Aprendizaje Supervisado

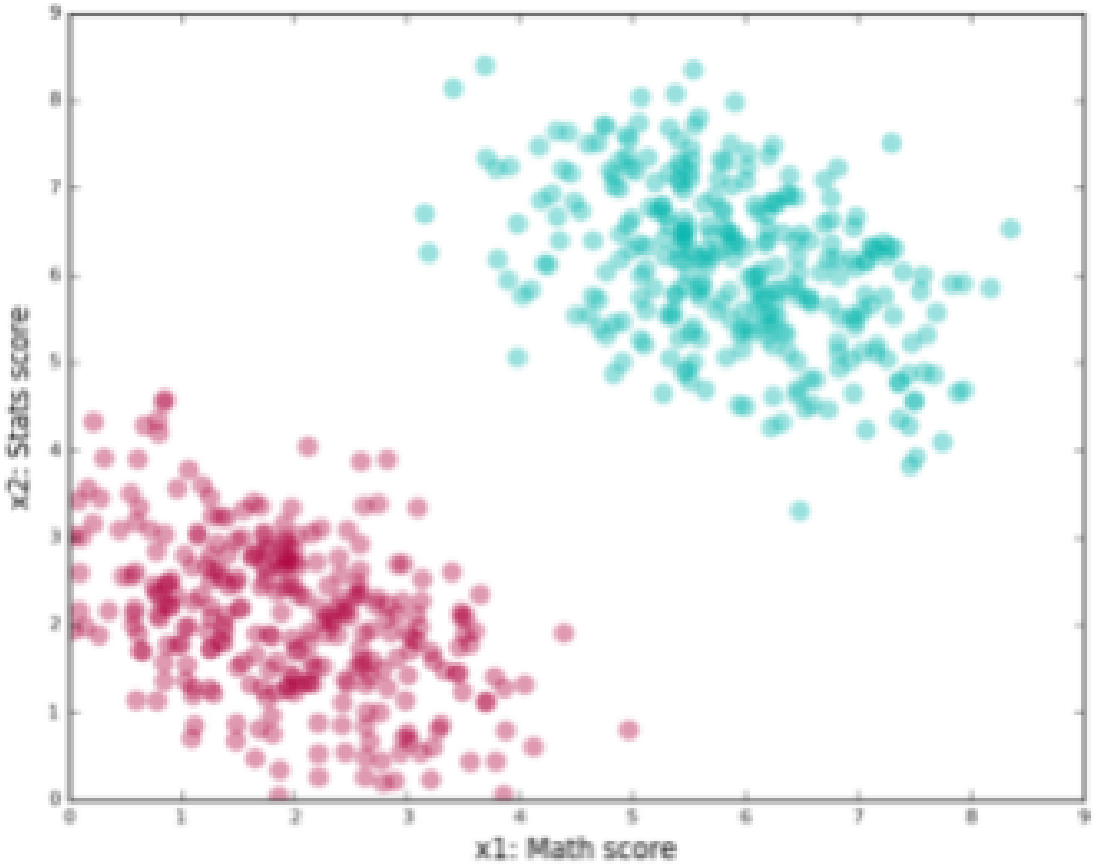
SVM

November 6, 2021

Resumen de clase

SVC

Problema: Clasificación (dos clases)



SVC

n datos $x_1, \ldots, x_n$ de p coordenadas con una respuesta $y_i \in \{1, -1\}$.
Tenemos 2 clases:

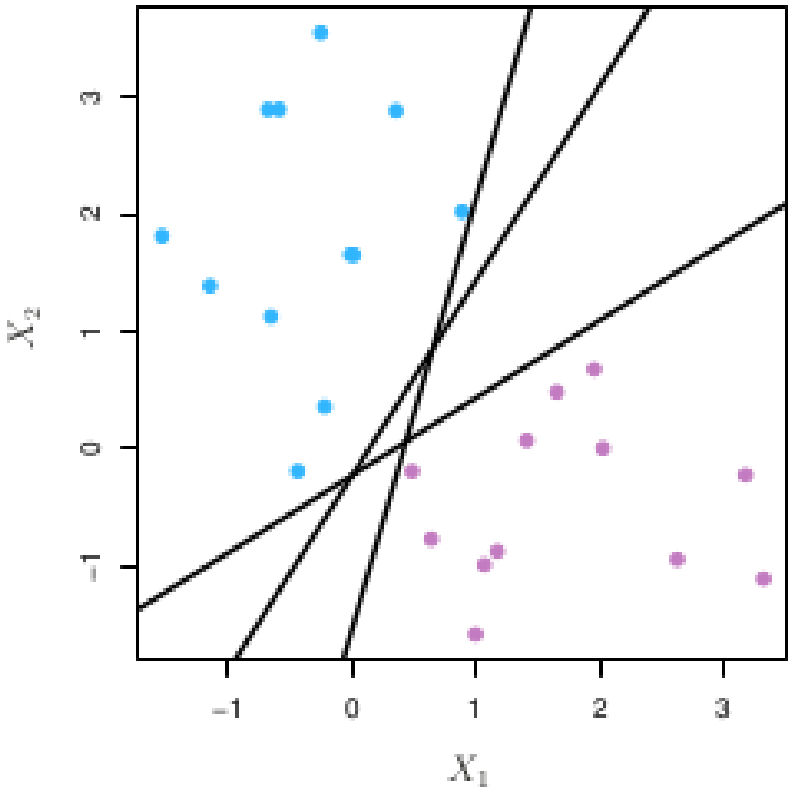
$$C_1 = \{y_i = 1\} \text{ y } C_2 = \{y_i = -1\}$$

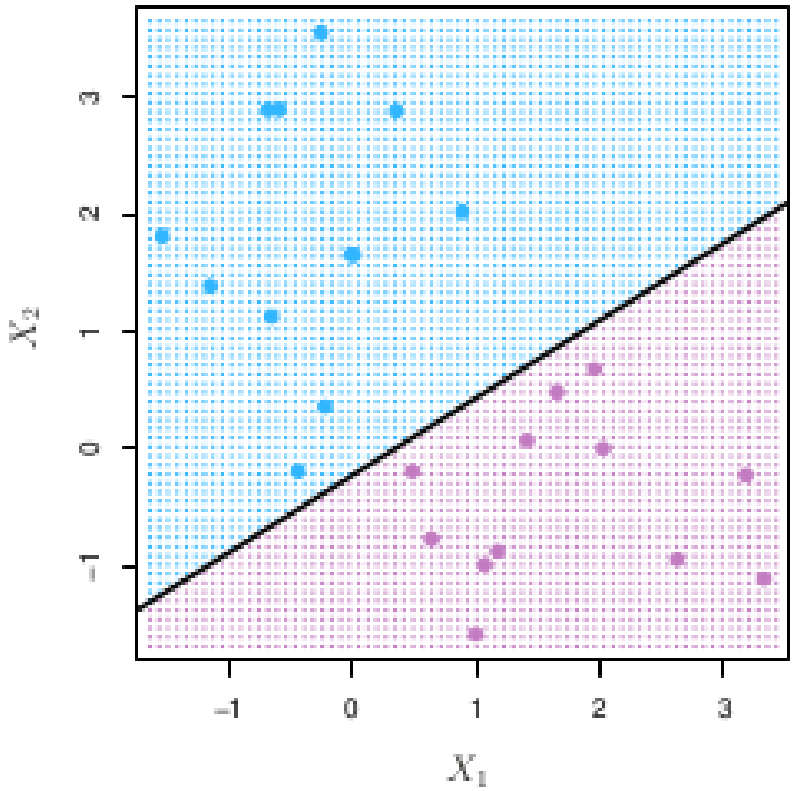
Supongamos que existe un hiperplano separador de las dos clases.

Recuerdo: un hiperplano es un "plano" en $\mathbb{R}^p$ determinado por una ecuación afin (lineal + constante):

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0$$

Hiperplano Separador





Hiperplano Separador

Pensamos $\mathbb{R}^p$ separado en dos "mitades" por el hiperplano. Asignando a

$$\begin{aligned} x \in C_1 \Leftrightarrow y = 1 \Leftrightarrow & \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p > 0 \\ \Leftrightarrow & y(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) > 0 \end{aligned}$$

$$\begin{aligned} x \in C_2 \Leftrightarrow y = -1 \Leftrightarrow & \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p < 0 \\ \Leftrightarrow & y(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p) > 0 \end{aligned}$$

O sea $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ y y tienen el mismo signo!

Hiperplano Separador

Si tal hiperplano existe clasificamos los puntos mirando a que mitad del espacio pertenece usando

$$signo(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)$$

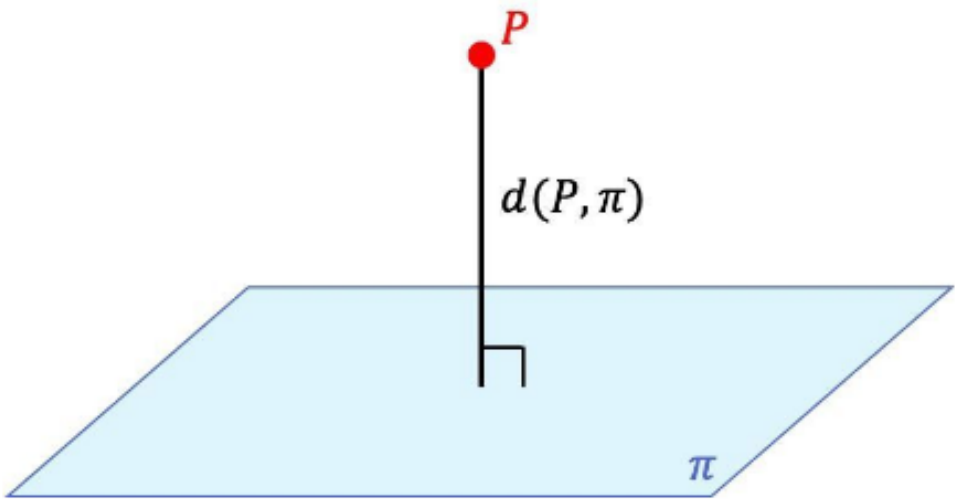
Buscamos (si existen) los coeficientes del hiperplano separador $\beta_0, \beta_1, \ldots, \beta_p$

Antes, un poco de cuentas...

Distancia de un punto a un hiperplano

Sea P un punto y π un hiperplano dado por la ecuación

$$\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = 0$$



$\beta := (\beta_1, \ldots, \beta_p)$ es normal a π. Entonces,

$$t \mapsto t\beta + P$$

es una recta perpendicular a π que pasa por P.
Intersección de π con la recta para hallar la proyección ortogonal: $t \in \mathbb{R}$ tal que $t\beta + P \in \pi$

Distancia de un punto a un hiperplano

Si $P = (x_1, \ldots, x_p)$ $t\beta + P \in \pi$:

$$\beta_0 + \sum_{i=1}^{p} \beta_i (t\beta_i + x_i) = 0$$

$$\beta_0 + \sum_{i=1}^{p} t\beta_i^2 + \sum_{i=1}^{p} \beta_i x_i = 0$$

$$\beta_0 + t\|\beta\|^2 + \beta.P = 0$$

$$t = -\frac{\beta_0 + \beta.P}{\|\beta\|^2}$$

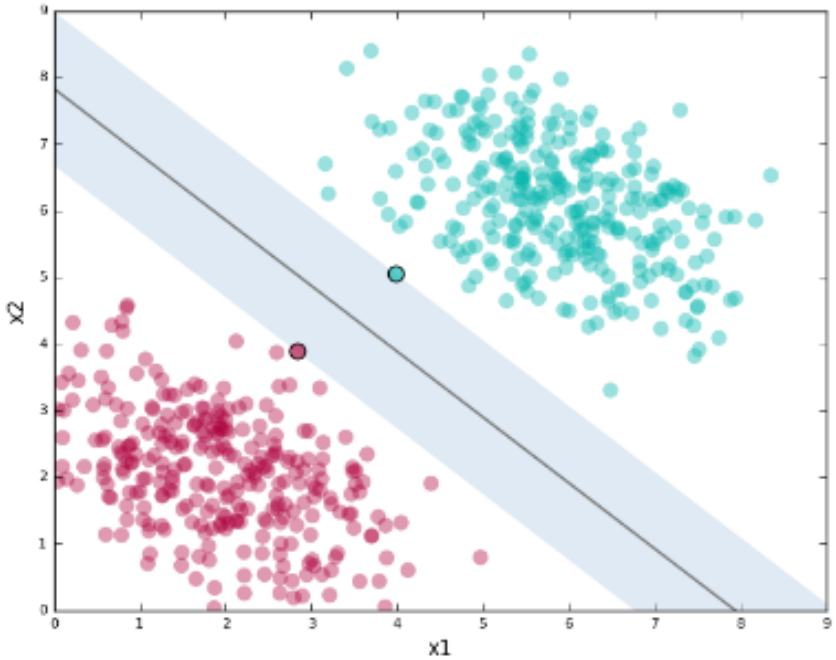
Luego,

$$d(t\beta + P, P) = \|t\beta + P - P\| = |t|\|\beta\| = \frac{|\beta_0 + \beta.P|}{\|\beta\|}$$

$$= \frac{1}{\|\beta\|}|\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p|$$

Clasificador de margen maximal

Si $\|\beta\| = 1$, el módulo de la ecuación del hiperplano determina la distancia de un punto a π.



Signo y nivel:

$$f(x) = f(x_1, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

cerca estamos del hiperplano, más información además de la clasificación.

Clasificador de margen maximal

Resolvemos el siguiente problema de optimización con restricciones:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,M} M$$

restringido a

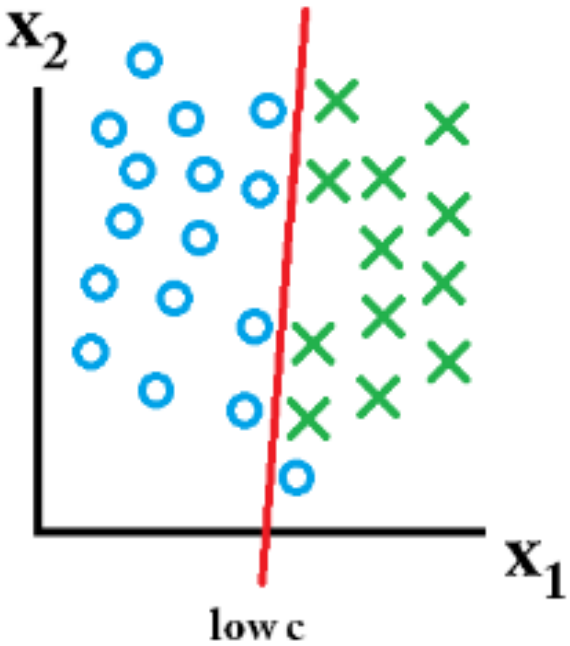
$$\sum_{j=1}^{p} \beta_j^2 = 1$$

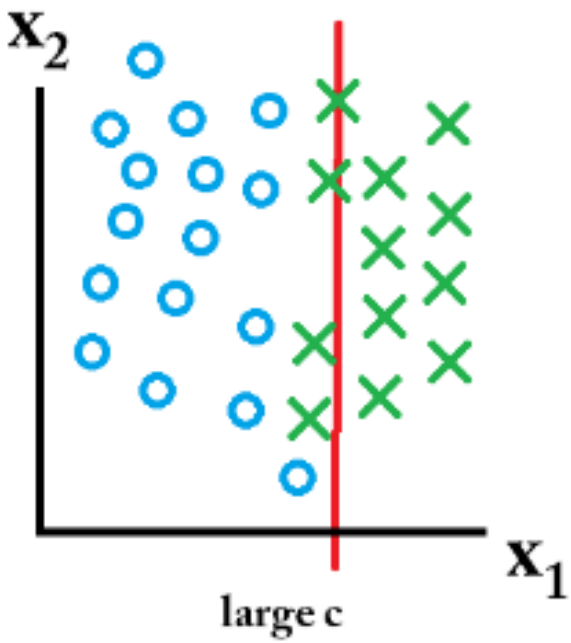
$$y_i(\beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i) \geq M \text{ para todo i}$$

- El problema tiene solución para $M > 0$ si existe un hiperplano separador.
- El valor óptimo hallado M es la medida del margen de la banda.

SVC

¿Si los datos no son separables mediante un hiperplano?





SVC

Modificamos el problema de optimización agegando variables que nos permitan incumplir las condiciones:

$$y_i(\beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i) \geq M(1 - \varepsilon_i)$$

- $\varepsilon_i = 0$: clasifico correctamente fuera del margen
- $0 < \varepsilon_i < 1$: clasifico correctamente pero dentro del margen
- $1 < \varepsilon_i$: permitimos clasificar incorrectamente

SVC

Para tener un control sobre la cantidad de error permitido agregamos un hiperparámetro C costo:

$$\sum_i \varepsilon_i \leq C$$

con $\varepsilon_i \geq 0$ para todo i.

Si $C = 0$, entonces $\varepsilon_i = 0$ para todo i. Estamos en el problema anterior sin modificaciones.

Si $C \leq 1$, entonces $\varepsilon_i < 1$ para todo i. Clasifico correctamente pero el margen es menor a M.

SVC

Resolvemos el siguiente problema de optimización con restricciones. Dado $C \geq 0$:

$$\max_{\beta_0,\beta_1,\ldots,\beta_p,\varepsilon_1,\ldots,\varepsilon_n,M} M$$

restringido a

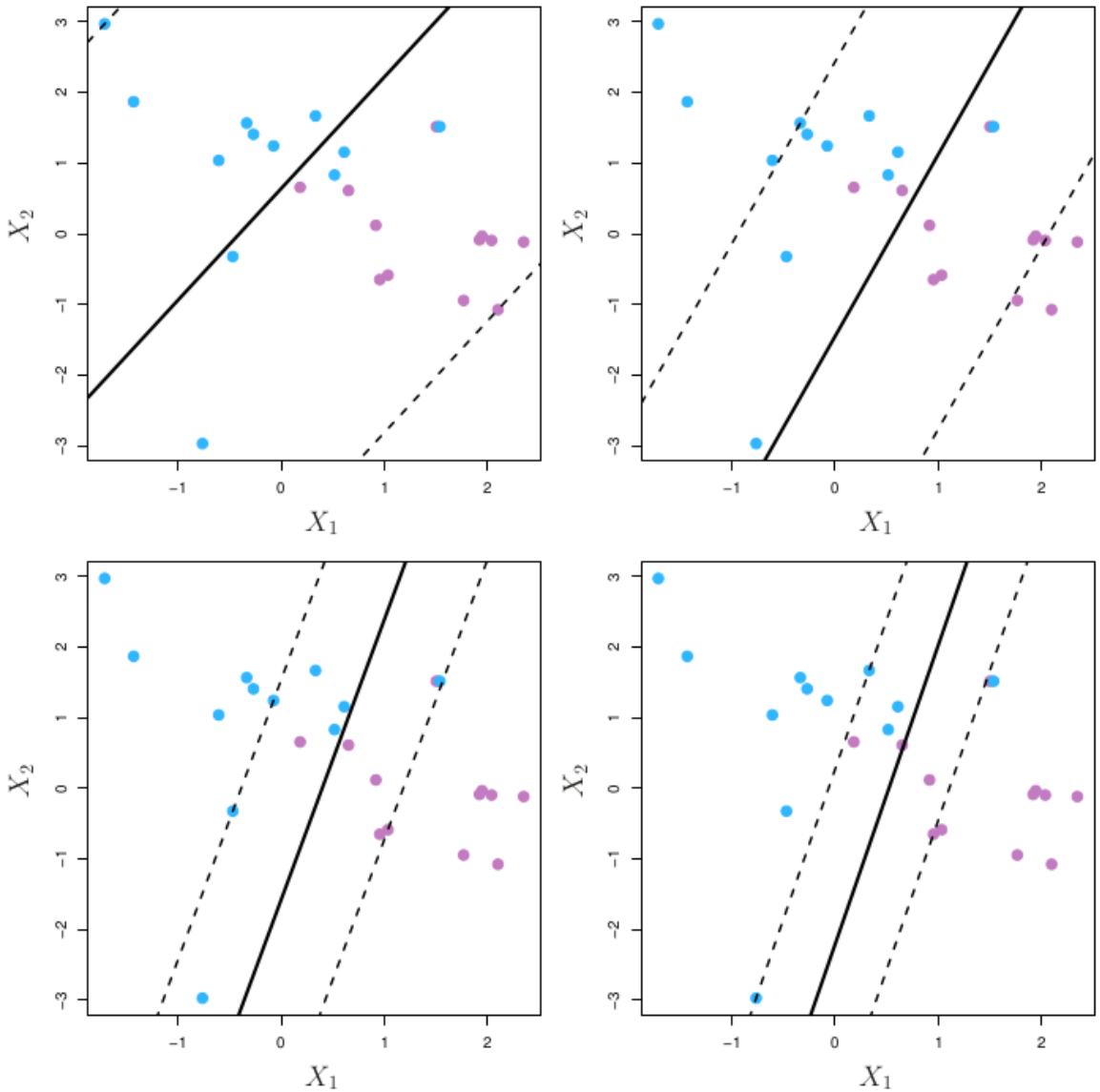
$$\sum_{j=1}^{p} \beta_j^2 = 1$$

$$y_i(\beta_0 + \beta_1 x_1^i + \cdots + \beta_p x_p^i) \geq M(1 - \varepsilon_i) \text{ para todo i}$$

$$\sum_{i=1}^{n} \varepsilon_i \leq C$$

$$\varepsilon_i \geq 0$$

Hiperparametro C



SVM

Otra forma de expresar la ecuación del hiperplano en términos de las observaciones $x^i = (x_1^i, \ldots, x_p^i)$. Para $i = 1, \ldots, n$

$$\begin{aligned} g(x) &= \beta_0 + \sum_{i=1}^{n} \alpha_i x.x^i \\ &= \beta_0 + \sum_{i=1}^{n} \alpha_i \left(\sum_{j=1}^{p} x_j x_j^i \right) \\ &= \beta_0 + \sum_{j=1}^{p} \left(\sum_{i=1}^{n} \alpha_i x_j^i \right) x_j \end{aligned}$$

Queremos $f = g$: buscamos α_i, $i = 1, \ldots, n$ tal que para todo $j = 1, \ldots, p$

$$\beta_j = \sum_{i=1}^{n} \alpha_i x_j^i$$

Obtenemos un sistema lineal de p ecuaciones y n incognitas.

Tenemos una escritura en función del producto interno contra las observaciones:

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i x.x^i$$

Para obtener los α_i alcanza con computar $x^j.x^i$ los productos internos entre las observaciones.

Si notamos $K(x, x^i) = x^j.x^i$, obtenemos

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x^i)$$

SVM

Podemos entonces generalizar el clasificador cambiando el producto interno por otras funciones **kernel** K que cuantifiquen similaridad entre dos observaciones.

- Kernel lineal:

$$K(x^j, x^i) = x^j.x^i = \sum_{k=1}^{p} x_k^j x_k^i$$

- Kernel polinomial grado d:

$$K_d(x^j, x^i) = (1 + x^j.x^i)^d$$

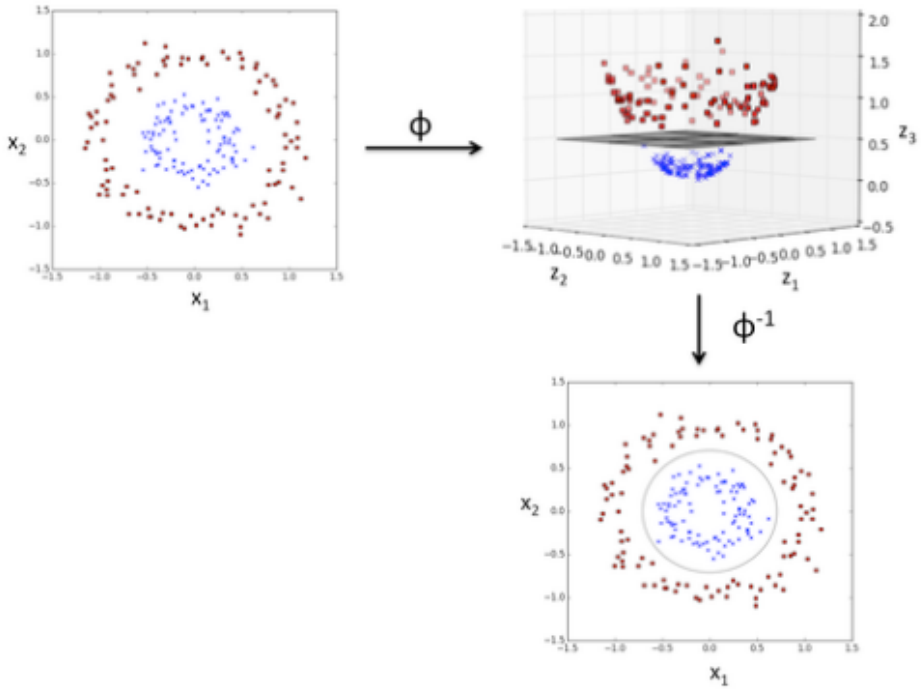
- Kernel radial:

$$K_\gamma(x^j, x^i) = \exp(-\gamma \sum_{k=1}^{p} (x_k^j - x_k^i)^2)$$

¿Qué rol cumple K?

$$f(x) = \beta_0 + \sum_{i=1}^{n} \alpha_i K(x, x^i)$$

1. Proyectamos a mayor dimensión
2. SVC separamos linealmente
3. Volvemos con la proyección



SVM polinomial

Kernel polinomial grado d:

$$K_d(x^j, x^i) = (1 + x^j.x^i)^d$$

- $d = 1$ obtenemos SVC
- $d > 1$ fiteamos un modelo SVC en el espacio de atributos de hasta dimensión d.
- obtenemos frontera de clasificación no lineal
- Ventaja sobre ir directamente al espacio de atributos polinomiales es que tenemos menos parámetros en el modelo y solamente necesitamos $K(x^j, x^i)$ para obtener f

SVM radial

Kernel radial:

$$K_\gamma(x^j, x^i) = \exp(-\gamma \|x^j - x^i\|^2)$$

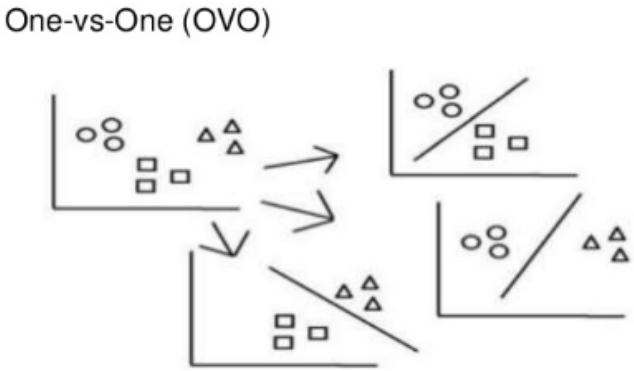
- exp en serie de Taylor (polinomio de grado infinito convergente)
- γ hiperparámetro del modelo
- K tiene comportamiento local: si una obs x^* está lejos del conjunto de entrenamiento $\|x^* - x^i\|^2$ es grande y luego $K(x^*, x^i)$ es chico. Es decir, x^i no tiene gran influencia en la clase de x^*.

SVM multiclase

$K > 2$ clases y se quiere aplicar el método de clasificación basado en SVM.

- One versus one
- One versus all
- DAGSVM

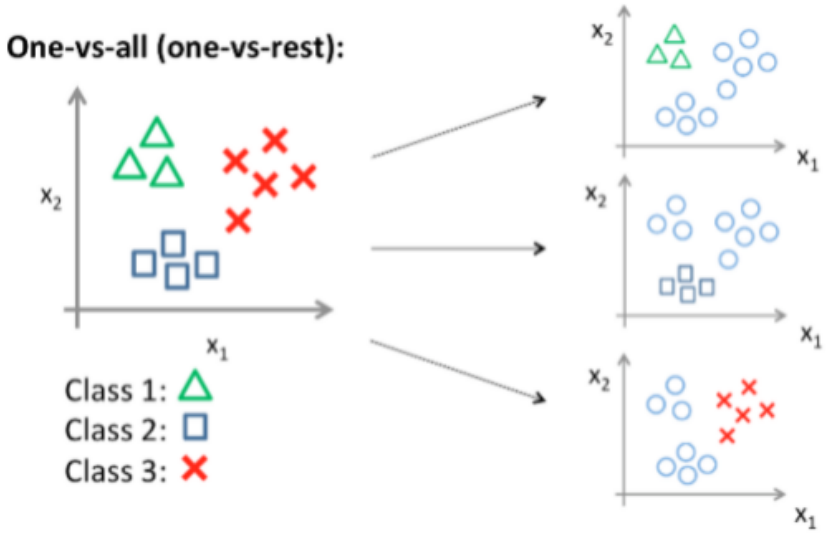
One versus one



1. Generar un total de $\frac{K(K-1)}{2}$ SVMs, comparando todos los posibles pares de clases.
2. Registramos el número de veces que la observación es asignada a cada una de las clases.
3. Clasificamos la observación a la clase a la que ha sido asignada con más frecuencia.

Desventaja:el número de modelos se dispara con K / no es aplicable en todos los escenarios.

One versus all



1. Generar un total de K SVMs, comparando cada clase frente a las $K-1$ clases restantes.
2. Clasificamos x^* a la clase que maximice $|f_k(x^*)|$

Desventaja: cada clasificador se entrena de forma no balanceada. Por ejemplo, si el conjunto de datos contiene 100 clases con 10 observaciones por clase, cada clasificador se ajusta con 10 observaciones positivas y 990 negativas.

DAGSVM

DAGSVM (Directed Acyclic Graph SVM) mejora del método one-versus-one. Datos con cuatro clases (A, B, C, D) y 6 clasificadores entrenados con cada posible par de clases (A-B, A-C, A-D, B-C B-D, C-D).

1. Se inician las comparaciones con el clasificador (A-D) y se obtiene como resultado que la observación pertenece a la clase A, o lo que es equivalente, que no pertenece a la clase D.
2. Excluimos todas las comparaciones que contengan la clase D, puesto que se sabe que no pertenece a este grupo.
3. En la siguiente comparación se emplea el clasificador (A-C) y se predice que es A. Con esta nueva información se excluyen todas las comparaciones que contengan C.
4. Finalmente solo queda emplear el clasificador (A-B) y asignar la observación al resultado devuelto.

Solo ha sido necesario emplear 3. DAGSVM tiene las mismas ventajas que el método one-versus-one pero mejorando mucho el rendimiento.

Ejercicios

1. ISLR labs 9.6
2. Notebook: https://colab.research.google.com/drive/1ReO_jbtOK3Y3aA1YinUaKOZKoJYOpCHJ?usp=sharing