \theta_0^*) = 1 \right\}$$

- Todos los vectores soporte equidistan del hiperplano separador:

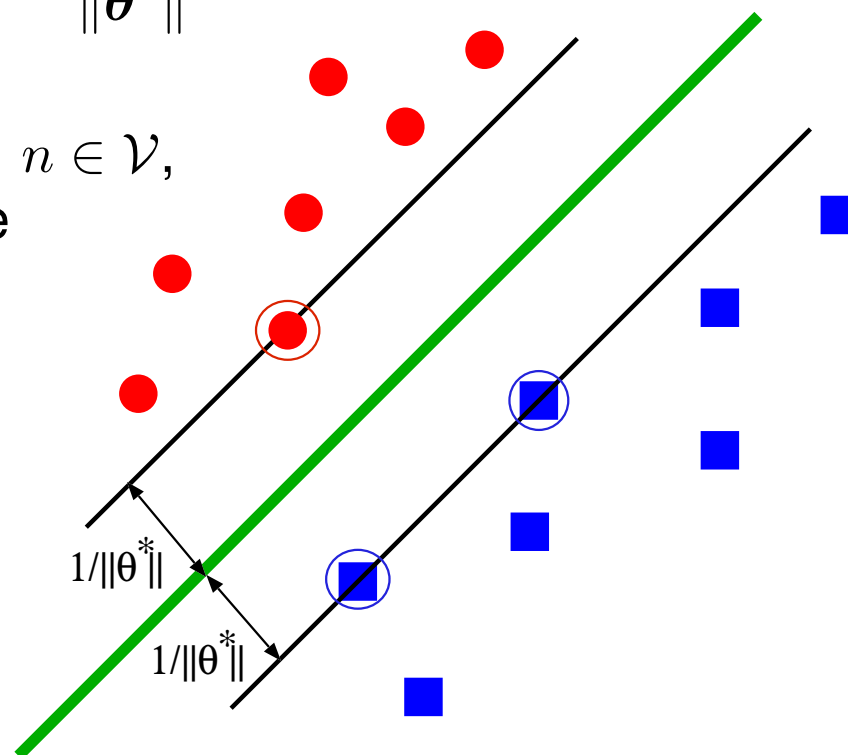
$$\forall n \in \mathcal{V}, r_{x_n} = \frac{|\theta^{*t} x_n + \theta_0^*|}{\|\theta^*\|} = \frac{|c_n|}{\|\theta^*\|} = \frac{1}{\|\theta^*\|}$$

- Las propiedades: $\sum_{n=1}^N \alpha_n^* c_n = 0$ y $\alpha_n^* > 0, n \in \mathcal{V}$, implican que hay al menos un vector soporte de cada clase; es decir, $\exists n, n' \in \mathcal{V}$ tales que

$$c_n = +1, c_{n'} = -1$$

$$\alpha_n^*, \alpha_{n'}^* > 0,$$

$$n \neq n'$$



Máquinas de vectores soporte

Un clasificador de máximo margen queda definido por la función discriminante lineal $\phi(\mathbf{x}; \Theta) \stackrel{\text{def}}{=} \boldsymbol{\theta}^{\star t} \mathbf{x} + \theta_0^{\star}$, donde $\boldsymbol{\theta}^{\star}, \theta_0^{\star}$ son parámetros óptimos del problema *original* (maximizar el margen), o de los correspondientes problemas *primal–dual*.

$\boldsymbol{\theta}^{\star}$ se obtiene mediante combinación lineal de vectores soporte, por lo que estos clasificadores también se denominan *máquinas de vectores soporte*.

- Por la primera propiedad: $\boldsymbol{\theta}^{\star} = \sum_{n=1}^N c_n \alpha_n^{\star} \mathbf{x}_n = \sum_{n \in \mathcal{V}} c_n \alpha_n^{\star} \mathbf{x}_n$
- Por KKT, para cualquier $m \in \mathcal{V}$: $\theta_0^{\star} = c_m - \boldsymbol{\theta}^{\star t} \mathbf{x}_m = c_m - \sum_{n \in \mathcal{V}} c_n \alpha_n^{\star} \mathbf{x}_n^t \mathbf{x}_m$

Función discriminante lineal que maximiza el margen:

$$\phi_{\text{svm}}(\mathbf{x}; \Theta) = \sum_{n \in \mathcal{V}} \alpha_n^{\star} c_n \mathbf{x}_n^t \mathbf{x} + \theta_0^{\star}$$

DEMO

Ejercicios

1. Sea $S = \{((1, 1)^t, +1), ((2, 2)^t, -1)\}$ una muestra de entrenamiento. Mediante el método de los multiplicadores de Lagrange, obtener (analíticamente) θ^* y θ_0^* que clasifiquen S con el máximo margen.
2. Sea S una muestra linealmente separable. Demostrar que el margen óptimo es:

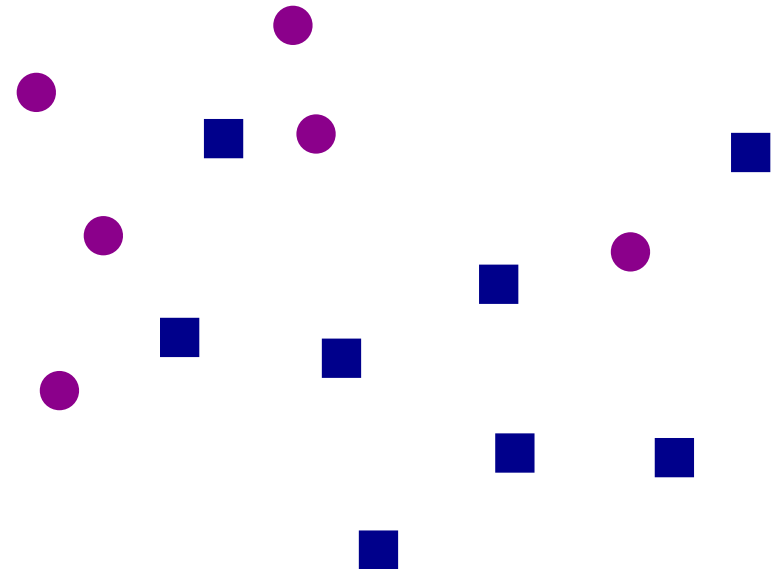
$$2 \left(\sum_{n \in \mathcal{V}} \alpha_n^* \right)^{-1/2}$$

Caso de no separabilidad lineal: “márgenes blandos”

A la función a minimizar, $\frac{1}{2}\|\boldsymbol{\theta}\|^2$, se le añade un término que pondera cómo de mal clasificado (o fuera del margen) se tolera que esté cada vector \boldsymbol{x}_n de S .

Dado $S = \{(\boldsymbol{x}_1, c_1), \dots, (\boldsymbol{x}_N, c_N)\}$ y una constante $\mathcal{C} > 0$, obtener $\boldsymbol{\theta} \in \mathbb{R}^d$, $\theta_0 \in \mathbb{R}$ y $\boldsymbol{\zeta} \in \mathbb{R}^N$ tales que:

- $\frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\theta} + \mathcal{C} \sum_{n=1}^N \zeta_n$ sea mínimo
- sujeto a:
 - $c_n (\boldsymbol{\theta}^t \boldsymbol{x}_n + \theta_0) \geq 1 - \zeta_n, \quad 1 \leq n \leq N$
 - $\zeta_n \geq 0, \quad 1 \leq n \leq N$

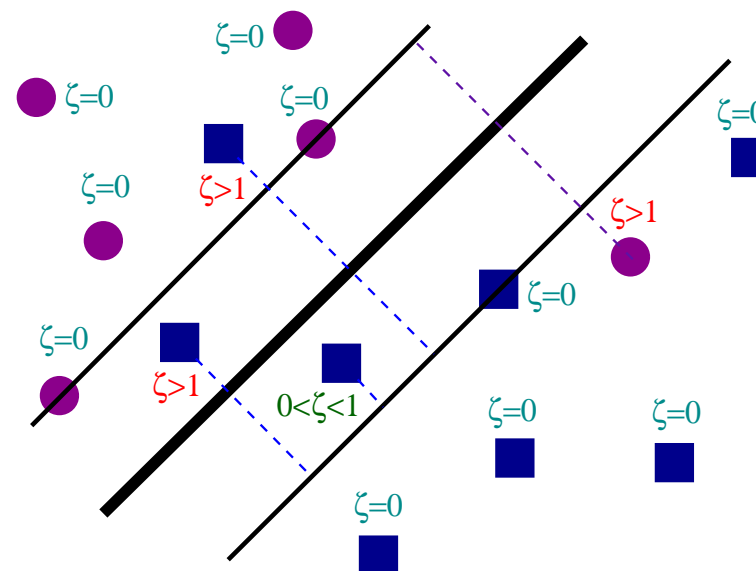


Caso de no separabilidad lineal: “márgenes blandos”

A la función a minimizar, $\frac{1}{2}\|\boldsymbol{\theta}\|^2$, se le añade un término que pondera cómo de mal clasificado (o fuera del margen) se tolera que esté cada vector \mathbf{x}_n de S .

Dado $S = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}$ y una constante $\mathcal{C} > 0$, obtener $\boldsymbol{\theta} \in \mathbb{R}^d$, $\theta_0 \in \mathbb{R}$ y $\zeta \in \mathbb{R}^N$ tales que:

- $\frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\theta} + \mathcal{C} \sum_{n=1}^N \zeta_n$ sea mínimo
- sujeto a:
 - $c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) \geq 1 - \zeta_n, \quad 1 \leq n \leq N$
 - $\zeta_n \geq 0, \quad 1 \leq n \leq N$



SVM en el caso de no separabilidad lineal

- *Lagrangiana primal:*

$$\text{Minimizar } \Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta} + C \sum_{n=1}^N \zeta_n - \sum_{n=1}^N \alpha_n (c_n (\boldsymbol{\theta}^t \mathbf{x}_n + \theta_0) + \zeta_n - 1) - \sum_{n=1}^N \beta_n \zeta_n$$

sujeito a $\alpha_n \geq 0$, $\beta_n \geq 0$ y $\zeta_n \geq 0$ para $1 \leq n \leq N$

- *Lagrangiana dual:*

$$\text{Maximizar } \Lambda_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \Lambda_D(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N c_n c_m \alpha_n \alpha_m \mathbf{x}_n^t \mathbf{x}_m$$

sujeito a $\alpha_n \geq 0$, $\alpha_n + \beta_n = C$ para $1 \leq n \leq N$ y a $\sum_{n=1}^N \alpha_n c_n = 0$

Desarrollo similar al caso separable (*ejercicio*)

SVM en el caso de no separabilidad lineal

- *Lagrangiana dual*:

$$\Lambda_D(\alpha) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n=1}^N \sum_{m=1}^N c_n c_m \alpha_n \alpha_m \mathbf{x}_n^t \mathbf{x}_m$$

- Las soluciones θ^* , θ_0^* , ζ^* , α^* , β^* verifican:

$$1. \theta^* = \sum_{n=1}^N c_n \alpha_n^* \mathbf{x}_n$$

$$2. \sum_{n=1}^N \alpha_n^* c_n = 0$$

$$3. 0 \leq \alpha_n^* \leq \mathcal{C} \quad 1 \leq n \leq N$$

$$4. \beta_n^* = \mathcal{C} - \alpha_n^* \quad 1 \leq n \leq N$$

- Condición complementaria de Karush-Kuhn-Tucker

$$\left. \begin{aligned} \alpha_n^* \left(c_n (\theta^{*t} \mathbf{x}_n + \theta_0^*) - 1 + \zeta_n^* \right) &= 0 \\ \beta_n^* \zeta_n^* &= 0 \end{aligned} \right\} 1 \leq n \leq N$$

Vectores soporte “erróneos”

$$1 \leq n \leq N \quad \left\{ \begin{array}{l} \alpha_n^* \left(c_n (\boldsymbol{\theta}^{*t} \mathbf{x}_n + \theta_0^*) - 1 + \zeta_n^* \right) = 0 \\ \beta_n^* \zeta_n^* = (\mathcal{C} - \alpha_n^*) \zeta_n^* = 0 \end{array} \right. \quad (1)$$

- (1) \rightarrow muestras \mathbf{x}_n tales que $\alpha_n^* \neq 0$ son **vectores soporte**.
En este caso, $c_n (\boldsymbol{\theta}^{*t} \mathbf{x}_n + \theta_0^*) = 1 - \zeta_n^*$
- (2) $\rightarrow (\mathcal{C} - \alpha_n^*) \zeta_n^* = 0$
 - $\mathcal{C} - \alpha_n^* > 0 \Rightarrow \zeta_n^* = 0 \Rightarrow c_n (\boldsymbol{\theta}^{*t} \mathbf{x}_n + \theta_0^*) = 1 \Rightarrow$ sin error de margen
 - $\zeta_n^* > 0 \Rightarrow \mathcal{C} = \alpha_n^* \Rightarrow$ **error de margen** $\left\{ \begin{array}{ll} \zeta_n^* > 1 & \text{error de clasificación} \\ \zeta_n^* \leq 1 & \text{dentro del margen} \end{array} \right.$
- \mathcal{C} se determina mediante validación cruzada y controla el compromiso entre el margen y los errores de margen
- Ejercicio: ¿Qué ocurre con el resto de muestras ($\alpha_n^* = 0$) ?

\mathcal{C} -SVM

- Calcular α_n^* , $1 \leq n \leq N$, que maximicen $\Lambda_D(\alpha)$,
sujeto a las restricciones: $\sum_{n=1}^N \alpha_n c_n = 0$ y $0 \leq \alpha_n \leq \mathcal{C}$, $1 \leq n \leq N$
- *Vectores soporte*: $\mathbf{x}_n \in S$, $n \in \mathcal{V}$, $\mathcal{V} = \{n \in \mathbb{N}, 1 \leq n \leq N \mid \alpha_n^* \neq 0\}$
- Coeficientes de la FDL:
 - $\boldsymbol{\theta}^* = \sum_{n \in \mathcal{V}} c_n \alpha_n^* \mathbf{x}_n$
 - $\theta_0^* = c_n - \boldsymbol{\theta}^{*t} \mathbf{x}_n$ para algún $n \in \mathcal{V}$ tal que $\alpha_n^* < \mathcal{C}$

Función discriminante lineal de margen máximo:

$$\phi_{\text{svm}}(\mathbf{x}; \boldsymbol{\Theta}) = \sum_{n \in \mathcal{V}} \alpha_n^* c_n \mathbf{x}_n^t \mathbf{x} + \theta_0^*$$

Métodos de optimización para SVM

Problema: *maximizar la Lagrangiana dual*:

$$\begin{aligned} \arg \max_{\boldsymbol{\alpha}} \quad & \Lambda_D(\boldsymbol{\alpha}) \\ \sum_{n=1}^N c_n \alpha_n = & 0 \\ 0 \leq \alpha_n \leq C, \quad & 1 \leq n \leq N \end{aligned}$$

- Solución analítica, si $N \lll$ (generalmente $N \leq 3$)
- Ascenso por gradiente, en general
- Algoritmos de descomposición, si $N \lesssim 5000$
- Optimización minimal secuencial, $N \gg 5000$
("Sequential minimal optimization algorithm", SMO)

Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 *Núcleos* ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 Aplicaciones ▷ 49
- 6 Notación ▷ 52

Funciones discriminantes lineales generalizadas (FDLG)

- FDLG para un problema de clasificación en dos clases:

$$\phi(\mathbf{x}; \Theta) = \sum_{i=1}^{d'} \theta_i \psi_i(\mathbf{x}) + \theta_0 = \boldsymbol{\theta}^t \boldsymbol{\psi}(\mathbf{x}) + \theta_0$$

donde: – $\mathbf{x} \in \mathbb{R}^d$, $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ (típicamente $d' \gg d$),
 – $\boldsymbol{\psi}$ es una función no lineal: $\boldsymbol{\psi} : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$

- Ejemplo: Linealización de funciones cuadráticas:

$$\phi(\mathbf{x}; \Theta) = \sum_{i=1}^d \sum_{j=1}^d a_{ij} x_i x_j + \sum_{j=1}^d b_j x_j + c$$

$$\phi(\mathbf{x}; \Theta) = \boldsymbol{\theta}^t \boldsymbol{\psi}(\mathbf{x}) + \theta_0 \quad \text{con } \boldsymbol{\psi} : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}, \quad d' = \frac{1}{2}d(d+3) :$$

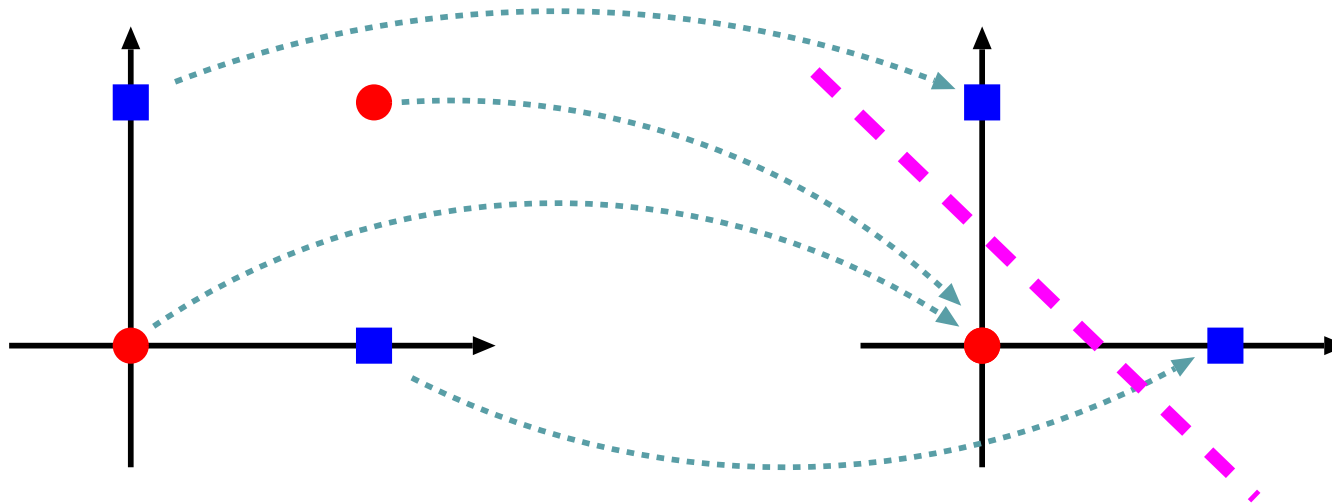
$$\boldsymbol{\psi}(x_1, \dots, x_d) = (x_1 x_1, x_1 x_2, \dots, x_2 x_1, x_2 x_2, \dots, x_d x_d, x_1, \dots, x_d)^t$$

$$\boldsymbol{\theta} = (a_{11}, a_{12}, \dots, a_{21}, a_{22}, \dots, a_{dd}, b_1, \dots, b_d)^t, \quad \theta_0 = c$$

Ejemplo: Problema de la 'O' exclusiva (XOR)

Mediante la función escalón $E : \mathbb{R} \rightarrow \{0, 1\}$ definida como $E(x) = \begin{cases} 1 & \text{si } x \geq 0 \\ 0 & \text{si } x < 0 \end{cases}$ el cambio de espacio de representación:

$$\psi_1(x_1, x_2) = E(x_1 - x_2 - 0.5) \quad \text{y} \quad \psi_2(x_1, x_2) = E(-x_1 + x_2 - 0.5)$$



permite definir una FDLG que linealiza el problema XOR:

$$\phi(x_1, x_2; \theta_0, \theta_1, \theta_2) = \sum_{k=1}^2 \theta_k \psi_k(x_1, x_2) + \theta_0$$

$$//\theta_1 = \theta_2 = 1, \theta_0 = -0.5// \quad = \quad E(x_1 - x_2 - 0.5) + E(-x_1 + x_2 - 0.5) - 0.5$$

Generalización de SVM¹: Núcleos

Se aprovecha la propiedad de que una SVM se puede expresar en base a productos escalares entre muestras de entrenamiento:

$$\begin{aligned}\phi(\mathbf{x}) &= \sum_{n=1}^N \alpha_n c_n \mathbf{x}_n^t \mathbf{x} + \theta_0 \\ &\Downarrow \\ \phi_\psi(\mathbf{x}) &= \sum_{n=1}^N \alpha_n c_n \psi(\mathbf{x}_n)^t \psi(\mathbf{x}) + \theta_0 \\ &\Downarrow \\ \phi_{\mathcal{K}}(\mathbf{x}) &= \sum_{n=1}^N \alpha_n c_n \mathcal{K}(\mathbf{x}_n, \mathbf{x}) + \theta_0\end{aligned}$$

donde $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ es una función que se denomina **núcleo** si $\exists \psi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ tal que $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^t \psi(\mathbf{x}')$, $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$.

A $\mathbb{R}^{d'}$ se le suele llamar **espacio de características**

¹ Una generalización similar puede hacerse también para el perceptrón (*Kernel perceptron*)

Núcleos: ejemplo

Sean $\mathbf{x} = (x_1, x_2, x_3)^t$, $\mathbf{y} = (y_1, y_2, y_3)^t$ y $\mathcal{K}: \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ definida como:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = (x_1 y_1 + x_2 y_2 + x_3 y_3)^2$$

¿Es $\mathcal{K}(\mathbf{x}, \mathbf{y})$ un núcleo?

Núcleos: ejemplo

Sean $\mathbf{x} = (x_1, x_2, x_3)^t$, $\mathbf{y} = (y_1, y_2, y_3)^t$ y $\mathcal{K}: \mathbb{R}^3 \times \mathbb{R}^3 \rightarrow \mathbb{R}$ definida como:

$$\mathcal{K}(\mathbf{x}, \mathbf{y}) = (x_1 y_1 + x_2 y_2 + x_3 y_3)^2$$

¿Es $\mathcal{K}(\mathbf{x}, \mathbf{y})$ un núcleo? ... Sí, ya que:

$$\begin{aligned} \mathcal{K}(\mathbf{x}, \mathbf{y}) &= (x_1 y_1 + x_2 y_2 + x_3 y_3)^2 \\ &= x_1^2 y_1^2 + x_2^2 y_2^2 + x_3^2 y_3^2 + 2 x_1 y_1 x_2 y_2 + 2 x_1 y_1 x_3 y_3 + 2 x_2 y_2 x_3 y_3 \end{aligned}$$

$\mathcal{K}(\mathbf{x}, \mathbf{y}) = \psi(\mathbf{x})^t \psi(\mathbf{y})$ si $\psi: \mathbb{R}^3 \rightarrow \mathbb{R}^6$ se define como:

$$\psi(\mathbf{x}) = (x_1^2, x_2^2, x_3^2, \sqrt{2} x_1 x_2, \sqrt{2} x_1 x_3, \sqrt{2} x_2 x_3)^t$$

$$\psi(\mathbf{y}) = (y_1^2, y_2^2, y_3^2, \sqrt{2} y_1 y_2, \sqrt{2} y_1 y_3, \sqrt{2} y_2 y_3)^t$$

Alternativas para calcular $\mathcal{K}(\mathbf{x}, \mathbf{y})$:

- *Directamente* en \mathbb{R}^3 , mediante $\mathcal{K}(\mathbf{x}, \mathbf{y})$: $3 + 2 + 1 = 6$ productos + sumas
- Obtener primero $\psi(\mathbf{x})$, $\psi(\mathbf{y})$ en \mathbb{R}^6 y calcular $\psi(\mathbf{x})^t \psi(\mathbf{y})$:
 $2 \cdot 6 + 6 + 5 = 23$ productos + sumas

Construcción de núcleos

- Elegir $\psi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ y el núcleo es $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^t \psi(\mathbf{x}')$
Es necesario trabajar en $\mathbb{R}^{d'}$, $d' \gg d$: ¡amenaza de la dimensionalidad!
- Elegir $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ y:
 - Demostrar que $\exists \psi : \mathbb{R}^d \rightarrow \mathbb{R}^{d'} : \mathcal{K}(\mathbf{x}, \mathbf{x}') = \psi(\mathbf{x})^t \psi(\mathbf{x}')$, $\forall \mathbf{x}, \mathbf{x}' \in \mathbb{R}^d$
 - *Condición de Mercer*: \mathcal{K} es un núcleo si y solo si, para cualquier conjunto de vectores $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$, la matriz $[\mathcal{K}(\mathbf{x}_n, \mathbf{x}_m)]_{1 \leq n, m \leq N}$ (llamada matriz de Gramm) es semidefinida positiva
 - Mediante “álgebra de núcleos”: construir \mathcal{K} a partir de núcleos simples. Si \mathcal{K}_1 y \mathcal{K}_2 son núcleos, entonces también son núcleos:
 - * La suma, el producto o cualquier polinomio con coeficientes no negativos de \mathcal{K}_1 y \mathcal{K}_2
 - * $\exp(\mathcal{K}_1(\mathbf{x}, \mathbf{x}'))$
 - *Núcleos de base radial* (RBK): $\mathcal{K}(\mathbf{x}, \mathbf{x}') \stackrel{\text{def}}{=} f(r)$, $r = \|\mathbf{x} - \mathbf{x}'\|$
Ejemplo: núcleos gaussianos: $\mathcal{K}(\mathbf{x}, \mathbf{x}') = \exp(-c \|\mathbf{x} - \mathbf{x}'\|^2)$
ver: http://en.wikipedia.org/wiki/Radial_basis_function
 - ... Etc.

Máquinas de vectores soporte y núcleos

- **Problema:** Dado un muestra de entrenamiento $S = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}$, una constante $\mathcal{C} \geq 0$ y un núcleo $\mathcal{K}(\mathbf{x}_n, \mathbf{x}_m) = \psi(\mathbf{x}_n)^t \psi(\mathbf{x}_m)$, obtener $\boldsymbol{\theta} \in \mathbb{R}^{d'}$ y $\theta_0 \in \mathbb{R}$ tales que:

- $\frac{1}{2} \boldsymbol{\theta}^t \boldsymbol{\theta} + \mathcal{C} \sum_{n=1}^N \zeta_n$ sea mínimo
- sujeto a $c_n(\boldsymbol{\theta}^t \psi(\mathbf{x}_n) + \theta_0) \geq 1 - \zeta_n$ y $\zeta_n \geq 0, 1 \leq n \leq N$

- **Solución:** Lagrangiana dual: $\Lambda_D(\boldsymbol{\alpha}) = \sum_{n=1}^N \alpha_n - \frac{1}{2} \sum_{n,m=1}^N c_n c_m \alpha_n \alpha_m \mathcal{K}(\mathbf{x}_n, \mathbf{x}_m)$

– Calcular $\boldsymbol{\alpha}^*$ que maximice $\Lambda_D(\boldsymbol{\alpha})$

sujeto a $\sum_{n=1}^N \alpha_n c_n = 0, 0 \leq \alpha_n \leq \mathcal{C}, 1 \leq n \leq N$

– **Vectores soporte:** $\mathbf{x}_n \in S : n \in \mathcal{V}, \mathcal{V} = \{n \in \mathbb{N}, 1 \leq n \leq N \mid \alpha_n^* \neq 0\}$

– **FDLG:** $\phi_{\mathcal{K}}(\mathbf{x}) = \sum_{n \in \mathcal{V}} c_n \alpha_n^* \mathcal{K}(\mathbf{x}_n, \mathbf{x}) + \theta_0^*$

con $\theta_0^* = c_n - \sum_{m \in \mathcal{V}} c_m \alpha_m^* \mathcal{K}(\mathbf{x}_m, \mathbf{x}_n)$ para algún $n \in \mathcal{V} : \alpha_n^* < \mathcal{C}$

Cuestión: ¿Dónde están ψ y $\boldsymbol{\theta}^*$?

Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 Núcleos ▷ 23
- 4 *SVM para problemas de C clases* ▷ 31
- 5 Aplicaciones ▷ 49
- 6 Notación ▷ 52

Clasificación en C clases con funciones discriminantes lineales

FUNCIONES DISCRIMINANTES LINEALES GENERALIZADAS

$$\phi_c(\mathbf{x}; \Theta) = \boldsymbol{\theta}_c^t \boldsymbol{\psi}(\mathbf{x}) + \theta_{c0} = \sum_{i=1}^{d'} \theta_{ci} \psi_i(\mathbf{x}) + \theta_{c0}, \quad 1 \leq c \leq C$$

donde:

- $\boldsymbol{\psi} : \mathbb{R}^d \rightarrow \mathbb{R}^{d'}$ es una *función de cambio de espacio*
- $\boldsymbol{\theta}_c \in \mathbb{R}^{d'}$ es el **vector de pesos** de la clase c
- $\theta_{c0} \in \mathbb{R}$ es el **umbral** de la clase c

REGLA DE CLASIFICACIÓN

$$f(\mathbf{x}) \stackrel{\text{def}}{=} \hat{c} = \arg \max_{1 \leq c \leq C} \phi_c(\mathbf{x}; \Theta) \iff \phi_{\hat{c}}(\mathbf{x}; \Theta) > \phi_c(\mathbf{x}; \Theta) \quad \forall c \neq \hat{c}$$

El problema de C clases: uno-contr-uno

$$S = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}, \text{ con } \mathbf{x}_n \in \mathbb{R}^d, c_n \in \{1, \dots, C\}$$

$$\text{Un núcleo } \mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$$

- $C(C-1)/2$ clasificadores uno-contr-uno

– *Aprendizaje:* Para $1 \leq c < c' \leq C$,

$$S_{cc'} : (\mathbf{x}_n, c_{ncc'}) \in S_{cc'} \text{ si } (\mathbf{x}_n, c_n) \in S \text{ con } c_n = c \text{ o } c_n = c' \text{ y } c_{ncc'} = \begin{cases} +1 & c_n = c \\ -1 & c_n = c' \end{cases}$$

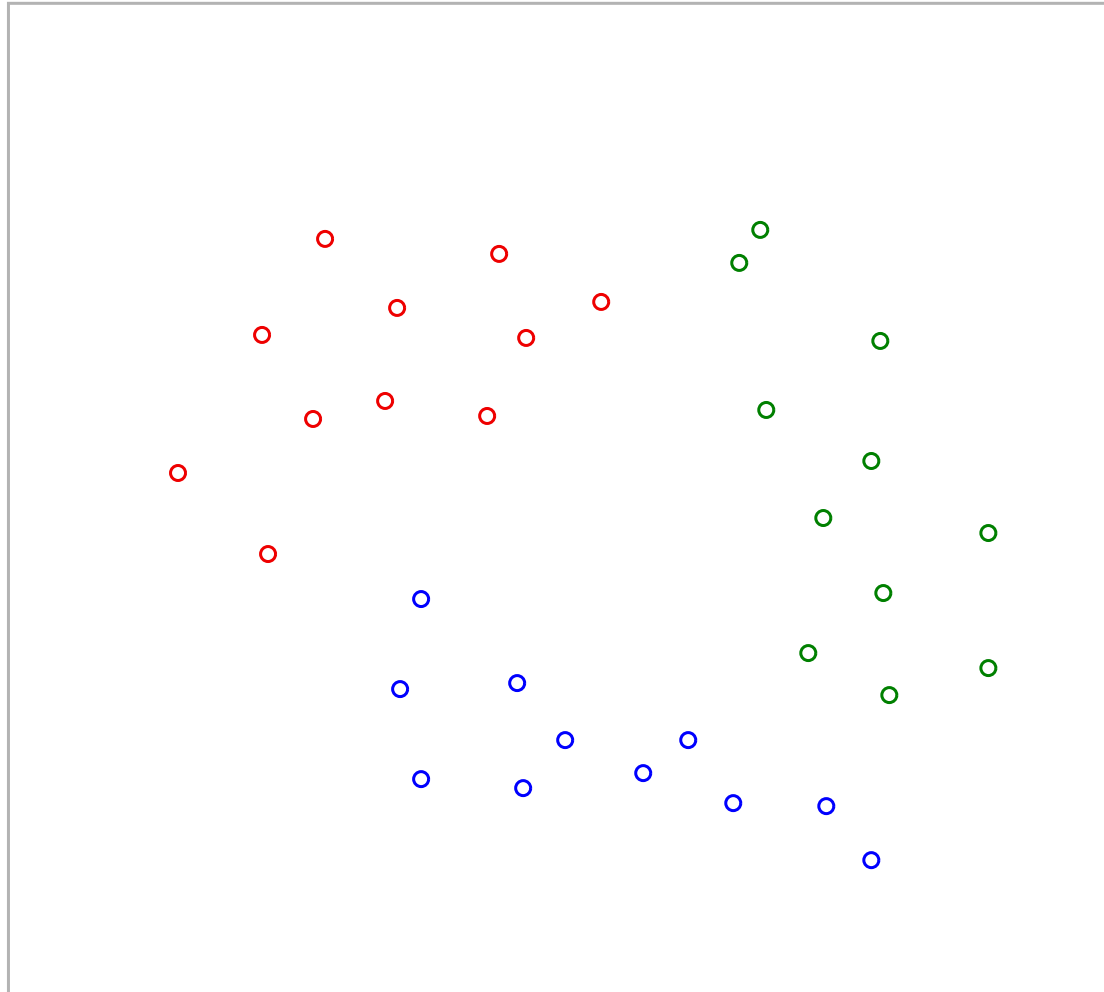
$$\phi_{cc'}(\mathbf{x}) = \sum_{\mathbf{x}_n \in S_{cc'}} \alpha_{ncc'}^* c_{ncc'} \mathcal{K}(\mathbf{x}_n, \mathbf{x}) + \theta_{cc'0}^*$$

$$f_{cc'}(\mathbf{x}) = \begin{cases} +1 & \text{si } \phi_{cc'}(\mathbf{x}) \geq 0 \\ -1 & \text{si } \phi_{cc'}(\mathbf{x}) < 0 \end{cases}$$

- *Clasificación por votación:* $O(C^2)$ (or $O(M^2 \cdot |\bar{\mathcal{V}}|)$ cálculos de kernels)
- *Clasificación utilizando DAGs (directed acyclic graphs):* $O(C)$

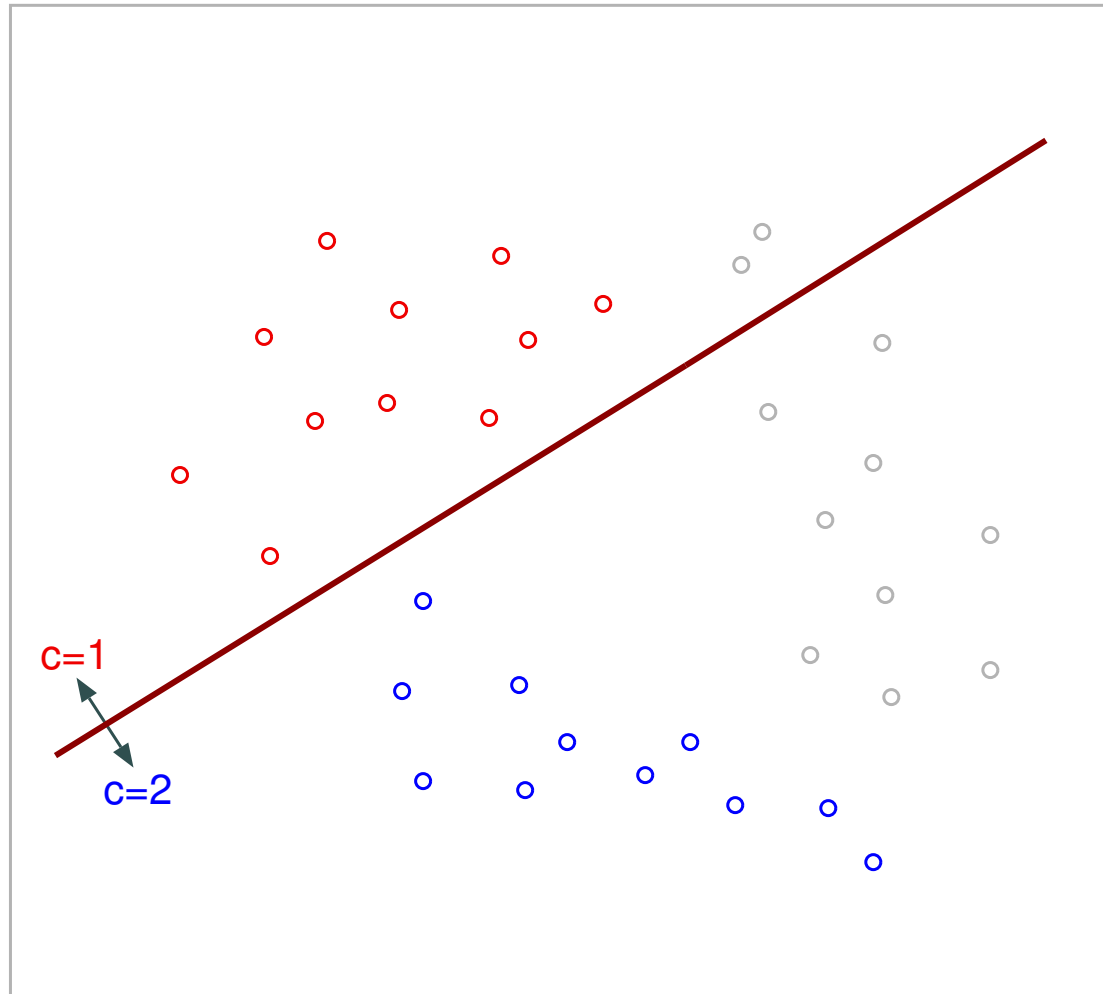
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



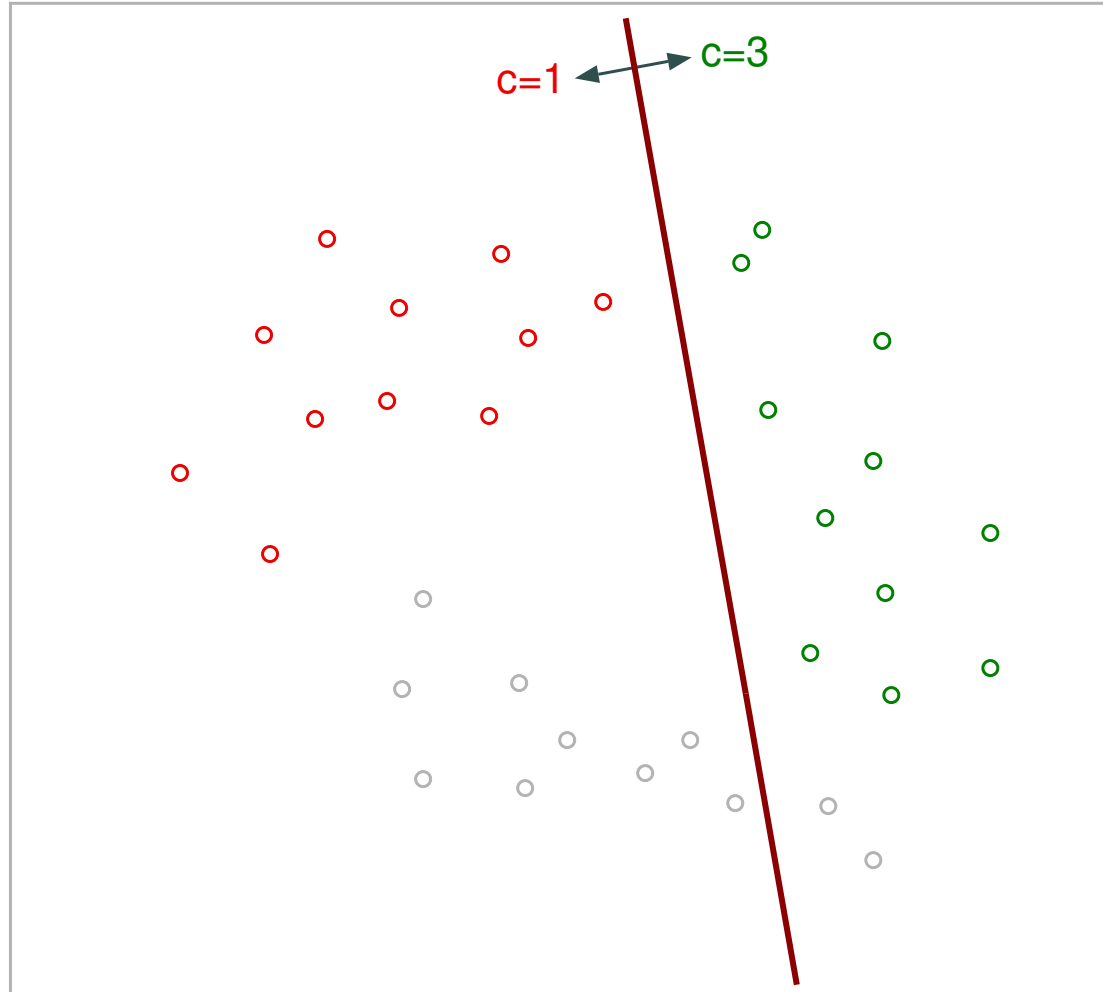
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



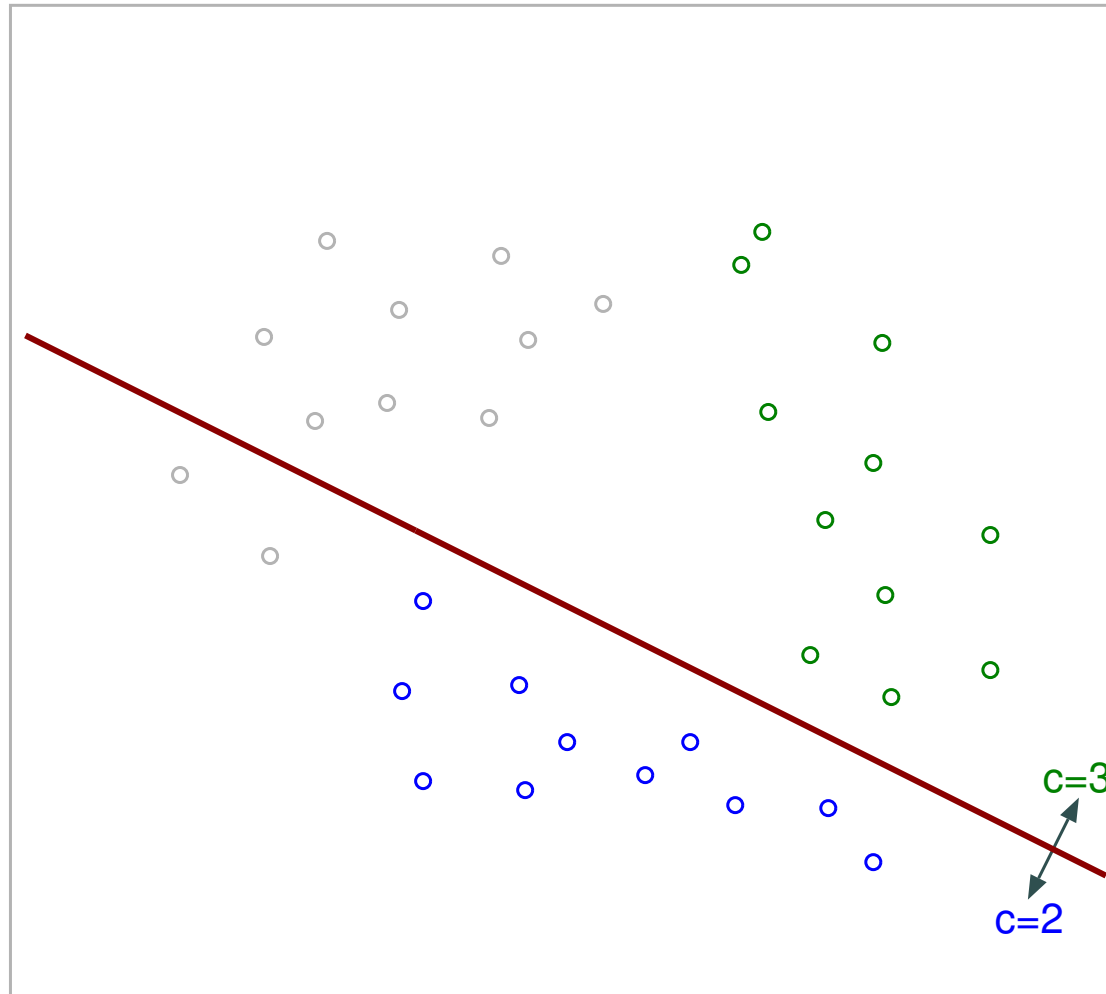
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



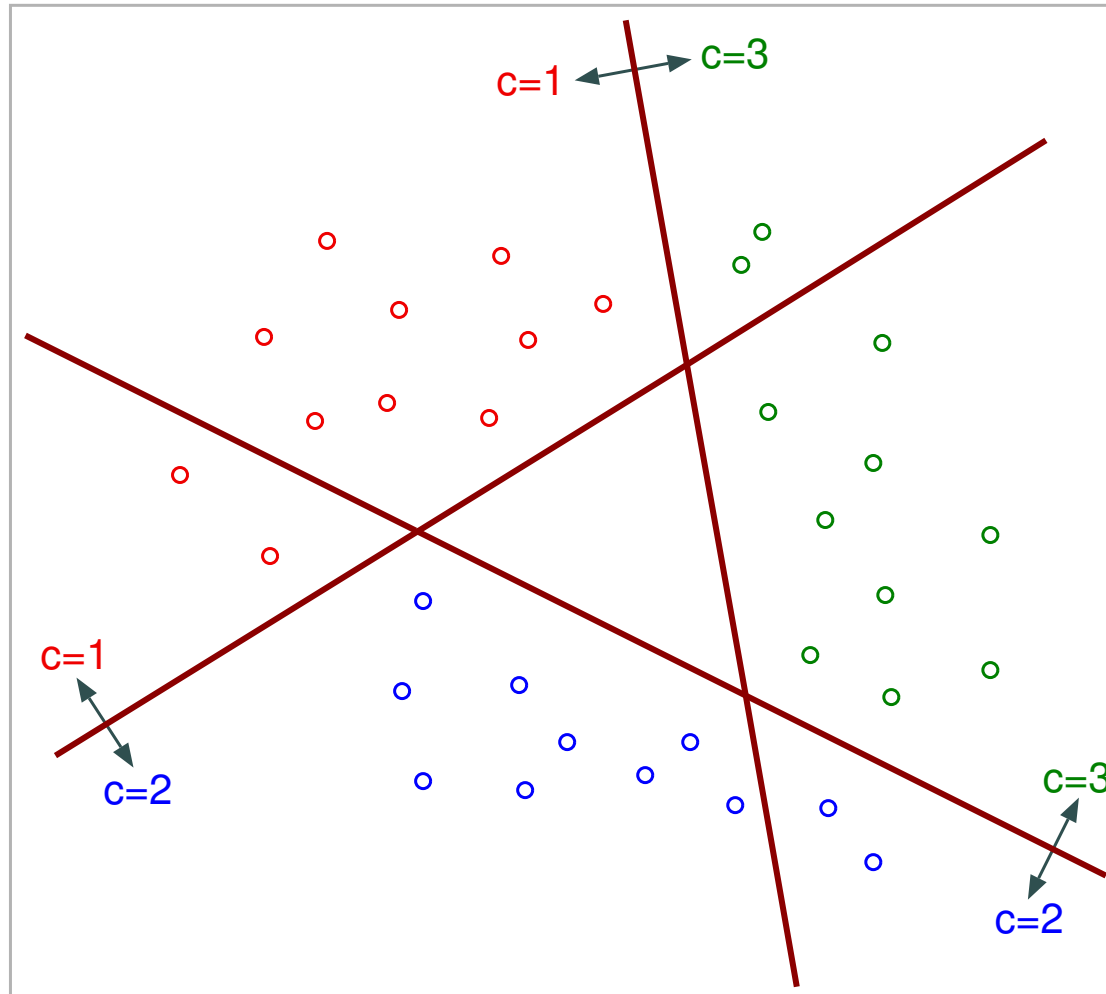
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



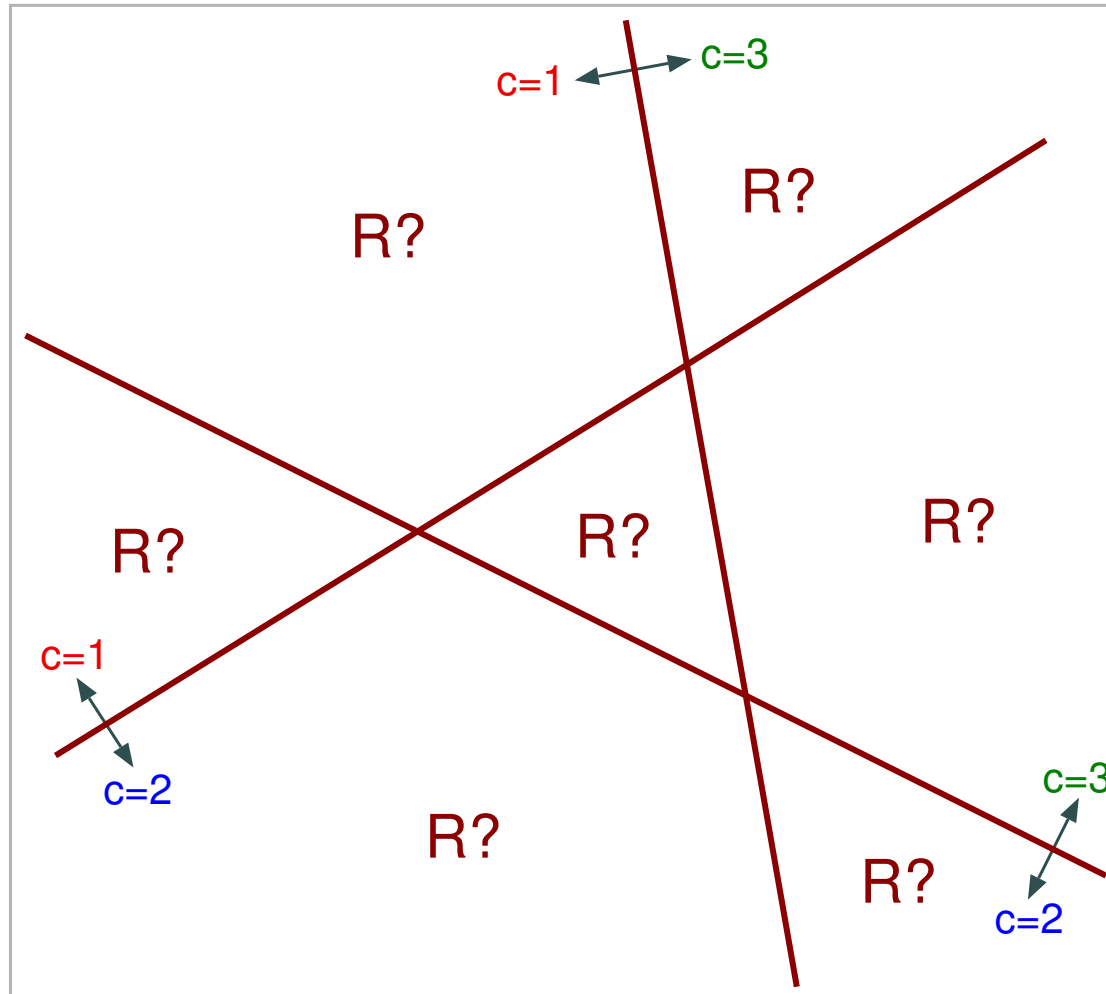
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



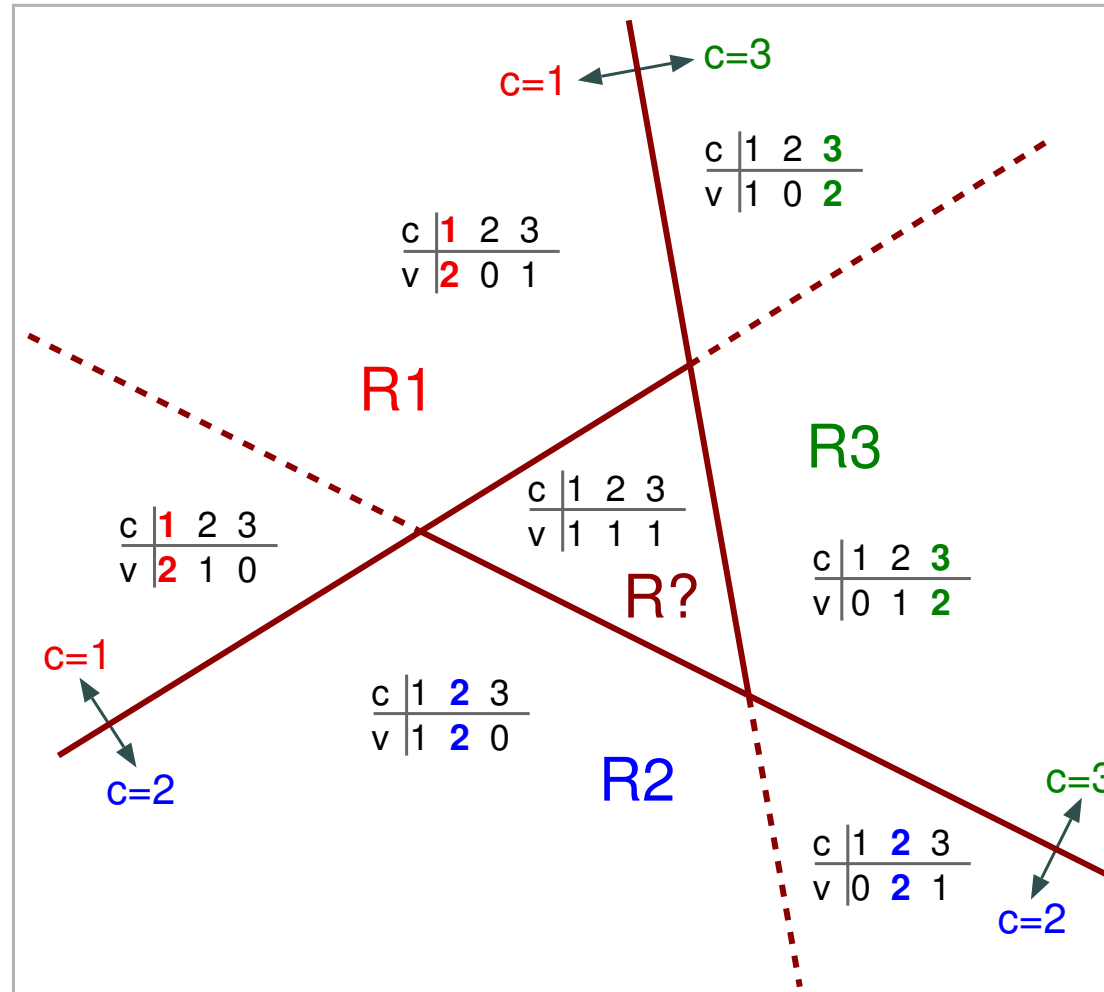
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación



Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-uno por votación

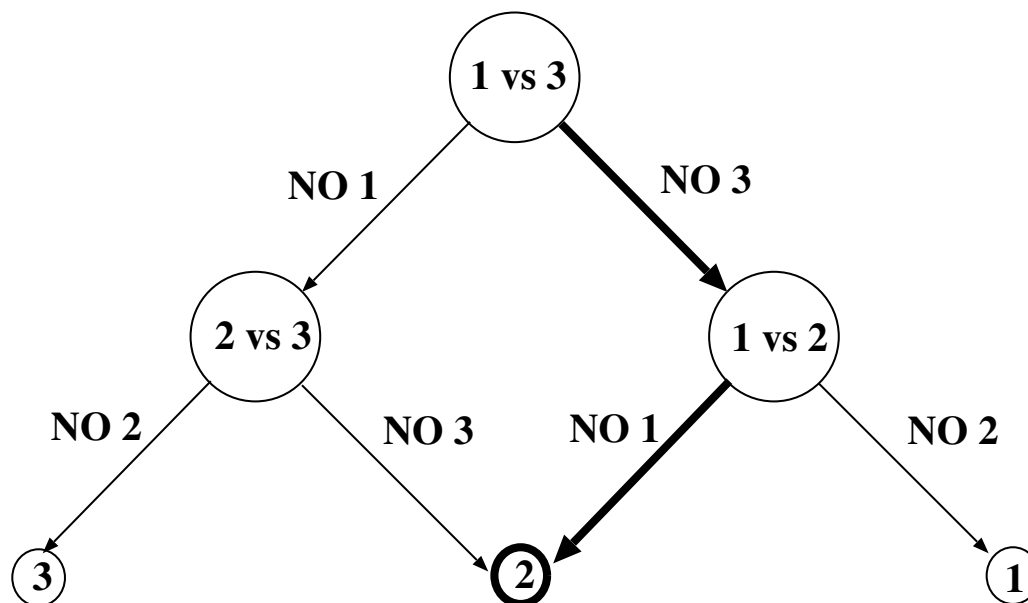


El problema de 3 clases: uno-contra-uno y DAGs

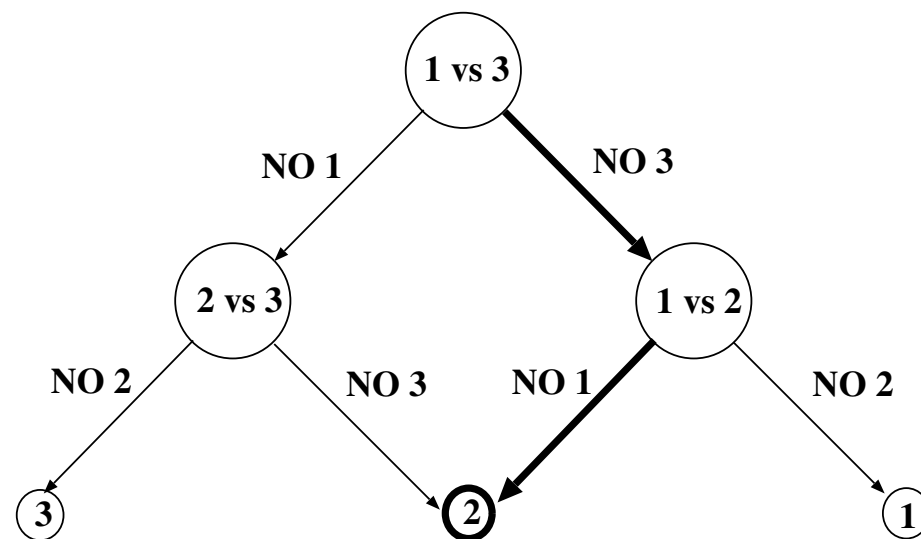
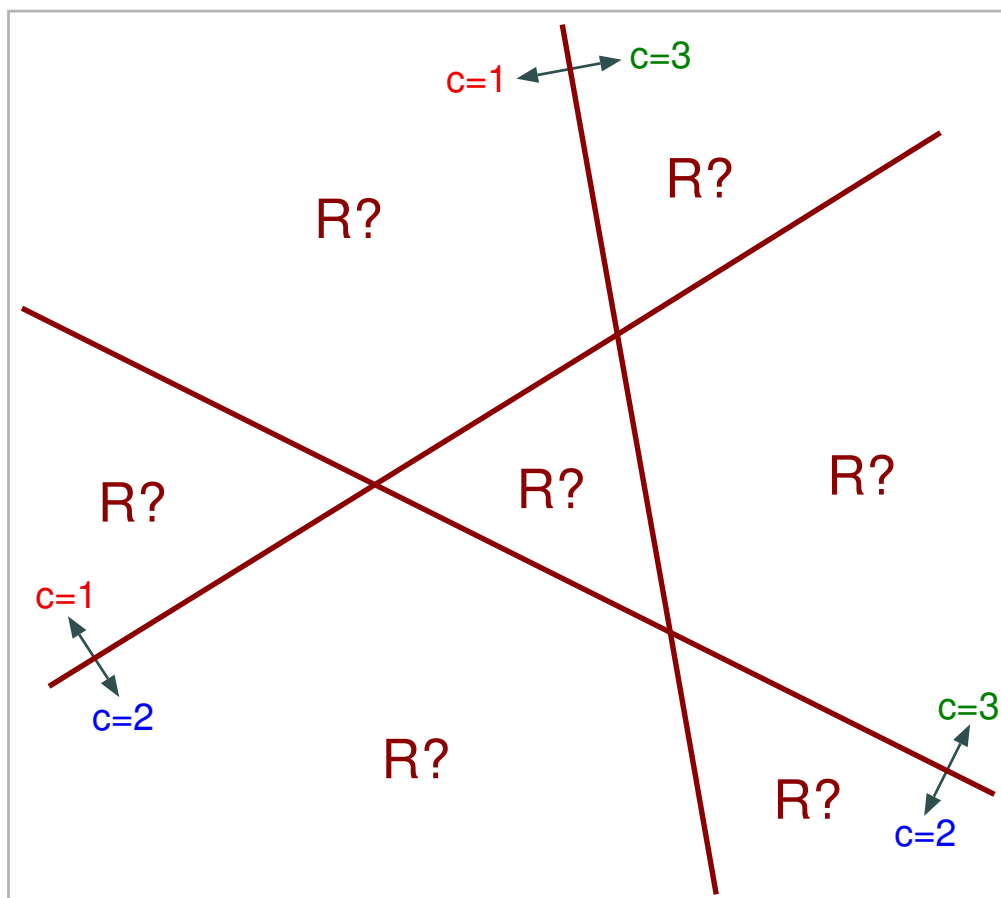
- Formulación equivalente a la clasificación por votación: A partir de los clasificadores binarios $f_{cc'}$ para $1 \leq c, c' \leq C$: La decisión multi-clase se define:

$$f(\mathbf{x}) = \arg \max_{1 \leq c \leq C} \sum_{c' \neq c} f_{cc'}(\mathbf{x})$$

- Clasificación utilizando grafos dirigidos y acíclicos* (DAGs): Para $C = 3$,



Clasificación multi-clase mediante clasificadores binarios: ejemplo anterior mediante DAGs

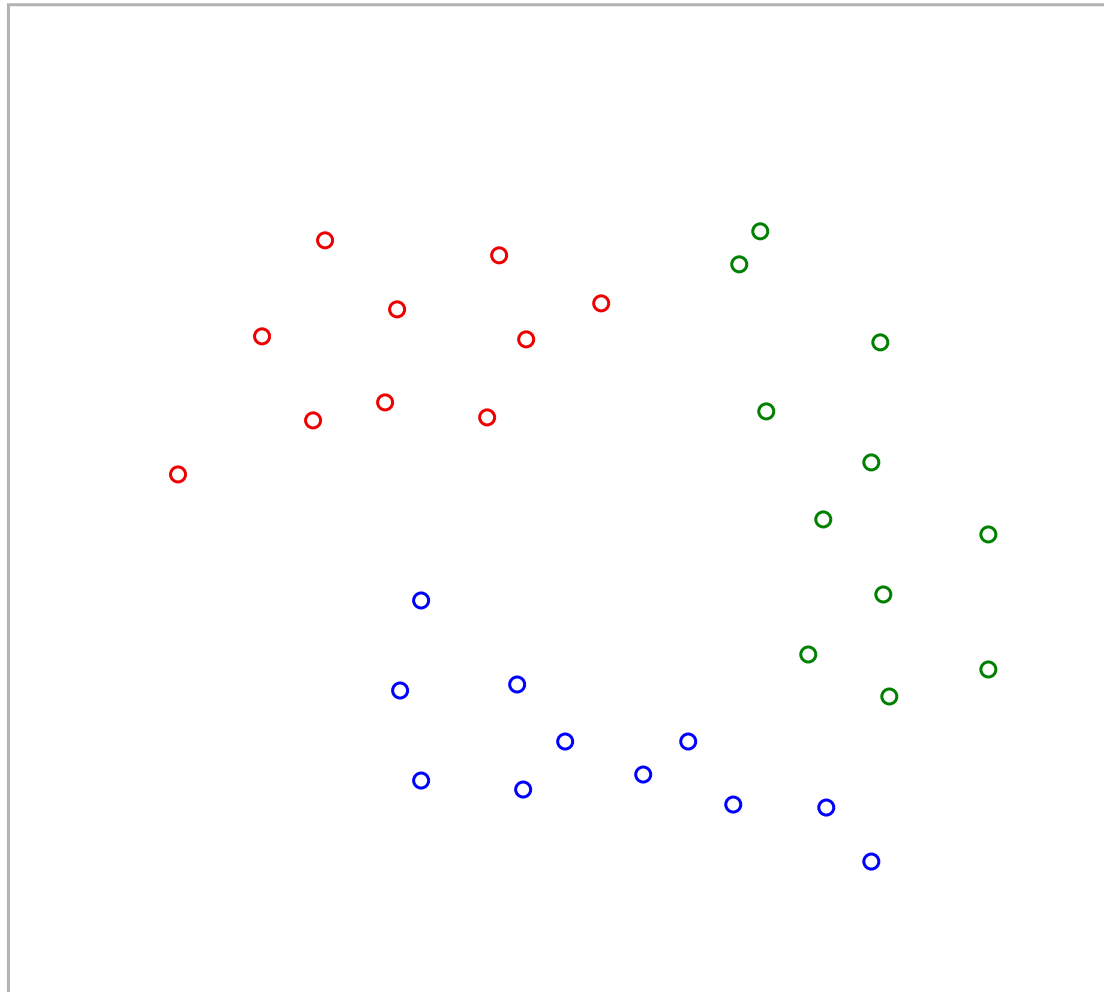


Otras técnicas de clasificación multi-clase para SVM

- *Uno-contra-el-resto*: Se entrenan C FDLs binarias, ϕ_c , $1 \leq c \leq C$, usando como muestras positivas *solo* los vectores de la clase c y como negativas el resto.
- $SVM^{\text{multi}C}$: Optimización directa de márgenes en C clases [Cramer & Singer, 01]
- *Construcción de Kesler*: Transformación de un problema de C clases en otro de 2 clases (aumentando la dimensionalidad). [Duda & Hart, 73], [Franc & Hlaváč, 02]
- *En la práctica*: Las técnicas que parecen más adecuadas son *uno-contra-uno* con *DAG*. [Hsu & Lin, 03]
- *Demo*: <http://www.csie.ntu.edu.tw/~cjlin/libsvm>

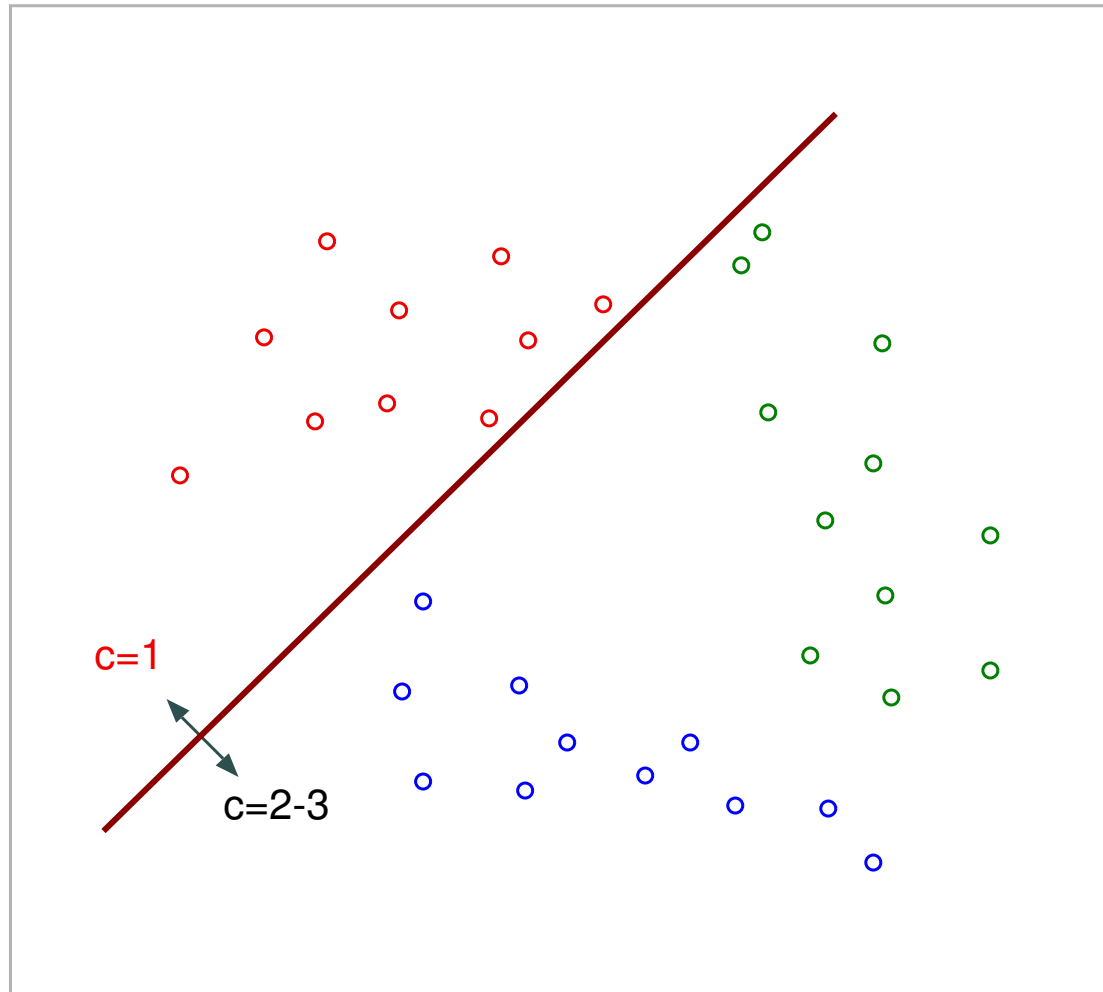
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-resto



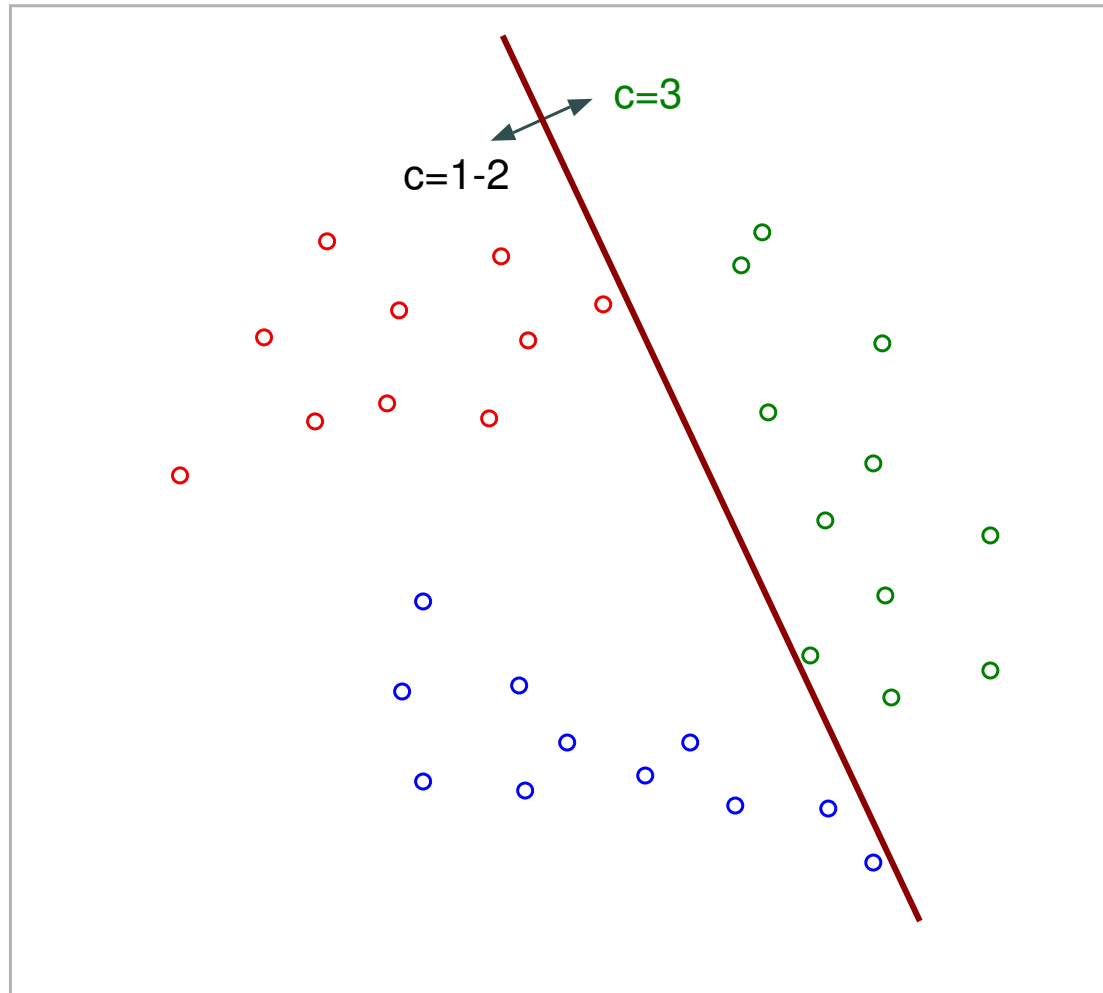
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-resto



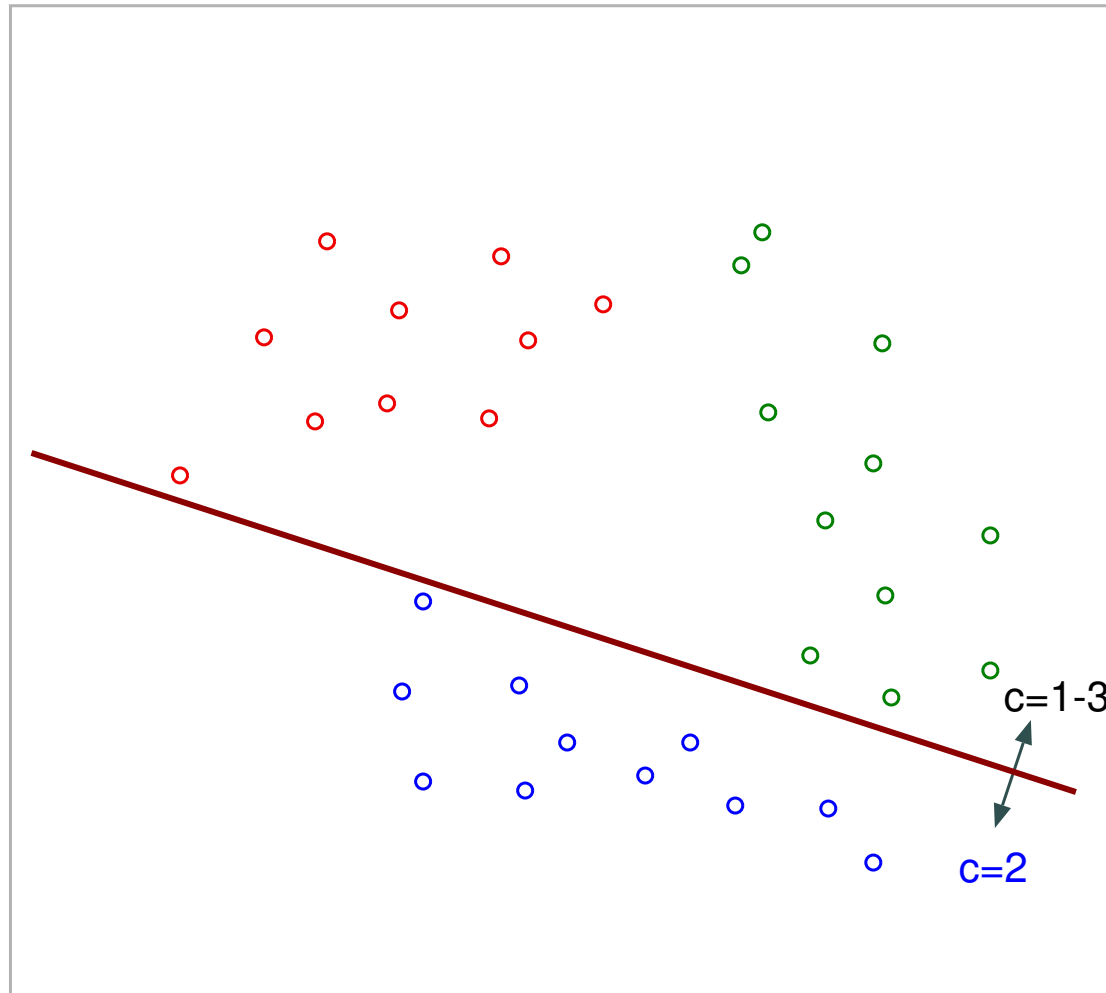
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-resto



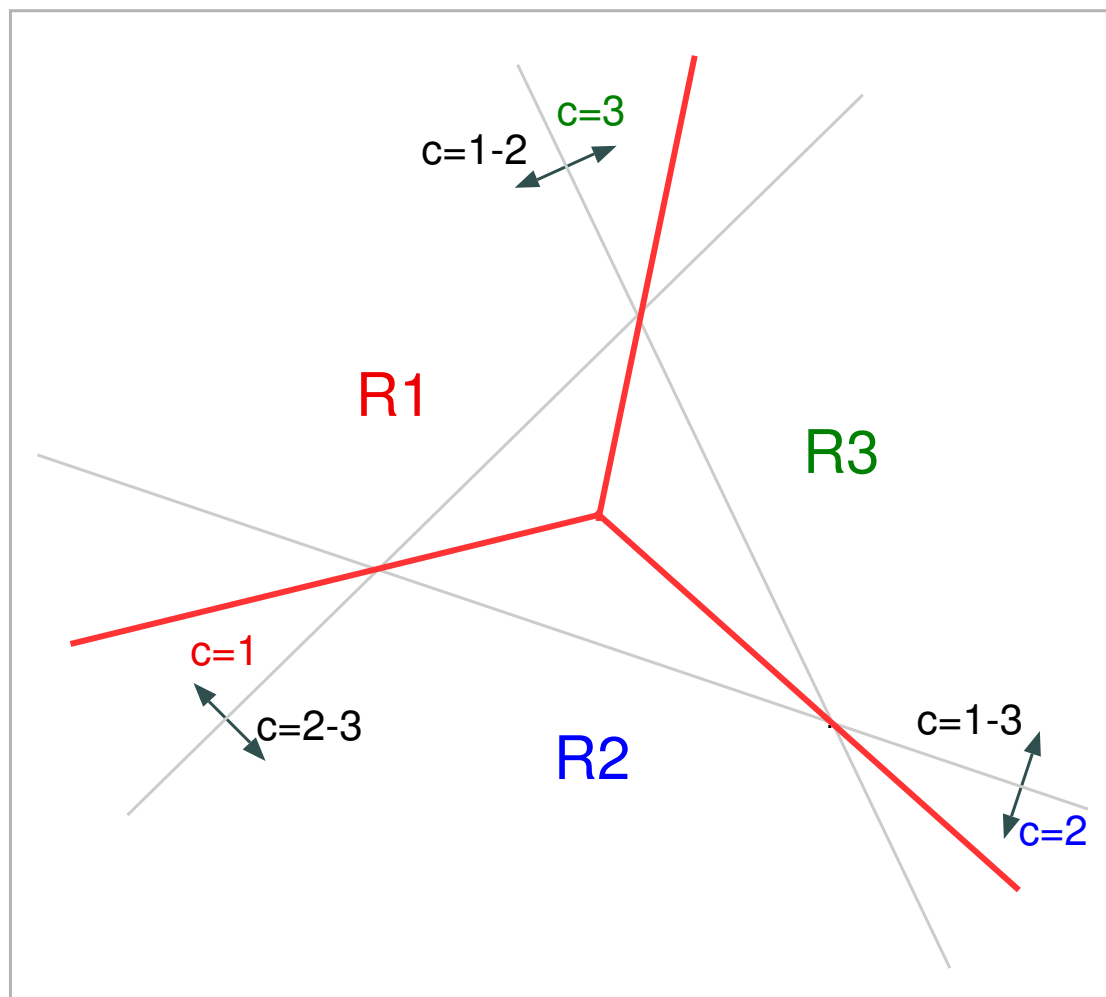
Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-resto



Clasificación multi-clase mediante clasificadores binarios: ejemplo

Uno-contra-resto



Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 Núcleos ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 *Aplicaciones* ▷ 49
- 6 Notación ▷ 52

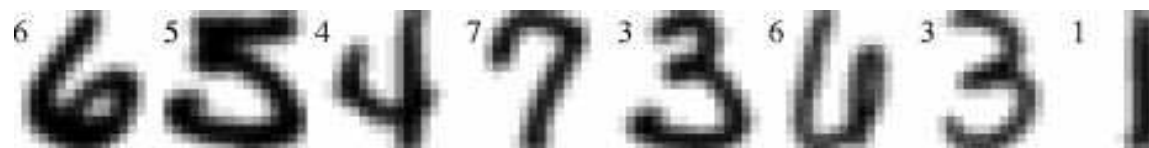
Aplicaciones: reconocimiento de caracteres manuscritos off-line

K.-R. Müller et al: An Introduction to Kernel-Based Learning Algorithms. IEEE Trans. on Neural networks. 2001¹

A.M. HafizK & G.M. Baht: Handwritten Digit Recognition using Slope Detail Features. International Journal of Computer Applications. 2014²

Q. Wang et al. Convolutional 2D LDA for Nonlinear Dimensionality Reduction. Joint Conference on Artificial Intelligence. 2017.³

Corpus USPS: 7291 muestras



Técnica	Tasa de error (%)
SVM sin kernel	8.7 ¹
k-vecino más próximo	5.7 ¹
Redes neuronales radiales	4.1 ¹
SVM virtuales	3.0 ¹
Vecino más próximo utilizando la distancia tangente	2.5 ¹
Humano	2.5 ¹
SVM (características especiales)	1.3 ²
Redes convolucionales	2.1 ³

Aplicaciones diversas

- Búsqueda de imágenes
- Detección de caras
- Localización de las matrículas
- Detección de texto en imágenes
- Detección de humanos en movimiento
- Detección de mensajes ocultos en imágenes
- Clasificación de texto
- Predicción del nivel del agua de un lago
- Series temporales financieras
- Reconocimiento de secuencias en texto genómico

Index

- 1 Funciones discriminantes lineales ▷ 2
- 2 Clasificadores de margen máximo: SVM ▷ 7
- 3 Núcleos ▷ 23
- 4 SVM para problemas de C clases ▷ 31
- 5 Aplicaciones ▷ 49
- 6 *Notación* ▷ 52

Notación

- Representación de un objeto y su clase: $\mathbf{x} \in \mathbb{R}$ y $c \in \{+1, -1\}$ (para problemas de dos clases).
- **Funciones discriminantes lineales**: $\phi(\mathbf{x}; \Theta)$ para una entrada \mathbf{x} y parámetros Θ compuestos por vector de pesos y umbral (θ, θ_0) .
- Conjunto de N muestras de entrenamiento:
 $S = \{(\mathbf{x}_1, c_1), \dots, (\mathbf{x}_N, c_N)\}$
- Función de Lagrange: $\Lambda(\theta, \theta_0, \alpha)$ con multiplicadores de Lagrange α_n
- Lagrangiana dual: $\Lambda_D(\alpha)$
- Conjunto de vectores soporte: \mathcal{V}
- Coeficientes de tolerancia para “márgenes blandos”: ζ_n
- Núcleo: \mathcal{K}