Máquinas de Vectores de Soporte

Sistemas Inteligentes

Dra. Graciela Meza Lovón
gmezal@ucsp.edu.pe

Maestría en Ciencia de la Computación
Universidad Católica San Pablo

Basadas en Lecture Notes de Andrew Ng[4]
Tutorial de SVM de Carmona [6].

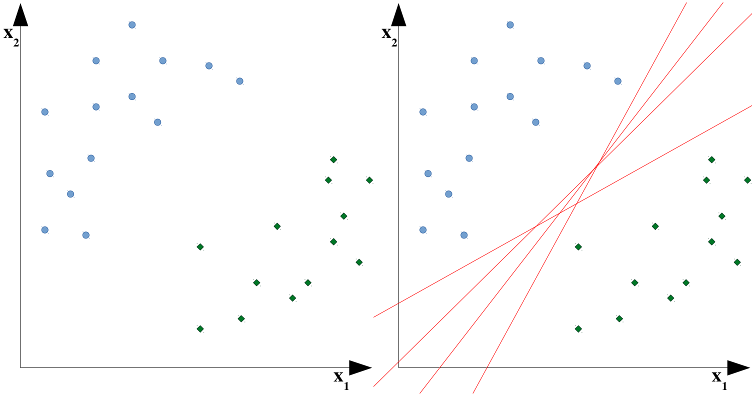
Noviembre, 2018

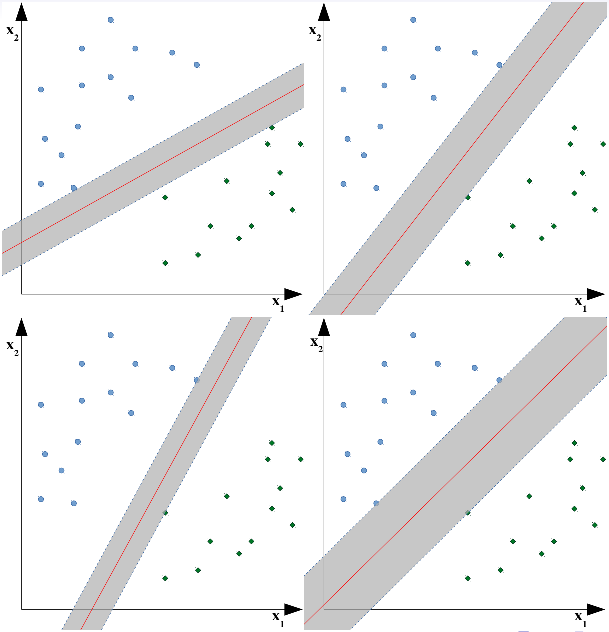
Introducción

- Las bases del algoritmo fueron dadas por Vladimir Vapnik y Alexey Chervonenkis [7] en 1964.
- En 1992, Bernhard E. Boser, Isabelle M. Guyon y Vladimir N. Vapnik [2] sugirieron una manera de crear clasificadores no lineales aplicando el truco del kernel a los hiperplanos de margen máximo.
- En 1995, el algoritmo fue propuesto por Corina Cortes y Vapnik [3].

Contenido

Introducción




x2
x1
x2
x1
x2
x1
x2
x1

- Dado un conjunto separable de ejemplos $S = \{(\boldsymbol{x}^{(1)}, y^{(1)}), \ldots, (\boldsymbol{x}^{(m)}, y^{(m)})\}$, donde $\boldsymbol{x}^{(i)} \in \mathbb{R}^d$ e $y^{(i)} \in \{+1, -1\}$, se define un hiperplano de separación (una función lineal) capaz de separar el conjunto sin errores:

$$D(\boldsymbol{x}) = (w_1 x_1) + \ldots + w_d x_d) + b = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b$$

donde $\boldsymbol{w}$ y b son coeficientes reales.

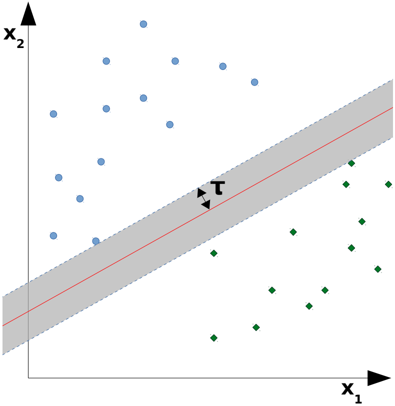
- Para todo $\boldsymbol{x}^{(i)}$ del conjunto, el hiperplano de separación cumplirá estas restricciones:
    - $\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b \geq 0$ si $y^{(i)} = +1$,
    - $\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b \leq 0$ si $y^{(i)} = -1$.

Margen de un Hiperplano de Separación

- De forma más compacta

$$y^{(i)}(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b) \geq 0, i = 1, \ldots, m$$

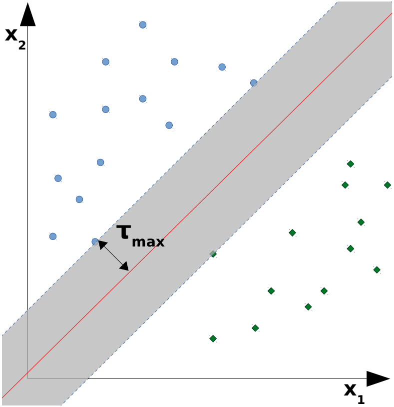
- Se define el concepto de **margen de un hiperplano de separación**, τ, como la mínima distancia entre dicho hiperplano y el ejemplo más cercano de cualquiera de las dos clases.



Margen de un Hiperplano Óptimo de Separación

- Un hiperplano de separación será óptimo si su margen es de tamaño máximo.
- La distancia entre un hiperplano de separación $D(\boldsymbol{x})$ a un ejemplo $\boldsymbol{x'}$ es:

$$\frac{|D(\boldsymbol{x'})|}{||\boldsymbol{w}||}$$



- Todos los ejemplos de entrenamiento cumplirán que:

$$\frac{y^{(i)}D(\boldsymbol{x}^{(i)})}{||\boldsymbol{w}||} \geq \tau_{max}, i = 1, \ldots, m$$

Formulación del Problema

- Encontrar el hiperplano óptimo es equivalente a encontrar el valor de $\boldsymbol{w}$ que maximiza el margen.
- Ya que existen un número infinito de hiperplanos óptimos, es necesario limitar los hiperplanos separables escalando el producto de τ_{max} y la norma de $\boldsymbol{w}$ a 1, i.e.,

$$\tau_{max}||\boldsymbol{w}|| = 1.$$

- Aumentar el margen equivale a disminuir la norma de $\boldsymbol{w}$, i.e.,

$$\tau_{max} = \frac{1}{||\boldsymbol{w}||}$$

- Un hiperplano de separación para cual se obtenga un valor mínimo de $||\boldsymbol{w}||$ y que esté restringido a $\frac{y^{(i)}D(\boldsymbol{x}^{(i)})}{||\boldsymbol{w}||} \geq \tau_{max}$ y a $\tau_{max}||\boldsymbol{w}|| = 1$ será óptimo, i.e., un hiperplano de separación óptimo cumple que:

$$y^{(i)}(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b) \geq 1, i = 1, \ldots, m$$

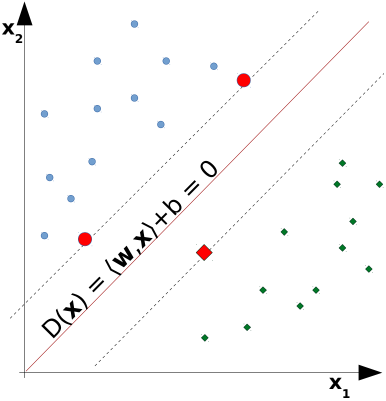
- El problema de encontrar el margen óptimo es un problema de optimización cuadrática:

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}||\boldsymbol{w}||^2 \\ \text{s.t.} \quad & y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1, i = 1, \ldots, m \end{aligned}$$

- Los ejemplos $\boldsymbol{x}^{(i)}$ que dan soporte al hiperplano óptimo son los mismos que definen el margen óptimo, i.e., aquellos para los que se cumple la igualdad

$$y^{(i)}(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b) = 1$$

son los vectores de soporte.



Dualidad de Lagrange

- Se llama "Problema Primal" de optimización cuadrática a:

$$\begin{aligned} \min_{\boldsymbol{w}} \quad & f(\boldsymbol{w}) \\ \text{s.t.} \quad & g_i(\boldsymbol{w}) \leq 0, i = 1, \ldots, k \\ & h_i(\boldsymbol{w}) = 0, i = 1, \ldots, l. \end{aligned}$$

- Para encontrar su "Problema Dual", encontramos la función Lagrangiana:

$$L(\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = f(\boldsymbol{w}) + \sum_{i=1}^{k} \alpha_i g_i(\boldsymbol{w}) + \sum_{i=1}^{k} \beta_i h_i(\boldsymbol{w}),$$

  donde α_i's β_i's son los multiplicadores de Lagrange.
- Bajo ciertas condiciones existen $\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ tales
  - $\boldsymbol{w}^*$ es la solución del problema primal
  - $\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*$ son la solución del problema dual
  - La solución del problema Dual es la solución del Primal.
  - $\boldsymbol{w}^*, \boldsymbol{\alpha}^*$ y $\boldsymbol{\beta}^*$ satisfacen las condiciones de Karush-Kuhn-Tucker (KKT).

- Para $\boldsymbol{w}^*, \boldsymbol{\alpha}^*$ y $\boldsymbol{\beta}^*$ las condiciones de KKT son:

$$
\begin{aligned}
\frac{\partial}{\partial w_i}\mathcal{L}(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= 0, i = 1, \ldots, n & (1)\\
\frac{\partial}{\partial \beta_i}\mathcal{L}(\boldsymbol{w}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) &= 0, i = 1, \ldots, l & (2)\\
\alpha_i^* g_i(\boldsymbol{w}^*) &= 0, i = 1, \ldots, k & (3)\\
g_i(\boldsymbol{w}^*) &\leq 0, i = 1, \ldots, k & (4)\\
\alpha_i^* &\geq 0, i = 1, \ldots, k & (5)
\end{aligned}
$$

- A la Ecuación 3 se le conoce como la condición complementaria dual KKT.
  - Si $\alpha_i^* > 0$, entonces $g_i(\boldsymbol{w}^*) = 0$. (i.e., " $g_i(\boldsymbol{w}) \leq 0$" es una restricción activa, i.e., se cumple la igualdad y no la desigualdad).

Formulación del Problema Dual

- Problema de optimización primal para el margen óptimo:

$$\begin{aligned} \min_{\boldsymbol{w},b} \quad & \frac{1}{2}||\boldsymbol{w}||^2 \\ \text{s.t.} \quad & y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) \geq 1, i = 1, \ldots, m \end{aligned}$$

- Las restricciones pueden reescribirse como:

$$g_i(\boldsymbol{w}) = -y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) + 1 \leq 0.$$

- De la condición de complementariedad dual KKT, $\alpha_i > 0$ solo para los ejemplos que tienen un margen igual a 1 (los correspondientes a restricciones con igualdad, $g_i(\boldsymbol{w}) = 0$).

- Para el encontrar el problema de optimización dual, se debe calcular el Lagrangiano:

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{m} \alpha_i [y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b) - 1] \quad (6)$$

- Ya que el problema solo tiene desigualdades solo se considera los multiplicadores de Lagrange "α_i " pero no "β_i"
- Para encontrar la forma dual, hay que aplicar las condiciones de KKT. En particular, aplicando la Ecuación 1 obtenemos:

$$\frac{\partial \mathcal{L}}{\boldsymbol{w}}(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}) = \boldsymbol{w}^* - \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} = 0. \quad (7)$$

  - Luego,

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)} \quad (8)$$

- Para las derivadas con respecto a b (aplicando 2), tenemos:

$$\frac{\partial}{\partial b}\mathcal{L}(\boldsymbol{w}^*, b^*, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i y^{(i)} = 0. \tag{9}$$

- Si la definición de $\boldsymbol{w}^*$ en la Ecuación 8, la reemplazamos en el lagrangiano (Ecuación 6) y simplificamos, obtenemos

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)} - b\sum_{i=1}^{m} \alpha_i y^{(i)}.$$

- El último término debe ser cero (por la la Ecuación 9), por lo que obtenemos la función objetivo dual a ser optimizada:

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j (\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)}. \tag{10}$$

- El problema de optimización dual se obtiene maximizando la función de la Ecuación 10 sujeta a las restricciones $\alpha_i \geq 0$ (que se mantienen de la formulación anterior) y la restricción obtenida en la Ecuación 9 (i.e., $\sum_{i=1}^{m} \alpha_i y^{(i)} = 0$):

$$
\begin{aligned}
\max_{\boldsymbol{\alpha}} \quad & W(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{j)} \rangle. \\
\text{s.t.} \quad & \alpha_i \geq 0, i = 1, \ldots, m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0
\end{aligned}
$$

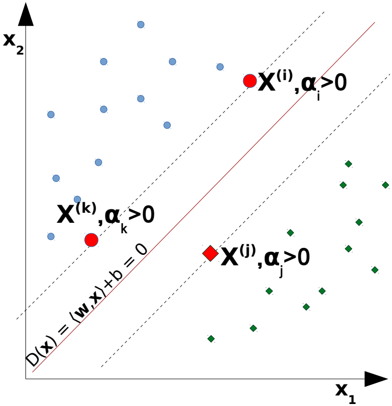
- Al solucionar el problema dual se encuentra $\boldsymbol{\alpha}^*$. Luego, se soluciona el problema primal que consiste en encontrar encontrar $\boldsymbol{w}^*$ y por ende $\boldsymbol{w}^T\boldsymbol{x} + b$.
- Se debe reescribir $\boldsymbol{w}^T\boldsymbol{x} + b$ usando 8, de lo que se obtiene:

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{x} + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}\right)^T \boldsymbol{x} + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle \boldsymbol{x}^{(i)}, \boldsymbol{x} \rangle + b.\end{aligned}$$

- Analizando la condición complementaria, se observa que para un ejemplo $(\boldsymbol{x^{(i)}}, y^{(i)})$, cuyo $\alpha_i > 0$ se tiene que

$$y^{(i)}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(i)} + b^*) = 1$$

- Note que para $(\boldsymbol{x}^{(i)}, y^{(i)})$, con $\alpha_i > 0$ se cumple la igualdad y no la desigualdad.
- Los ejemplos para los que se cumple la igualdad son los vectores de soporte.

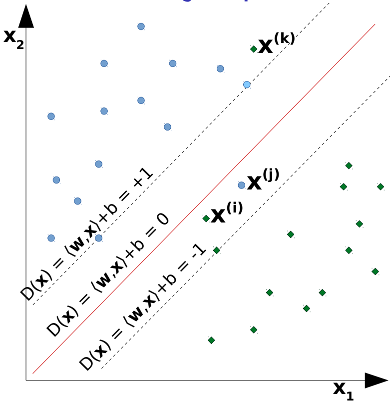


- Finalmente, se calcula b^*:

$$b^* = y^{(i)} - (\boldsymbol{w}^*)^T \boldsymbol{x}^{(i)},$$

donde $\boldsymbol{x}^{(i)}$ es cualquier vector de soporte.

Caso: Ejemplos Casi Separables Linealmente

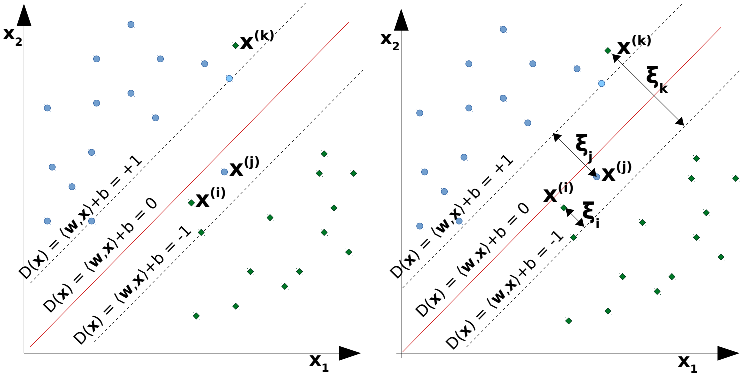


- Los problemas reales se caracterizan normalmente por poseer ejemplos ruidosos, i.e., el caso de ejemplos linealmente separables es un ideal que muy difícilmente se cumplirá.
- Un ejemplo $(y^{(i)}, \boldsymbol{x}^{(i)})$ es no separable si no cumple:

$$y_i(\boldsymbol{w}^T \boldsymbol{x}^{(i)}) \geq 1$$

- Dos casos de ejemplos no separables:
  - El ejemplo cae dentro del margen asociado a la clase correcta, e.g., $\boldsymbol{x}^{(i)}$.
  - El ejemplo cae al otro lado de dicho hiperplano. No es clasificado correctamente. e.g., $\boldsymbol{x}^{(j)}$ y $\boldsymbol{x}^{(k)}$.

Variables de Holgura



- Solución: Relajar el grado de separabilidad del conjunto de ejemplos, permitiendo que haya errores de clasificación. Para ello se usan variables de holgura, $\xi_i, i = 1, \ldots m$, $\xi_i \geq 0, \xi \in \mathbb{R}$.

- Para un ejemplo $(\boldsymbol{x}^{(i)}, y^{(i)})$, su variable de holgura, ξ_i, representa la desviación del caso separable, medida desde $\boldsymbol{x}^{(i)}$ hasta borde del margen correspondiente a la clase $y^{(i)}$.
  - Si $\xi_i = 0$: Ejemplos separables
  - Si $\xi_i \geq 0$: Ejemplos no separables bien clasificados
  - Si $\xi_i \geq 1$: Ejemplos no separables y mal clasificados.
- La suma de todas las variables de holgura, $\sum_{i=1}^{m} \xi_i$: mide el costo asociado al número de ejemplos no separables.
- Una variable de holgura mayor que cero ($\xi_i > 0$) permite que el margen del ejemplo $\boldsymbol{x}^{(i)}$ sea menor que 1.

$$y^{(i)}(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)} \rangle + b) \geq 1 - \xi_i, \xi_i \geq 0, \xi \in \mathbb{R}, i = 1, \ldots, m$$

- Sin embargo, es necesario modificar la función objetivo a fin de que el hecho de incrementar el valor de ξ_i tenga asociado un costo proporcional al valor de ξ_i.
- Para incluir dicho costo, se define una nueva función objetivo a optimizar:

$$f(\boldsymbol{w}, \boldsymbol{\xi}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m} \xi_i$$

  donde C es una constante que controla el grado en que el costo de ejemplos no separables incluye en la minimización de la norma.

Problema Primal

- El nuevo problema de optimización incluyendo las variables de holgura es:

$$
\begin{array}{ll}
\underset{\boldsymbol{w}}{min} & \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} & y^{(i)}(\langle \boldsymbol{w}, \boldsymbol{x}^{(i)}\rangle + b) + \xi_i - 1 \geq 0 \\
 & \xi_i \geq 0, i = 1, \ldots, m
\end{array}
$$

- Al hiperplano obtenido inlcuyendo variables de holgura se le llama **hiperplano de separación de margen blando**. Al hiperplano obtenido en el caso separable, se le llama **hiperplano de separación de margen duro**.
- Como en el caso separable, el problema de optimización puede ser transformado a su forma dual.

Problema Dual

- Obtención de la función Lagrangiana

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i[y^{(i)}(\boldsymbol{w}^T\boldsymbol{x}^{(i)} + b) + \xi_i - 1] - \sum_{i=1}^{m}\beta_i\xi_i \quad (11)$$

- Aplicación de las condiciones de KKT:

$$\frac{\partial\mathcal{L}(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial\boldsymbol{w}} = \boldsymbol{w}^* - \sum_{i=1}^{m}\alpha_i y^{(i)}\boldsymbol{x}^{(i)} = 0 \quad (12)$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial b} = \sum_{i=1}^{m}\alpha_i y^{(i)} = 0 \quad (13)$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}, \boldsymbol{\beta})}{\partial\xi_i} = C - \alpha_i - \beta_i = 0 \quad (14)$$

$$\alpha_i[1 - y^{(i)}[((\boldsymbol{w}^*)^T\boldsymbol{x}^{(i)} + b^*) - \xi_i] = 0, i = 1, \ldots, m \quad (15)$$

$$\beta_i\xi_i = 0, i = 1, \ldots, m \quad (16)$$

- La Ecuación 12 implica que la relación entre las variables del problema primal $(\boldsymbol{w}^*, b^*, \boldsymbol{\xi_i})$ con las del problema dual $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ es:

$$\boldsymbol{w}^* = \sum_{i=1}^{m} \alpha_i y^{(i)} \boldsymbol{x}^{(i)}$$

- De las ecuaciones 13 y 14 se establecen las restricciones adicionales KKT:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0 \qquad\qquad C = \alpha_i + \beta_i$$

- Eliminando las variables primales del Lagrangiano (Ec. 11) se obtiene que:

$$\mathcal{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (\boldsymbol{x}^{(i)})^T \boldsymbol{x}^{(j)} \tag{17}$$

- De Ec. 17, se obtiene la formalización del problema dual:

$$\max_{\boldsymbol{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle \boldsymbol{x}^{(i)}, \boldsymbol{x}^{j)} \rangle.$$

$$\text{s.t.} \qquad \sum_{i=1}^{m} \alpha_i y^{(i)} = 0$$

$$0 \leq \alpha_i \leq C, i = 1, \ldots, m$$

Algunas Deducciones

- El hiperplano de separación óptima en términos de α^*:

$$D(\boldsymbol{x}) = \sum_{i=1}^{m} \alpha_i^* y^{(i)} \langle \boldsymbol{x}, \boldsymbol{x}^{(i)} \rangle + b^*$$

- Si $\alpha_i = 0$, entonces $\xi_i = 0$ (16). Cada ejemplo $\boldsymbol{x}^{(i)}$ cuyo α_i sea igual a cero corresponde a un ejemplo separable ($\xi_i = 0$) (Por $C = \alpha_i + \beta_i$).
- Todo ejemplo no separable, $\boldsymbol{x}^{(i)}$, tiene un $\xi_i > 0$. Se deduce que $\alpha_i = C$.
  - Cuando $\alpha_i = C$, se deduce que $\alpha_i \neq 0$, se deduce por la Ec.15 que:

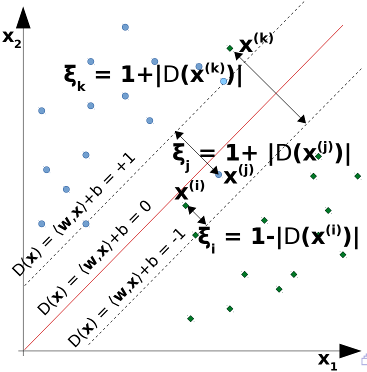
$$1 - y^{(i)}(\langle \boldsymbol{w}^*, \boldsymbol{x}^{(i)} \rangle + b^*) - \xi_i = 0, \quad \text{i.e.} \quad 1 - y^{(i)} D(\boldsymbol{x}^{(i)}) = \xi_i$$

- Cuando $\alpha_i = C$, se pueden considerar dos casos:
  1) El ejemplo $\boldsymbol{x}^{(i)}$, aunque es no separable, está bien clasificado, i.e.,

$$y^{(i)} D(\boldsymbol{x}^{(i)}) \geq \text{ entonces } \xi_i = 1 - |D(\boldsymbol{x}^{(i)})|$$

  2) El ejemplo $\boldsymbol{x}^{(i)}$, no es separable y está mal clasificado, i.e.,

$$y^{(i)} D(\boldsymbol{x}^{(i)}) < 0, \text{ entonces } \xi_i = 1 + |D(\boldsymbol{x}^{(i)})|$$



- Si $0 < \alpha_i < C$
  - De la restricción "$C = \alpha_i + \beta_i$" se deduce que $\beta_i \neq 0$ y a su vez de la restricción "$\beta_i \xi_i = 0$", se deduce que $\xi_i = 0$
  - De la restricción "$\alpha_i[1 - y^{(i)}(\boldsymbol{w}^T \boldsymbol{x}^{(i)} + b^*) - \xi_i] = 0$" y ya que $(\xi_i = 0)$, se deduce que

$$1 - y^{(i)}(\langle \boldsymbol{w}^*, \boldsymbol{x}^{(i)} \rangle + b^*) = 0$$

- Un ejemplo $\boldsymbol{x}^{(i)}$, es un vector soporte si y solo si $0 < \alpha_i < C$.
- Se calcula el valor b^*:

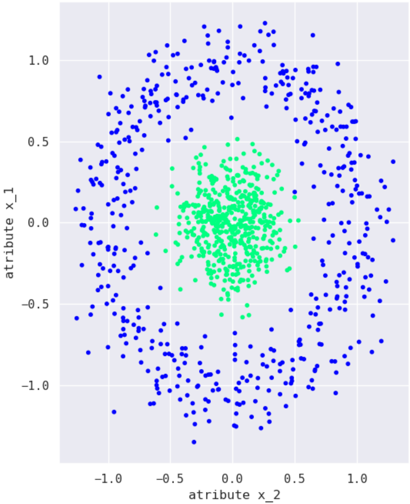
$$b^* = y^{(i)} - \langle \boldsymbol{w}^*, \boldsymbol{x}^{(i)} \rangle$$

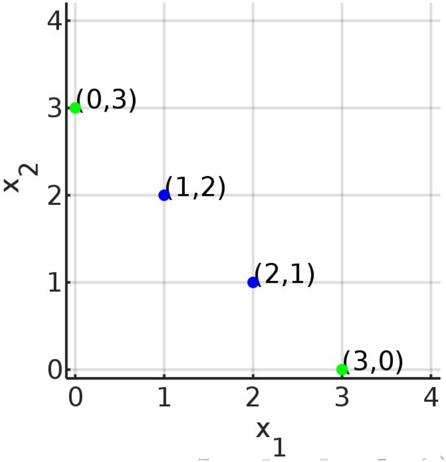
- Se puede expresar b^* en términos de las variables duales:

$$b^* = y^{(i)} - \sum_{j=1}^{m} \alpha_i^* y^{(i)} \langle \boldsymbol{x}^{(j)}, \boldsymbol{x}^{(i)} \rangle$$

Kernels

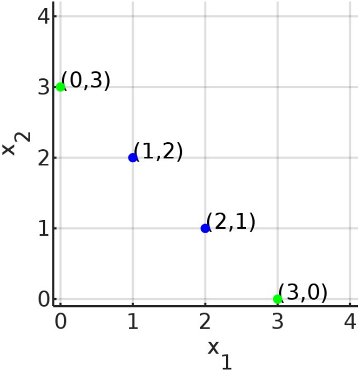
- El conjunto no puede ser separado por una función lineal.

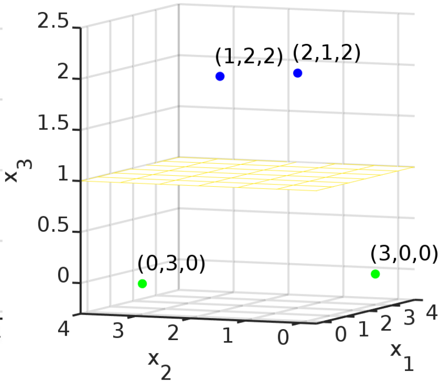




Kernels

- El conjunto no puede ser separado por una función lineal.





Kernels

- La idea es obtener una separación lineal usando una función, $\Phi(\boldsymbol{x})$, que mapee de los datos de entrada (del espacio de los ejemplos o espacio - $\mathcal{X}$) a un espacio dimensional superior (espacio de las características o espacio - $\mathcal{F}$).
- La frontera de decisión obtenida en el espacio de las características será lineal pero al mapearla al espacio de los ejemplos será una frontera de decisión no lineal.
- E.g. sea un $\boldsymbol{x} \in \mathbb{R}^n$ donde $n = 2$, se define una función de mapeo Φ que lleva a $\boldsymbol{x}$ de $\mathbb{R}^2$ a $\mathbb{R}^3$ como sigue:

$$\Phi(\boldsymbol{x}) = \begin{bmatrix} \phi_1(\boldsymbol{x}) \\ \phi_2(\boldsymbol{x}) \\ \phi_3(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} x_1 \\ x_2 \\ x_1 x_2 \end{bmatrix}$$

- Sea Φ una función de mapeo. Definimos una función Kernel, K tal que $K(\boldsymbol{x},\boldsymbol{z}) = \langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle = \Phi(\boldsymbol{x})^T \Phi(\boldsymbol{z})$.
- E.g., sean $\boldsymbol{x} \in \mathbb{R}^3$ y $\boldsymbol{z} \in \mathbb{R}^3$, para calcular $\langle \Phi(\boldsymbol{x}), \Phi(\boldsymbol{z}) \rangle$ usando la función de mapeo Φ definida como:

$$\Phi(\boldsymbol{x}) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix} \text{ obtenemos que } \Phi(\langle \boldsymbol{x}), \Phi(\boldsymbol{z} \rangle) = \langle \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}, \begin{bmatrix} z_1z_1 \\ z_1z_2 \\ z_1z_3 \\ z_2z_1 \\ z_2z_2 \\ z_2z_3 \\ z_3z_1 \\ z_3z_2 \\ z_3z_3 \end{bmatrix} \rangle =$$

$$\sum_{i,j=1}^{3} (x_ix_j)(z_iz_j) = \sum_{i=1}^{3}\sum_{j=1}^{3} x_ix_jz_iz_j = \left(\sum_{i=1}^{3} x_iz_i\right)\left(\sum_{j=1}^{3} x_jz_j\right) = (\boldsymbol{x}^T\boldsymbol{z})^2$$

- De manera general (n dimensiones),

$$\begin{aligned}
\langle \phi(\boldsymbol{x}), \phi(\boldsymbol{z}) \rangle &= \sum_{i,j=1}^{n} (x_i x_j)(z_i z_j) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} x_i x_j z_i z_j \\
&= \left(\sum_{i=1}^{n} x_i z_i \right) \left(\sum_{j=1}^{n} x_j z_j \right) \\
&= (\boldsymbol{x}^T \boldsymbol{z})^2 \\
&= K(\boldsymbol{x}, \boldsymbol{z})
\end{aligned}$$

- Para calcular $\Phi(\boldsymbol{x})$ se requiere $O(n^2)$ mientras que $K(\boldsymbol{x}, \boldsymbol{z})$ solo requiere $O(n)$.
- IMPORTANTE: Más general aún, para calcular $(\boldsymbol{x}^T \boldsymbol{z})^d = K(\boldsymbol{x}, \boldsymbol{z})$ se requiere $O(n^d)$ mientras que $K(\boldsymbol{x}, \boldsymbol{z})$ solo requiere $O(n)$.

Kernels Válidos

- Otra interpretración: $K(\boldsymbol{x},\boldsymbol{z})$ puede ser considerado como una medida de similaridad entre $\boldsymbol{x}$ y $\boldsymbol{z}$.
  - Si $\Phi(\boldsymbol{x})$ y $\Phi(\boldsymbol{z})$ son cercanos, $K(\boldsymbol{x},\boldsymbol{z}) = \Phi(\boldsymbol{x})^T\Phi(\boldsymbol{z})$ debería ser una cantidad grande,
  - Si $\Phi(\boldsymbol{x})$ y $\Phi(\boldsymbol{z})$ están alejados entonces $K(\boldsymbol{x},\boldsymbol{z}) = \Phi(\boldsymbol{x})^T\Phi(\boldsymbol{z})$ debería ser una cantidad pequeña.
- Suponga que:
  - K es un kernel válido que se corresponde con la función de mapeo Φ,
  - existe un conjunto de m puntos $\{\boldsymbol{x}^{(1)},\ldots,\boldsymbol{x}^{(m)}\}$,
  - existe una matriz cuadrada $m \times m$ llamada matrix Kernel K (hemos sobrecargado la notación).

Luego,

- Si K es un kernel válido entonces se cumple que $K_{ij} = K(\boldsymbol{x}^{(i)},\boldsymbol{x}^{(j)}) = \Phi(\boldsymbol{x}^{(i)})^T\Phi(\boldsymbol{x}^{(j)}) = \Phi(\boldsymbol{x}^{(j)})^T\Phi(\boldsymbol{x}^{(i)}) = K(\boldsymbol{x}^{(j)},\boldsymbol{x}^{(i)}) = K_{ji}$, i.e., K es **simétrica**.

- Además, al calcular $\boldsymbol{z}^T K \boldsymbol{z}$ para cualquier vector $\boldsymbol{z}$ tenemos que:

$$
\begin{aligned}
\boldsymbol{z}^T K \boldsymbol{z} &= \sum_i \sum_j z_i K_{ij} z_j \\
&= \sum_i \sum_j z_i \Phi(\boldsymbol{x}^{(i)})^T \Phi(\boldsymbol{x}^{(j)}) z_j \\
&= \sum_i \sum_j z_i \sum_k \phi_k(\boldsymbol{x}^{(i)}) \phi_k(\boldsymbol{x}^{(j)}) z_j \\
&= \sum_k \sum_i \sum_j z_i \phi_k(\boldsymbol{x}^{(i)}) \phi_k(\boldsymbol{x}^{(j)}) z_j \\
&= \sum_k \left(\sum_i z_i \phi_k(\boldsymbol{x}^{(i)}) \right)^2 \\
&\geq 0.
\end{aligned}
$$

- I.e., K es **semidefinida positiva**.

Teorema de Mercer

- Los resultados obtenidos se deben al teorema de Mercer: *Sea* $K : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$. *Para que* K *sea un kernel válido, es necesario y suficiente que para cualquier* $\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$, $(m < \infty)$, *la correspondiente matriz kernel sea simétrica y semidefinida positiva.*

Tipos de Kernels

- Algunas funciones kernel que cumplen con el Teorema de Mercer son:
  - Kernel Lineal: $K(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
  - Kernel Polinómico de grado p: $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle + a)^p$
  - Kernel Sigmoidal: $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma \langle \mathbf{x}_i, \mathbf{x}_j \rangle + a)$
  - Kernel Gaussiano: $K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma ||\mathbf{x}_i - \mathbf{x}_j||^2), \gamma > 0$

Software

- LIBSVM Chih-Chung Chang and Chih-Jen Lin Both C++ and Java sources CHANG, C.C. and C.J. LIN, 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/cjlin/libsvm.
- SVMlight Thorsten Joachims Written in C JOACHIMS, T., 1999. SVMLight: Support Vector Machine. Vector Machine http://svmlight.joachims.org/
- MATLAB Support Vector Machine Toolbox Gavin Cawley MATLAB toolbox CAWLEY, (2000) http://theoval.sys.uea.ac.uk/gcc/svm/toolbox. MATLAB Support Vector Machine Toolbox.
- R Language Chang and Lin 2001 - Package e1071 provides an interface to libsvm
- Python's Scikit-Learn

Referencias

Asa Ben-Hur et al. "Support Vector Machines and Kernels for Computational Biology.". In: *PLoS Computational Biology* 4.10 (2008). URL: http://dblp.uni-trier.de/db/journals/ploscb/ploscb4.html#Ben-HurOSSR08.

Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. "A Training Algorithm for Optimal Margin Classifiers". In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. COLT '92. Pittsburgh, Pennsylvania, USA: ACM, 1992, pp. 144–152. ISBN: 0-89791-497-X. DOI: 10.1145/130385.130401. URL: http://doi.acm.org/10.1145/130385.130401.

Corinna Cortes and Vladimir Vapnik. "Support-vector networks". In: *Machine Learning* 20.3 (Sept. 1995), pp. 273–297. ISSN: 1573-0565. DOI: 10.1007/BF00994018. URL: https://doi.org/10.1007/BF00994018.

Andrew Ng. *CS229 Lecture notes, Part V Support Vector Machines*. URL: http://cs229.stanford.edu/notes/cs229-notes3.pdf.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001. ISBN: 0262194759.

Enrique J. Carmona Suárez. *Tutorial sobre Máquinas de Vectores Soporte (SVM)*. URL: http://www.ia.uned.es/~ejcarmona/publicaciones/[2013-Carmona]%20SVM.pdf.

V. Vapnik and A. Chervonenkis. "A note on one class of perceptrons". In: *Automation and Remote Control* 25 (1964).