Ayudantía Examen

Inteligencia Artificial 2017-2

Antonio Ossa (aaossa@uc.cl)

Informaciones varias

- Fecha / hora / sala examen
- Martes 28 de noviembre, 9:00. Mismas salas de las ies.

- N° de preguntas?
- 3 (1 de la primera parte, 2 de la segunda)

- Algo de las tareas e interrogaciones?
- Recorrecciones T1 listas, notas T2 en proceso, T3 para el día del examen. Recorrecciones de interrogaciones aun no actualizadas en el excel (pero ya hechas).

- Algo más?

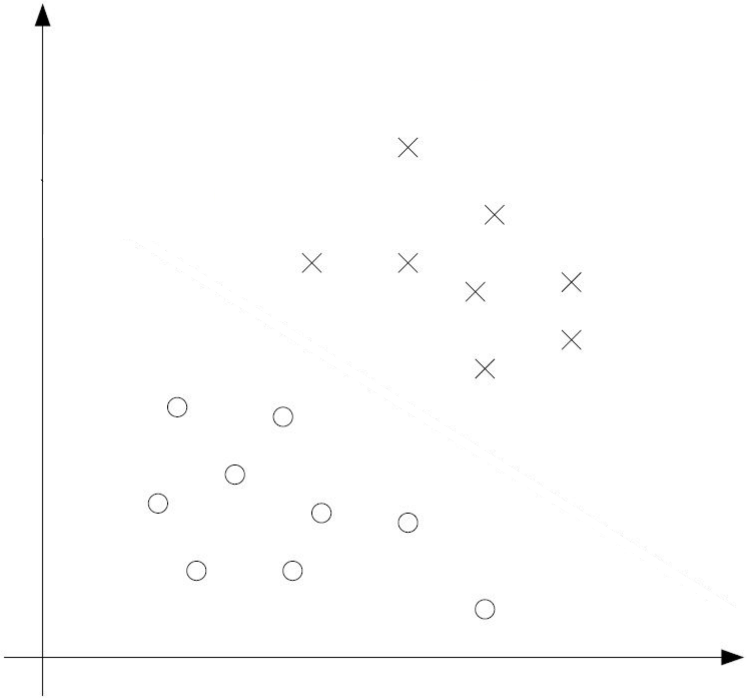
Contenidos a revisar hoy

- *Support Vector Machines*
  - Primal v/s dual
  - Kernel
  - *Soft-margin*
- Dudas sobre *SVM*
- Dudas sobre *Deep Learning*

Preparación para el examen:

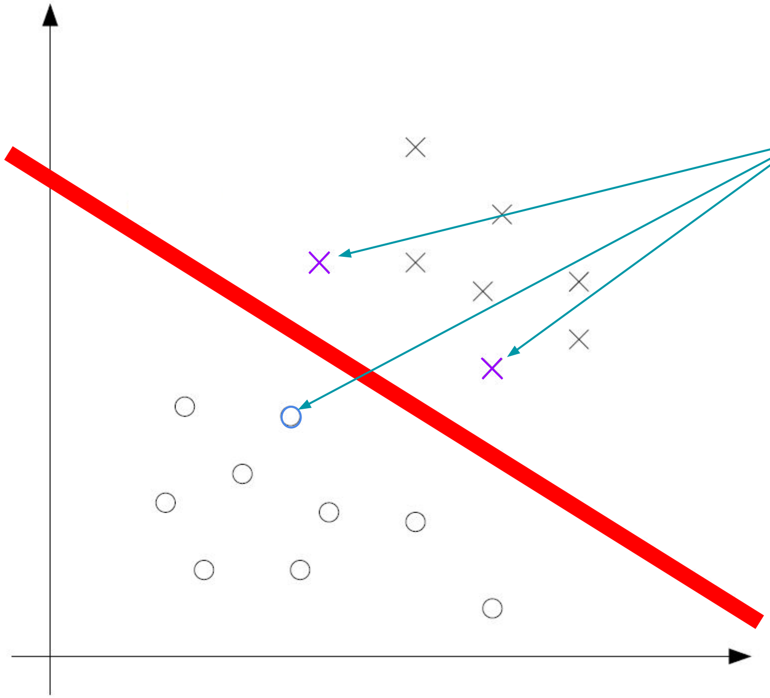
- Exámenes anteriores
  - Primera parte del curso
- Libros que aparecen en el programa
  - Segunda parte del curso
  - (+ libros que no aparecen)

SVM



Clases

SVM



Vectores de soporte:

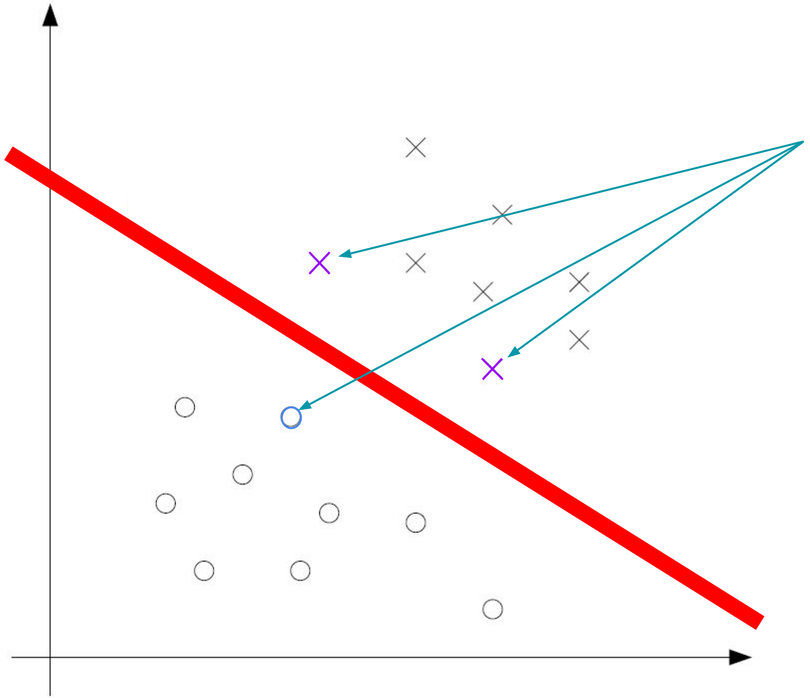
Puntos más cercanos al hiperplano

Hiperplano:

El mejor separador de los datos.

$\mathbf{w}^T x + \mathbf{b}$

SVM

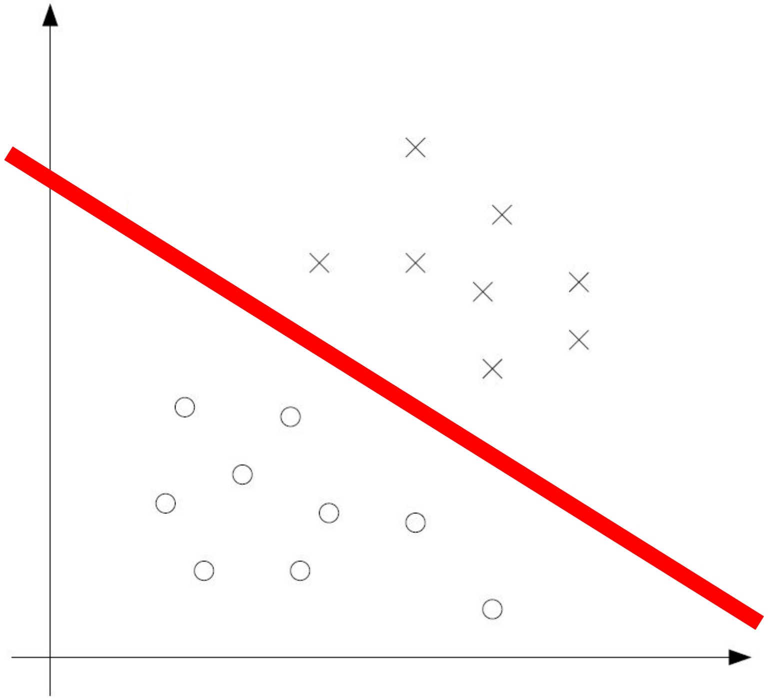


Vectores de soporte:
Puntos más cercanos al hiperplano

Este **hiperplano** es el que mejor separa los datos porque está lo más lejos posible de los **vectores de soporte**, o sea, es el que maximiza el **margen**

Hiperplano:
El mejor separador de los datos.
$\mathbf{w}^T x + \mathbf{b}$

SVM



$$h_{w,b}(x) = g(w^T x + b)$$

- $y \in \{-1, 1\}$ $\times$ $\bigcirc$
- $g(z) = 1$ if $z \geq 0$
  $g(z) = -1$ otherwise.

h: clasificador (output 1 o -1)
w: parámetros (se aprenden)
b: intercepto (se aprende)

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen funcional del dataset

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

Cada ejemplo tiene un margen funcional de al menos γ

Margen funcional igual a margen geométrico

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen funcional del dataset

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

Cada ejemplo tiene un margen funcional de al menos γ

Margen funcional igual a margen geométrico

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen relacionando parte geométrica y funcional

$$\max_{\hat{\gamma},w,b} \quad \frac{\hat{\gamma}}{||w||}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

Cada ejemplo tiene un margen funcional de al menos γ

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen relacionando parte geométrica y funcional

$$\max_{\hat{\gamma},w,b} \cancel{\frac{\hat{\gamma}}{||w||}}$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq \hat{\gamma}, \quad i = 1, \ldots, m$$

Cada ejemplo tiene un margen funcional de al menos γ

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen

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

Cada ejemplo tiene un margen funcional de al menos γ

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano

Margen

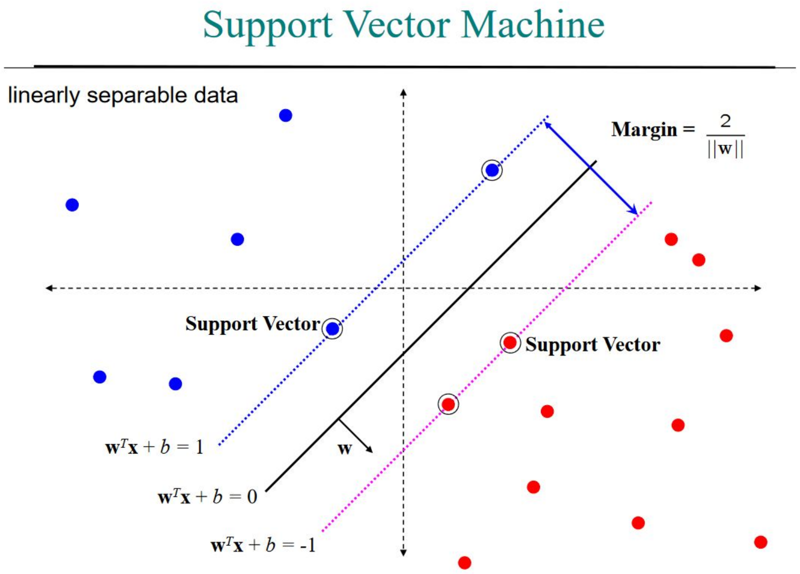
$$\min_{\gamma,w,b} \quad \frac{1}{2}||w||^2$$
$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \ldots, m$$

Cada ejemplo tiene un margen funcional de al menos γ

SVM como problema de optimización

max margen

s.a. que los puntos estén lejos del hiperplano



SVM: Por qué el dual?

- ¿Qué es el dual?
- ¿Por qué el primal no es suficiente?
- ¿Qué da el dual que no da el primal?
- ¿Qué son los alpha?

SVM: Por qué el dual?

- ¿Qué es el dual?
- ¿Por qué el primal no es suficiente?
- ¿Qué da el dual que no da el primal?
- ¿Qué son los alpha?

Resolviendo el primal obtenemos **w**, pero no sabemos nada sobre los **α**. Para clasificar un nuevo punto necesitamos calcular explicitamente el producto escalar $\mathbf{w}^\mathbf{T}\mathbf{x}$, que podría ser caro si el número de variables es grande.

Al resolver el dual, obtenemos los **α**, que son 0 para todos los puntos, **excepto para los vectores de soporte*.***

$$w^T x + w_0 = \left(\sum_{i=1}^{n} \alpha_i y_i x_i\right)^T x + w_0 = \sum_{i=1}^{n} \alpha_i y_i \langle x_i, x\rangle + w_0$$

SVM: Por qué el dual?

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

SVM: Por qué el dual?

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

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

SVM: Por qué el dual?

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

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

$$\nabla_w \mathcal{L}(w,b,\alpha) = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0$$

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

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

SVM: Por qué el dual?

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

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

SVM: Kernel?

Cuando nuestros datos no son linealmente separables, o sea, no podemos separarlos por un hiperplano, tenemos que “doblar” el espacio para poder trazar un hiperplano que separe los datos en esa representación.

$$K(x, z) = \phi(x)^T \phi(z)$$

SVM: Kernel?

Cuando nuestros datos no son linealmente separables, o sea, no podemos separarlos por un hiperplano, tenemos que “doblar” el espacio para poder trazar un hiperplano que separe los datos en esa representación.

$$K(x, z) = \phi(x)^T \phi(z)$$

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

SVM: Kernel?

Cuando nuestros datos no son linealmente separables, o sea, no podemos separarlos por un hiperplano, tenemos que “doblar” el espacio para poder trazar un hiperplano que separe los datos en esa representación.

$$K(x,z) = \phi(x)^T \phi(z)$$

$$K(x,z) = (x^T z)^2$$

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

SVM: Kernel?

Cuando nuestros datos no son linealmente separables, o sea, no podemos separarlos por un hiperplano, tenemos que "doblar" el espacio para poder trazar un hiperplano que separe los datos en esa representación.

$$K(x,z) = \phi(x)^T \phi(z)$$

$$\phi(x) = \begin{bmatrix} x_1x_1 \\ x_1x_2 \\ x_1x_3 \\ x_2x_1 \\ x_2x_2 \\ x_2x_3 \\ x_3x_1 \\ x_3x_2 \\ x_3x_3 \end{bmatrix}$$

SVM: Kernel?

Cuando nuestros datos no son linealmente separables, o sea, no podemos separarlos por un hiperplano, tenemos que “doblar” el espacio para poder trazar un hiperplano que separe los datos en esa representación.

$$K(x, z) = \phi(x)^T \phi(z)$$

- Polynomials of degree d

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v})^d$$

- Polynomials of degree up to d

$$K(\mathbf{u}, \mathbf{v}) = (\mathbf{u} \cdot \mathbf{v} + 1)^d$$

- Gaussian/Radial kernels (polynomials of all orders – recall series expansion)

$$K(\mathbf{u}, \mathbf{v}) = \exp\left(-\frac{||\mathbf{u} - \mathbf{v}||^2}{2\sigma^2}\right)$$

- Sigmoid

$$K(\mathbf{u}, \mathbf{v}) = \tanh(\eta \mathbf{u} \cdot \mathbf{v} + \nu)$$

SVM: *soft-margin?*

Es posible que nuestros datos no sean separables. Es algo que asumimos inicialmente. En este caso reformulamos la optimización (usando **regularización L1**) y permitimos que los datos no cumplan con el margen, pero esto penaliza nuestra solución.

$$
\begin{aligned}
\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\
\text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1,\ldots,m \\
& \xi_i \geq 0, \quad i = 1,\ldots,m.
\end{aligned}
$$

SVM: *soft-margin?*

Es posible que nuestros datos no sean separables. Es algo que asumimos inicialmente. En este caso reformulamos la optimización (usando **regularización L1**) y permitimos que los datos no cumplan con el margen, pero esto penaliza nuestra solución.

Penalización

$$\begin{aligned}\min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m.\end{aligned}$$

Margen “flexible”

SVM: *soft-margin?*

Es posible que nuestros datos no sean separables. Es algo que asumimos inicialmente. En este caso reformulamos la optimización (usando **regularización L1**) y permitimos que los datos no cumplan con el margen, pero esto penaliza nuestra solución.

Primal

$$\begin{aligned} \min_{\gamma,w,b} \quad & \frac{1}{2}||w||^2 + C\sum_{i=1}^{m}\xi_i \\ \text{s.t.} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m \\ & \xi_i \geq 0, \quad i = 1, \ldots, m. \end{aligned}$$

$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2}w^T w + C\sum_{i=1}^{m}\xi_i - \sum_{i=1}^{m}\alpha_i\left[y^{(i)}(x^T w + b) - 1 + \xi_i\right] - \sum_{i=1}^{m} r_i\xi_i$$

SVM: *soft-margin?*

Es posible que nuestros datos no sean separables. Es algo que asumimos inicialmente. En este caso reformulamos la optimización (usando **regularización L1**) y permitimos que los datos no cumplan con el margen, pero esto penaliza nuestra solución.

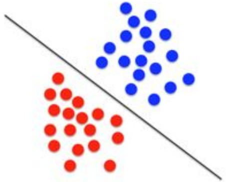
Dual

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

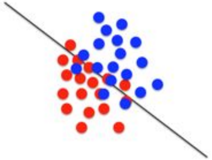
$$\mathcal{L}(w, b, \xi, \alpha, r) = \frac{1}{2} w^T w + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \left[y^{(i)} (x^T w + b) - 1 + \xi_i \right] - \sum_{i=1}^{m} r_i \xi_i$$

SVM: Resumen

1) Linear with perfect separation



2) Linear with no perfect separation



3) Non linear