2016-2017

Aprendizaje Automático

4. Máquinas de vectores soporte

Escola Tècnica Superior d'Enginyeria Informàtica

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

Enrique Vidal Ruiz
(evidal@dsic.upv.es)

Departament de Sistemas Informàtics i Computació (DSIC)

Universitat Politècnica de València (UPV)

Septiembre, 2016

Index

Clasificación en dos clases con funciones discriminantes lineales

FUNCIONES DISCRIMINANTES LINEALES (FDL)

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

$\boldsymbol{\Theta} = (\boldsymbol{\theta}, \theta_0)$: $\boldsymbol{\theta} \in \mathbb{R}^d$ es un *vector de pesos* y $\theta_0 \in \mathbb{R}$ se denomina *umbral* ($D = d + 1$).

REGLA DE CLASIFICACIÓN

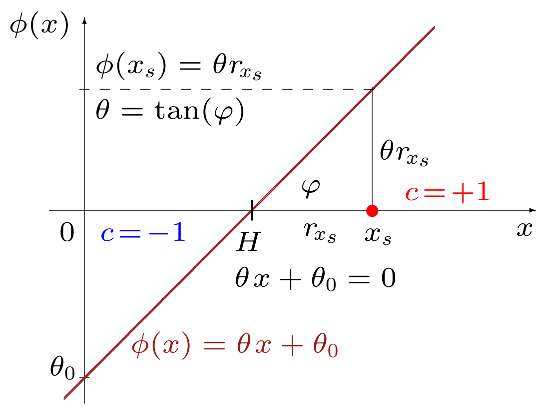
Asumiendo que las etiquetas de clase, c, son $+1$ y -1:

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

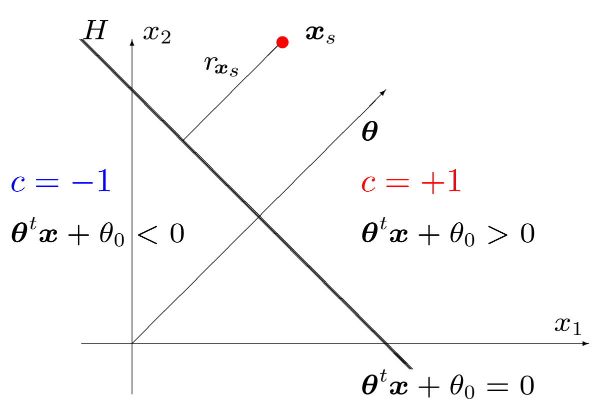
Propiedades de las funciones discriminantes lineales

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

Ejemplo con $d = 1$



Ejemplo con $d = 2$



Aprendizaje de funciones discriminantes lineales

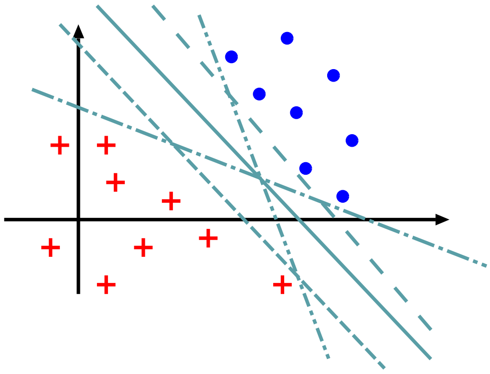
$S = \{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}$, with $\boldsymbol{x}_n \in \mathbb{R}^d$, $c_n \in \{+1, -1\},\ 1 \leq n \leq N$.

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

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

Aprendizaje: Dada una muestra linealmente separable S, encontrar $\boldsymbol{\Theta} = (\boldsymbol{\theta}, \theta_0)$ 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* $b \in \mathbb{R}^{\geq 0}$: $c_n \left(\boldsymbol{\theta}^t \boldsymbol{x}_n + \theta_0\right) \geq b$

Index

Forma canónica y margen de un clasificador

Para un hiperplano separador H dado, hay múltiples posibilidades de definirlo mediante diferentes FDLs $\phi(\boldsymbol{x}; \boldsymbol{\Theta})$. La *FDL canónica* con respecto a un conjunto S de N puntos se define por $\check{\boldsymbol{\Theta}} \equiv (\check{\boldsymbol{\theta}}, \check{\theta}_0)$, tal que:

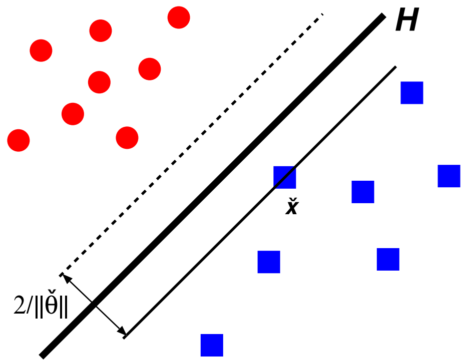
$$\min_{1 \leq n \leq N} c_n \phi(\boldsymbol{x}_n; \check{\boldsymbol{\Theta}}) \;\equiv\; \min_{1 \leq n \leq N} |\check{\boldsymbol{\theta}}^t \boldsymbol{x}_n + \check{\theta}_0 \,| \;=\; 1$$

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

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

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

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



Pregunta: En el ejemplo de la figura ¿puede haber otro H' dado por $(\boldsymbol{\theta}', \theta_0')$ tal que $\frac{2}{\|\boldsymbol{\theta}'\|} > \frac{2}{\|\check{\boldsymbol{\theta}}\|}$?

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

Clasificadores de margen máximo

- *Aprendizaje*: dada una muestra linealmente separable S, encontrar $\boldsymbol{\theta} \in \mathbb{R}^d$ y $\theta_0 \in \mathbb{R}$ que:
  - *maximicen:* $\dfrac{2}{\|\boldsymbol{\theta}\|}$
  - *sujetas a:* $c_n \;(\boldsymbol{\theta}^t \boldsymbol{x}_n + \theta_0) \;\geq\; 1, \quad 1 \leq n \leq N$
- Equivalentemente, *buscar* $\boldsymbol{\theta} \in \mathbb{R}^d$ *y* $\theta_0 \in \mathbb{R}$ *que:*
  - *minimicen:* $\dfrac{1}{2}\, \boldsymbol{\theta}^t \boldsymbol{\theta}$
  - *sujetas a:* $c_n \;(\boldsymbol{\theta}^t \boldsymbol{x}_n + \theta_0) \;\geq\; 1, \quad 1 \leq n \leq N$

Aplicación de la técnica de los multiplicadores de Lagrange

- Función de Lagrange:

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

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

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

$$\boldsymbol{\nabla}_{\boldsymbol{\theta}}\Lambda = \mathbf{0} \;\Rightarrow\; \boldsymbol{\theta}^\star = \sum_{n=1}^{N} c_n\,\alpha_n\,\boldsymbol{x}_n\,; \qquad \frac{\partial\Lambda}{\partial\theta_0} = 0 \;\Rightarrow\; \sum_{n=1}^{N}\alpha_n\,c_n \;=\; 0$$

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

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

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

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: $c_n(\boldsymbol{\theta}^t\boldsymbol{x}_n + \theta_0) \geq 1,\ 1 \leq n \leq N$

- *Primal:* minimizar $\Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\alpha}) \;=\; \frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\theta} - \sum_{n=1}^{N} \alpha_n \left(c_n\,(\boldsymbol{\theta}^t\boldsymbol{x}_n + \theta_0) - 1\right)$

  sujeto a $\alpha_n \geq 0,\ 1 \leq n \leq N \qquad\longrightarrow\quad \boldsymbol{\theta}^\star(\boldsymbol{\alpha})$

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

  sujeto a: $\sum_{n=1}^{N}\alpha_n\,c_n = 0;\quad \alpha_n \geq 0,\ 1 \leq n \leq N \qquad\longrightarrow\quad \boldsymbol{\alpha}^\star$

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

Ventajas de la formulación *dual*: Solo aparecen productos escalares de vectores de S y desaparecen $\boldsymbol{\theta}$ y θ_0 (todo se puede expresar solo mediante las α_n). Además permite soluciones computacionales más eficientes.

Resumen de propiedades de los clasificadores de máximo margen

Las soluciones $\boldsymbol{\theta}^\star, \theta_0^\star, \boldsymbol{\alpha}^\star$ verifican:

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

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

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

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

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

   Esto implica que hay dos posibilidades para cada n:

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

Vectores soporte

- *Vectores soporte:* muestras de entrenamiento $\boldsymbol{x}_n$ para las que $\alpha_n^\star \neq 0$

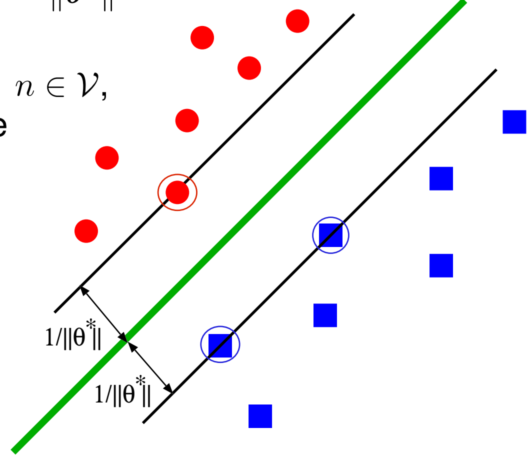
  $$\mathcal{V} = \left\{ n \in \mathbb{N}, \ 1 \leq n \leq N \mid (\boldsymbol{x}_n, c_n) \in S, \ c_n(\boldsymbol{\theta}^{\star t} \boldsymbol{x}_n + \theta_0^\star) = 1 \right\}$$

- Todos los vectores soporte equidistan del hiperplano separador:

  $$\forall n \in \mathcal{V}, \quad r_{\boldsymbol{x}_n} = \frac{|\, \boldsymbol{\theta}^{\star t} \boldsymbol{x}_n + \theta_0^\star \,|}{\|\boldsymbol{\theta}^\star\|} = \frac{|\, c_n \,|}{\|\boldsymbol{\theta}^\star\|} = \frac{1}{\|\boldsymbol{\theta}^\star\|}$$

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

  $$c_n = +1, \ c_{n'} = -1$$
  $$\alpha_n^\star, \alpha_{n'}^\star > 0,$$
  $$n \neq n'$$



Máquinas de vectores soporte

Un clasificador de máximo margen queda definido por la función discriminante lineal $\phi(\boldsymbol{x};\boldsymbol{\Theta}) \stackrel{\text{def}}{=} \boldsymbol{\theta}^{\star t}\boldsymbol{x} + \theta_0^\star$, 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: $\boldsymbol{\theta}^\star = \sum_{n=1}^{N} c_n\, \alpha_n^\star\, \boldsymbol{x}_n = \sum_{n\in\mathcal{V}} c_n\, \alpha_n^\star\, \boldsymbol{x}_n$
- Por KKT, para cualquier $m \in \mathcal{V}$: $\theta_0^\star = c_m - \boldsymbol{\theta}^{\star t}\boldsymbol{x}_m = c_m - \sum_{n\in\mathcal{V}} c_n\, \alpha_n^\star\, \boldsymbol{x}_n^t\boldsymbol{x}_m$

Función discriminante lineal que maximiza el margen:

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

DEMO DEMO

http://svm.cs.rhul.ac.uk/pagesnew/GPat.shtml

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) $\boldsymbol{\theta}^\star$ y $\theta_0^\star$ que clasifiquen S con el máximo margen.

2. Sea S una muestra linealmente separable. Demostrar que el margen óptimo es:

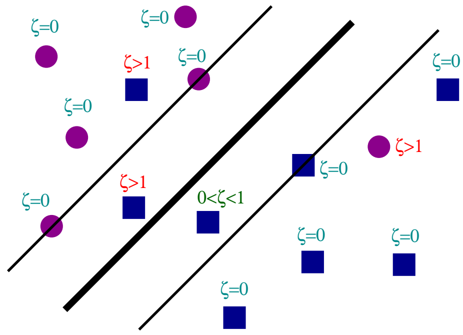
$$2\left(\sum_{n\in\mathcal{V}} \alpha_n^\star\right)^{-1/2}$$

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 $S = \{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}$ 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:
  - $c_n\,(\boldsymbol{\theta}^t\boldsymbol{x}_n + \theta_0) \;\geq\; 1 - \zeta_n\,, \;\; 1 \leq n \leq N$
  - $\zeta_n \;\geq\; 0\,,\; 1 \leq n \leq N$



SVM en el caso de no separabilidad lineal

- *Lagrangiana primal*:

  Minimizar $\Lambda(\boldsymbol{\theta}, \theta_0, \boldsymbol{\zeta}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \; =$
  $$\frac{1}{2}\boldsymbol{\theta}^t\boldsymbol{\theta} \; + \; \mathcal{C}\sum_{n=1}^{N}\zeta_n - \sum_{n=1}^{N}\alpha_n\left(c_n\,(\boldsymbol{\theta}^t\boldsymbol{x}_n + \theta_0) + \zeta_n - 1\right) - \sum_{n=1}^{N}\beta_n\zeta_n$$
  sujeto a $\alpha_n \geq 0, \;\; \beta_n \geq 0$ y $\zeta_n \geq 0$ para $1 \leq n \leq N$

- *Lagrangiana dual*:

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

  sujeto 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*:

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

- Las soluciones $\boldsymbol{\theta}^\star$, $\theta_0^\star$, $\zeta^\star$, $\boldsymbol{\alpha}^\star$, $\boldsymbol{\beta}^\star$ verifican:

1. $\boldsymbol{\theta}^\star = \sum_{n=1}^{N} c_n \, \alpha_n^\star \, \boldsymbol{x}_n$
2. $\sum_{n=1}^{N} \alpha_n^\star \, c_n = 0$
3. $0 \leq \alpha_n^\star \leq \mathcal{C} \quad 1 \leq n \leq N$
4. $\beta_n^\star = \mathcal{C} - \alpha_n^\star \quad 1 \leq n \leq N$
5. Condición complementaria de Karush-Kuhn-Tucker

$$\left.\begin{array}{rcl} \alpha_n^\star \left(c_i \, ({\boldsymbol{\theta}^\star}^t \boldsymbol{x}_n + \theta_0^\star) - 1 + \zeta_n^\star \right) & = & 0 \\ \beta_n^\star \zeta_n^\star & = & 0 \end{array}\right\} \; 1 \leq n \leq N$$

Vectores soporte “erróneos”

$$1 \leq n \leq N \quad \left\{ \begin{array}{rcll} \alpha_n^\star \left(c_n ({\boldsymbol{\theta}^\star}^t \boldsymbol{x}_n + \theta_0^\star) - 1 + \zeta_n^\star \right) & = & 0 & (1) \\ \beta_n^\star \, \zeta_n^\star \;=\; (\mathcal{C} - \alpha_n^\star) \, \zeta_n^\star & = & 0 & (2) \end{array} \right.$$

- (1) $\rightarrow$ muestras $\boldsymbol{x}_n$ tales que $\alpha_n^\star \neq 0$ son vectores soporte.
  En este caso, $c_n({\boldsymbol{\theta}^\star}^t \boldsymbol{x}_n + \theta_0^\star) \;=\; 1 - \zeta_n^\star$

- (2) $\rightarrow$ $(\mathcal{C} - \alpha_n^\star) \, \zeta_n^\star = 0$
  - $\mathcal{C} - \alpha_n^\star > 0 \;\Rightarrow\; \zeta_n^\star = 0 \;\Rightarrow\; c_n \, ({\boldsymbol{\theta}^\star}^t \boldsymbol{x}_n + \theta_0^\star) = 1 \;\Rightarrow$ sin error de margen
  - $\zeta_n^\star > 0 \;\Rightarrow\; \mathcal{C} = \alpha_n^\star \;\Rightarrow$ error de margen $\left\{ \begin{array}{lll} \zeta_n^\star & > 1 & \text{error de clasificación} \\ \zeta_n^\star & \leq 1 & \text{dentro del margen} \end{array} \right.$

- $\mathcal{C}$ 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 ($\alpha_n^\star = 0$) ?

$\mathcal{C}-$SVM

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

Función discriminante lineal de margen máximo:

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

Demo

http://svm.cs.rhul.ac.uk/pagesnew/GPat.shtml

Métodos de optimización para SVM

Problema: *maximizar la Lagrangiana dual:*

$$\underset{\substack{\boldsymbol{\alpha} \\ \sum_{n=1}^{N} c_n \alpha_n = 0 \\ 0 \leq \alpha_n \leq C, \; 1 \leq n \leq N}}{\arg\max} \quad \Lambda_D(\boldsymbol{\alpha})$$

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

Index

Funciones discriminantes lineales generalizadas (FDLG)

- FDLG para un problema de clasificación en dos clases:

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

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

- Ejemplo: Linealización de funciones cuadráticas:

$$\phi(\boldsymbol{x};\boldsymbol{\Theta}) \;=\; \sum_{i=1}^{d}\sum_{j=1}^{d} a_{ij}x_ix_j + \sum_{j=1}^{d} b_jx_j + c$$

$\phi(\boldsymbol{x};\boldsymbol{\Theta}) \;=\; \boldsymbol{\theta}^t\boldsymbol{\psi}(\boldsymbol{x}) + \theta_0$ con $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{d'},\; d' = d^2 + d :$

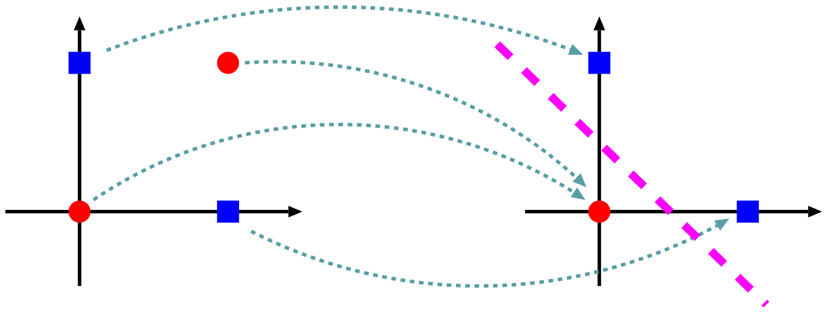
$$\boldsymbol{\psi}(x_1,\dots,x_d) \;=\; (x_1x_1, x_1x_2,\dots,x_2x_1,x_2x_2,\dots,x_dx_d,\; x_1,\dots,x_d)^t$$

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

Ejemplo: Problema de la o exclusiva (XOR)

Mediante la función escalón $E : \mathbb{R} \to \{0, 1\}$ 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:

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



permite definir una FDLG que linealiza el problema XOR:

$$\begin{aligned} \phi(x_1, x_2; \theta_0, \theta_1, \theta_2) &= \sum_{k=1}^{2} \theta_k \psi_k(x_1, x_2) + \theta_0 \\ //\theta_1 = \theta_2 = 1,\ \theta_0 = -0.5// &= E(x_1 - x_2 - 0.5) + E(-x_1 + x_2 - 0.5) - 0.5 \end{aligned}$$

Generalización de SVM [1]: Núcleos

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

$$\phi(\boldsymbol{x}) = \sum_{n=1}^{N} \alpha_n\, c_n\, \boldsymbol{x}_n{}^t \boldsymbol{x} \;+\; \theta_0$$

$$\Downarrow$$

$$\phi_\psi(\boldsymbol{x}) \;=\; \sum_{n=1}^{N} \alpha_n\, c_n\, \boldsymbol{\psi}(\boldsymbol{x}_n)^t \boldsymbol{\psi}(\boldsymbol{x}) \;+\; \theta_0$$

$$\Downarrow$$

$$\phi_{\mathcal{K}}(\boldsymbol{x}) \;=\; \sum_{n=1}^{N} \alpha_n\, c_n\, \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}) \;+\; \theta_0$$

donde $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ es una función que se denomina núcleo si $\exists \boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{d'}$ tal que $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\psi}(\boldsymbol{x})^t \boldsymbol{\psi}(\boldsymbol{x}'), \;\; \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$; es decir:
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 $\boldsymbol{x} = (x_1, x_2, x_3)^t$, $\boldsymbol{y} = (y_1, y_2, y_3)^t$ y $\mathcal{K}\colon \mathbb{R}^3 \to \mathbb{R}$ definida como:

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

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

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

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

$$\begin{aligned}\boldsymbol{\psi}(\boldsymbol{x}) &= (x_1^2,\ x_2^2,\ x_3^2,\ \sqrt{2}\, x_1\, x_2,\ \sqrt{2}\, x_1\, x_3,\ \sqrt{2}\, x_2\, x_3)^t \\ \boldsymbol{\psi}(\boldsymbol{y}) &= (y_1^2,\ y_2^2,\ y_3^2,\ \sqrt{2}\, y_1\, y_2,\ \sqrt{2}\, y_1\, y_3,\ \sqrt{2}\, y_2\, y_3)^t\end{aligned}$$

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

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

Construcción de núcleos

- Elegir $\boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{d'}$ y el núcleo es $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\psi}(\boldsymbol{x})^t \boldsymbol{\psi}(\boldsymbol{x}')$
  Es necesario trabajar en $\mathbb{R}^{d'}$, $d' \gg d$: ¡amenaza de la dimensionalidad!
- Elegir $\mathcal{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ y:
  - Demostrar que $\exists \boldsymbol{\psi} : \mathbb{R}^d \to \mathbb{R}^{d'}$: $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\psi}(\boldsymbol{x})^t \boldsymbol{\psi}(\boldsymbol{x}'),\ \ \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{R}^d$
  - *Condición de Mercer*: $\mathcal{K}$ es un núcleo *si* y *solo si*, para cualquier conjunto de vectores $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$, la matriz $[\mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}_m)]_{1 \leq n,m \leq N}$ (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 $\mathcal{K}_2$
    * $\exp\left(\mathcal{K}_1(\boldsymbol{x}, \boldsymbol{x}')\right)$
  - *Núcleos de base radial* (RBK): $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') \stackrel{\text{def}}{=} f(r),\ r = || \boldsymbol{x} - \boldsymbol{x}' ||$
    Ejemplo: núcleos gaussianos: $\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}') = \exp\left(-c\, || \boldsymbol{x} - \boldsymbol{x}' ||^2\right)$
    ver: http://en.wikipedia.org/wiki/Radial_basis_function
  - $\cdots$ Etc.

Máquinas de vectores soporte y núcleos

- *Problema*: Dado un muestra de entrenamiento $S = \{(\boldsymbol{x}_1, c_1), \ldots, (\boldsymbol{x}_N, c_N)\}$, una constante $\mathcal{C} \geq 0$ y un núcleo $\mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}_m) = \boldsymbol{\psi}(\boldsymbol{x}_n)^t \boldsymbol{\psi}(\boldsymbol{x}_m)$, 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 $c_n(\boldsymbol{\theta}^t\boldsymbol{\psi}(\boldsymbol{x}_n) + \theta_0) \geq 1 - \zeta_n$ y $\zeta_n \geq 0,\ 1 \leq n \leq N$
- *Solución:* Lagrangiana dual: $\Lambda_D(\boldsymbol{\alpha}) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2}\sum_{n,m=1}^{N} c_n\, c_m\, \alpha_n\, \alpha_m\, \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}_m)$
  - Calcular $\boldsymbol{\alpha}^\star$ 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:* $\boldsymbol{x}_n \in S : n \in \mathcal{V},\ \ \mathcal{V} = \{n \in \mathbb{N}, 1 \leq n \leq N \mid \alpha_n^\star \neq 0\}$
  - *FDLG*: $\phi_{\mathcal{K}}(\boldsymbol{x}) = \sum_{n \in \mathcal{V}} c_n\, \alpha_n^\star\, \mathcal{K}(\boldsymbol{x}_n, \boldsymbol{x}) + \theta_0^\star$

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

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

Index

Clasificación en C clases con funciones discriminantes lineales

FUNCIONES DISCRIMINANTES LINEALES GENERALIZADAS

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

donde:

- $\boldsymbol{\psi}:\mathbb{R}^d\to\mathbb{R}^{d'}$ 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

$$f(\boldsymbol{x}) \overset{\text{def}}{=} \hat{c} \;=\; \underset{1\le c\le C}{\arg\max}\,\phi_c(\boldsymbol{x};\boldsymbol{\Theta}) \iff \phi_{\hat{c}}(\boldsymbol{x};\boldsymbol{\Theta}) > \phi_c(\boldsymbol{x};\boldsymbol{\Theta})\;\forall c\neq\hat{c}$$

El problema de C clases: uno-contra-uno

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

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

- $C\,(C-1)/2$ clasificadores uno-contra-uno
  - *Aprendizaje:* Para $1\le c<c'\le C$,

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

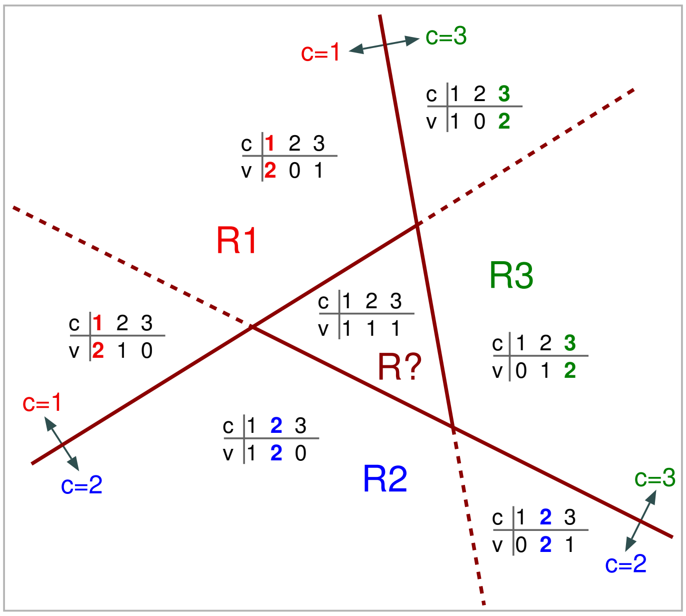
$$\phi_{cc'}(\boldsymbol{x}) \;=\; \sum_{\boldsymbol{x}_n\in\mathcal{SV}_{cc'}}\alpha^\star_{ncc'}\,c_{ncc'}\,\mathcal{K}(\boldsymbol{x}_n,\boldsymbol{x})+\theta^\star_{cc'0}$$

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

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

Clasificación multi-clase mediante clasificadores binarios: ejemplo

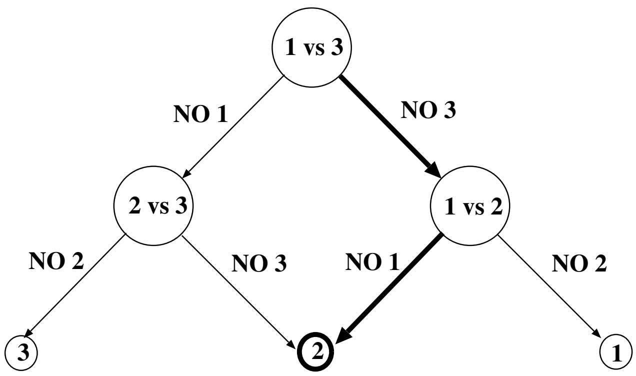
Uno-contra-uno por votación



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:
$$f(\boldsymbol{x}) = \underset{1 \leq c \leq C}{\arg\max} \sum_{c' \neq c} f_{cc'}(\boldsymbol{x})$$

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



Otras técnicas de clasificación multi-clase para SVM

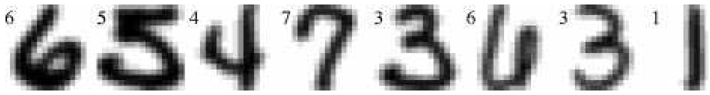
- *Uno-contra-el-resto:* Se entrenan C FDLs binarias, $\phi_c, 1 \leq c \leq C$, usando como muestras positivas *solo* los vectores de la clase c y como negativas el resto.
- *SVM*$^{\text{multiC}}$*:* Optimización directa de márgenes en C clases [Cramer & Singer, 01]
- *Construcción de Kesler:* Transformación de un problema de C clases on 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`

Index

Aplicaciones: reconocimiento de caracteres manuscritos off-line

K.-R. Müller et al: An Introduction to Kernel-Based Learning Algorithms. IEEE Trans. on Neural networks. 2001

Corpus USPS: 7291 muestras



Técnica	Tasa de error (%)
SVM sin kernel	8.7
k-vecino más próximo	5.7
Redes neuronales radiales	4.1
SVM virtuales	3.0
Vecino más próximo utilizando la distancia tangente	2.5
Humano	2.5

Aplicaciones: datos DNA

K.-R. Müller et al: An Introduction to Kernel-Based Learning Algorithms. IEEE Trans. on Neural networks. 2001

Reconocimiento de secuencias en texto genómico:
8000 muestras para entrenamiento y 3000 para test

Técnica	Tasa de error (%)
Redes neuronales	15.4
SVM sin kernel	13.2
SVM con un kernel particular	11.4

Aplicaciones diversas

- Búsqueda de imágenes

  E. Chang et al.: Support Vector Machine Concept-Dependent Active Learning For Image Retrieval. IEEE Trans. on Multimedia, 2005.
- Detección de caras

  E. Osuna et al.: Training support vector machines:an application to face detection. Proc. of CVPR'97, 1997
- Localización de las matrículas

  K. In Kim et .al: Fast Color Texture-Based Object Detection in Images: Application to License Plate Localization. Springer 2005
- Detección de texto en imágenes

  K. In Kim et .al: Support vector machine-based text detection in digital video. Pattern Recognition, 34(2), 2001
- Detección de humanos en movimiento

  H. Sidenbladh: Detecting Human Motion with Support Vector Machines, Proceedings of the 17th ICPR, 2:188–191, 2004
- Detección de mensajes ocultos en imágenes

  S. Lyu, H. Farid: Detecting Hidden Messages Using Higher-Order Statistics and Support Vector Machines. 5th Int. Work. on Inform. Hiding, 2002
- Clasificación de texto
- Predicción del nivel del agua de un lago

  M. S. Khan and P. Coulibaly. Application of Support Vector Machine in Lake Water Level Prediction. J. of Hydrologic Engineering 2006
- Series temporales financieras

  E.H. Tay, L.J. Cao: Modified support vector machines in financial time series forecasting. In *Neurocomputing*. 2002

Index

Notación

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