



UNIVERSIDAD DE CHILE

# Minería de Datos

Welcome to the Machine Learning class

---

Valentin Barriere

Universidad de Chile – DCC

CC5205, Otoño 2024

# **Aprendizaje supervisado – SVM**

# Outline : Máquinas de Vectores de Soporte

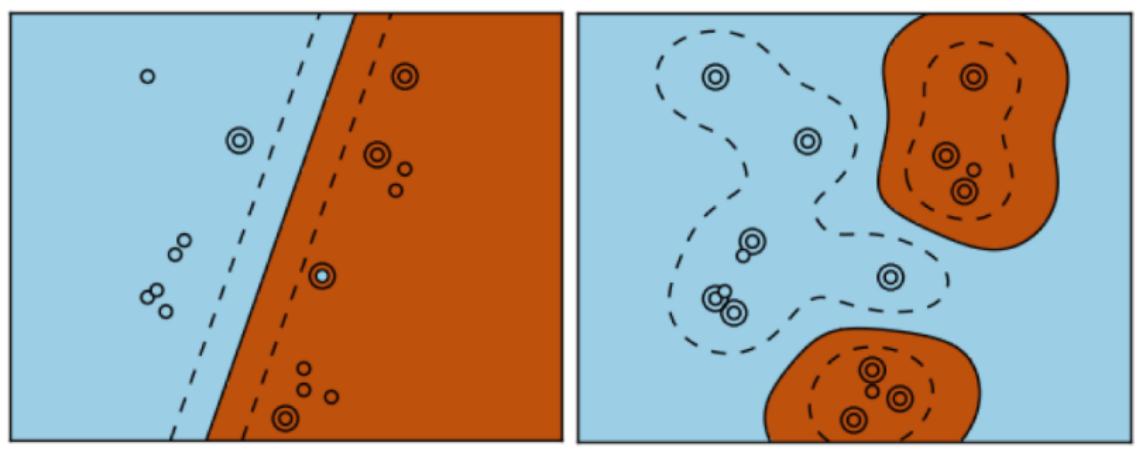
Máquinas de Vectores de Soporte

SVR

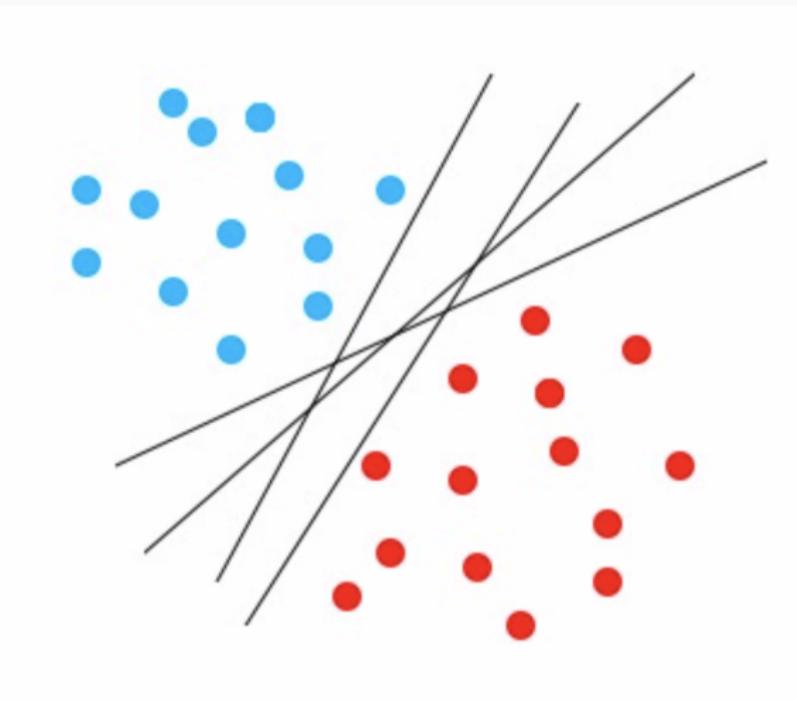
# Support Vector Machines – Separadores de Gran Márgen

## Máquinas de Vectores de Soporte

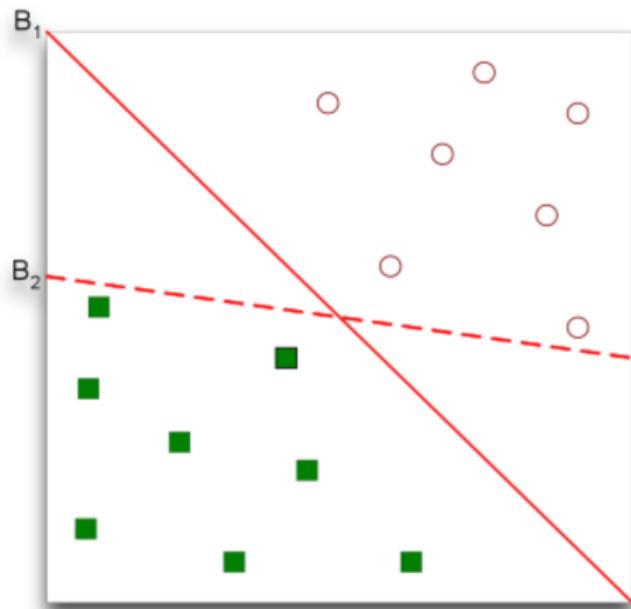
Buscan encontrar el margen que maximiza la separación entre las diferentes clases de datos.



## SVM lineal: ¿Qué plano elegir?

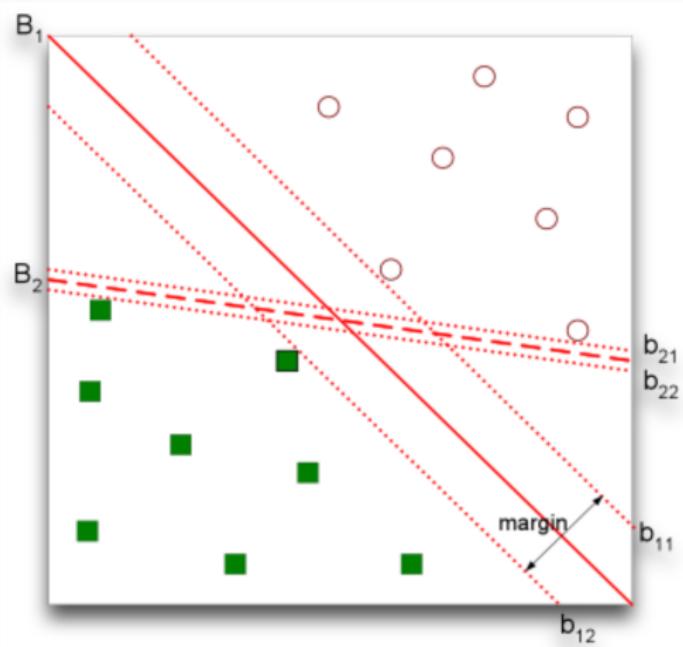


## SVM lineal: ¿Qué plano elegir?



- Cual es el mejor?  $B_1$  o  $B_2$ ?

## SVM lineal: ¿Qué plano elegir?

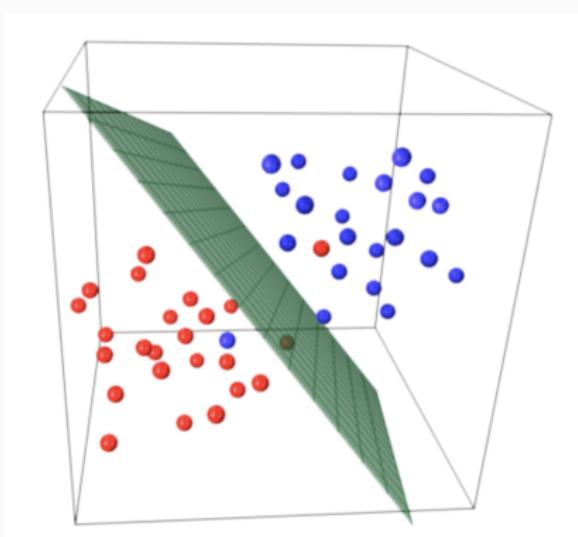


- Cual es el mejor?  $B_1$  o  $B_2$ ?  $\rightarrow B_1$
- Encontrar un hiperplano que **maximice el margen de entrenamiento** (menores errores de generalización, i.e., **menos específico a los datos**)

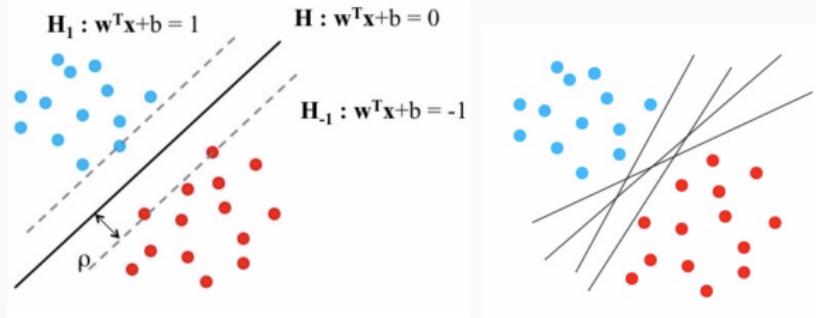
# Clasificador lineal: recordatorio

## Resumen

- $\mathbf{X} \in \mathbb{R}^d$  es el vector de descriptores.
- La ecuación  $\mathbf{W}^T \mathbf{X} + b = 0$  define un hiperplano en  $\mathbb{R}^d$ .
- $f_{\mathbf{W}, b}(\mathbf{X}) = \text{signo}(\mathbf{W}^T \mathbf{X} + b)$  da la clase de  $\mathbf{X}$ .



# SVM lineal

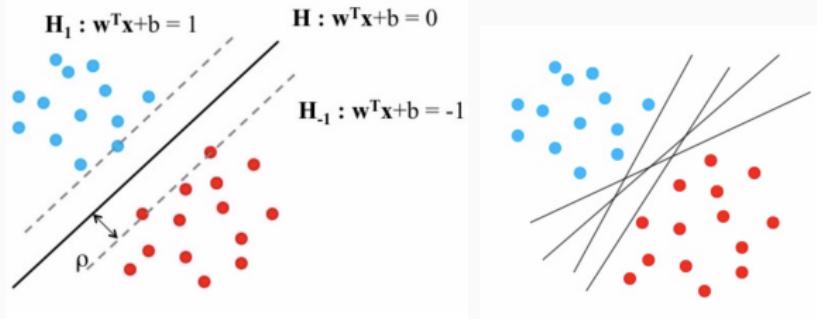


- Clasificador lineal:  $\text{signo}(\mathbf{W}^T \mathbf{X} + b)$ .
- Funciona en el **caso separable**:  $\exists(\mathbf{W}, b), \forall i, Y_i(\mathbf{W}^T \mathbf{X} + b) > 0$ .

## Cómo elegir $\mathbf{W}$ y $b$ para maximizar el margen $\rho$

- Separación estricta:  $\exists(\mathbf{W}, b), \forall i, Y_i(\mathbf{W}^T \mathbf{X} + b) \geq 1$ .
- Maximización de la distancia entre  $\mathbf{W}^T \mathbf{X} + b = 1$  y  $\mathbf{W}^T \mathbf{X} + b = -1$ .
- Equivale a minimizar  $\|\mathbf{W}\|$

# SVM lineal

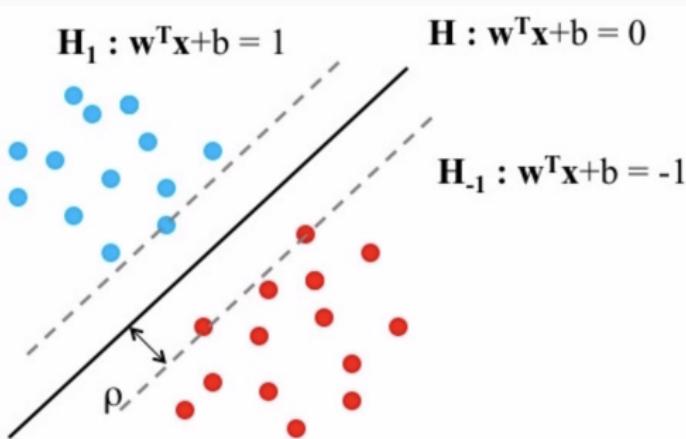


- Clasificador lineal:  $\text{signo}(\mathbf{W}^T \mathbf{X} + b)$ .
- Funciona en el **caso separable**:  $\exists(\mathbf{W}, b), \forall i, Y_i(\mathbf{W}^T \mathbf{X} + b) > 0$ .

## Cómo elegir $\mathbf{W}$ y $b$ para maximizar el margen $\rho$

- Separación estricta:  $\exists(\mathbf{W}, b), \forall i, Y_i(\mathbf{W}^T \mathbf{X} + b) \geq 1$ .
- Maximización de la distancia entre  $\mathbf{W}^T \mathbf{X} + b = 1$  y  $\mathbf{W}^T \mathbf{X} + b = -1$ .
- Equivale a minimizar  $\|\mathbf{W}\|$  **Pregunta → ¿Por qué?**

# SVM lineal: margen



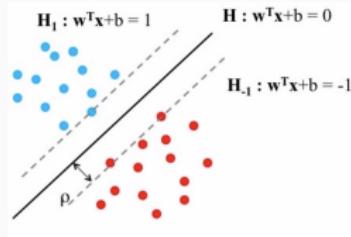
## Concepto de margen geométrico

- Para separar los datos, consideramos un trío de hiperplanos:

$$H : \mathbf{W}^T \mathbf{X} + b = 0, \quad H_1 : \mathbf{W}^T \mathbf{X} + b = 1, \quad H_{-1} : \mathbf{W}^T \mathbf{X} + b = -1,$$

- El margen geométrico  $\rho(\mathbf{W})$  es la distancia más pequeña entre los datos y  $H$ , es decir, la mitad de la distancia entre  $H_1$  y  $H_{-1}$ .
- Un cálculo simple da  $\rho(\mathbf{W}) = \frac{1}{\|\mathbf{W}\|}$ .

# SVM lineal



- La distancia  $\rho$  es igual a  $\frac{1}{\|w\|}$ : maximizar  $\rho$  es equivalente a minimizar  $\|w\|$ .
- **Atención:** siempre bajo la restricción de clasificar correctamente, ¡con un margen!

## Optimización: Problema primal

$$\underset{\mathbf{w}}{\text{minimizar}}$$

$$\frac{1}{2} \|\mathbf{w}\|^2$$

sujeto a

$$\forall i, Y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1$$

# Lagrangiano

## Problema primal de optimización

minimizar  
 $\mathbf{w}, b$

$$\frac{1}{2} \|\mathbf{W}\|^2$$

sujeto a

$$\forall i, Y_i(\mathbf{W}^T \mathbf{X}_i + b) \geq 1$$

## Solución: Método de los multiplicadores de Lagrange

El método de los multiplicadores de Lagrange permite encontrar los **puntos estacionarios** (máximo, mínimo...) de una función derivable de una o varias variables, bajo restricciones.

Problema del tipo:

$$(\mathcal{P}) : J(\mathbf{X}), \quad \text{bajo restricciones} \quad g_i(\mathbf{X}) \leq 0, \quad i = 1..p$$
$$\mathbf{x} \in \mathbb{R}^n$$

- Lagrangiano:  $\mathcal{L}(\mathbf{X}, \lambda) = J(\mathbf{X}) + \lambda^T g(\mathbf{X}), \lambda \geq 0, \lambda \in \mathbb{R}^q$
- $\lambda$  son los parámetros de Lagrange

# SVM: Lagrangiano aplicado al SVM I

## Lagrangiano

$$\mathcal{L}(\mathbf{X}, \alpha) = \frac{1}{2} \|\mathbf{W}\|^2 + \sum_i \alpha_i (1 - Y_i (\mathbf{W}^T \mathbf{X}_i + b))$$
$$\forall i, \quad \alpha_i \geq 0$$

## Condiciones de KKT (Karush-Kuhn-Tucker)

En el extremo (lo que buscamos, porque queremos el mínimo), tenemos:

$$\nabla_{\mathbf{W}} \mathcal{L} = 0 \quad \text{(estacionaridad)}$$

$$\nabla_b \mathcal{L} = 0 \quad \text{(estacionaridad)}$$

$$\forall i, \quad \alpha_i (1 - Y_i (\mathbf{W}^T \mathbf{X}_i + b)) = 0 \quad \text{(complementariedad)}$$

Nota: La complementariedad implica que si  $1 \neq Y_i (\mathbf{W}^T \mathbf{X}_i + b)$  entonces  $\alpha_i = 0$ . Esto significa que o bien  $\mathbf{X}_i$  está en uno de los planos de margen, o bien  $\alpha_i = 0$ .

# SVM: Lagrangiano aplicado al SVM I

## Lagrangiano

$$\mathcal{L}(\mathbf{X}, \alpha) = \frac{1}{2} \|\mathbf{W}\|^2 + \sum_i \alpha_i (1 - Y_i (\mathbf{W}^T \mathbf{X}_i + b))$$
$$\forall i, \quad \alpha_i \geq 0$$

## Condiciones de KKT (Karush-Kuhn-Tucker)

En el extremo (lo que buscamos, porque queremos el mínimo), tenemos:

$$\mathbf{W} = \sum_i \alpha_i Y_i \mathbf{X}_i \quad (\text{estacionaridad})$$

$$\sum_i \alpha_i Y_i = 0 \quad (\text{estacionaridad})$$

$$\forall i, \quad \alpha_i (1 - Y_i (\mathbf{W}^T \mathbf{X}_i + b)) = 0 \quad (\text{complementariedad})$$

Nota: La complementariedad implica que si  $1 \neq Y_i (\mathbf{W}^T \mathbf{X}_i + b)$  entonces  $\alpha_i = 0$ . Esto significa que o bien  $\mathbf{X}_i$  está en uno de los planos de margen, o bien  $\alpha_i = 0$ .

# SVM: Lagrangiano aplicado al SVM II

## Condiciones de KKT (Karush-Kuhn-Tucker)

Significa que:

- O bien  $\mathbf{X}_i$  está en uno de los planos de margen, o bien  $\alpha_i = 0$
- Hay al final pocos multiplicadores de Lagrange  $\alpha_k$  que no son cero
- Estos  $\mathbf{X}_k$  se llaman **vectores de soporte**
- Los parámetros  $\mathbf{W}$  y  $b$ , que definen el límite de decisión, dependen sólo de los vectores de soporte

# SVM: Lagrangiano y dual

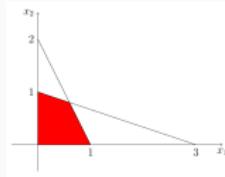
## Solución del problema dual

Mucho mas práctico de encontrar la solución del dual que del primal:

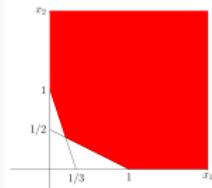
$$\underset{\alpha}{\text{Maximizar}} \quad \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_j y_i \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{sujeto a} \quad \forall i, \alpha_i \geq 0, \text{ ademas} \sum_i \alpha_i y_i = 0$$

$$\begin{aligned} \min \quad & z = c^t x, \\ \text{sujeto a} \quad & Ax \geq b, \\ & x \geq 0. \end{aligned}$$



$$\begin{aligned} \max \quad & z = b^t y, \\ \text{sujeto a} \quad & A^t y \leq c, \\ & y \geq 0. \end{aligned}$$



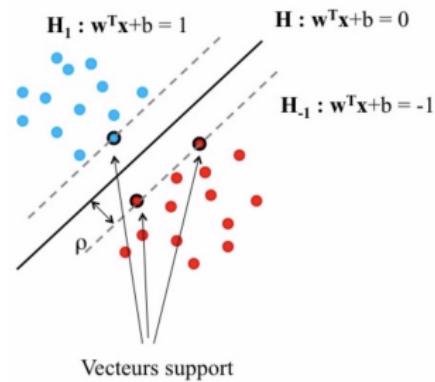
- El Lagrangiano dual involucra sólo los multiplicadores de Lagrange y los datos de entrenamiento (depende de menos parámetros).
- Mientras que el Lagrangiano primal involucra los multiplicadores de Lagrange así como los parámetros del hiperplano separador  $\mathbf{W}$  y  $b$ .
- El problema de **minimización** original para el Lagrangiano primal  $\mathcal{L}$  se convierte en un problema de **maximización** para el Lagrangiano dual  $\mathcal{L}_d(\mathbf{X}, \alpha)$ .
- Como este es un problema de optimización convexo las soluciones para ambos problemas de optimización son **equivalentes** (siempre y cuando nuestra solución satisfaga las condiciones KKT).

# SVM lineal: Vectores de soporte

Los  $\alpha_i$  siendo determinados, podemos reescribir la función del clasificador:

$$f(\mathbf{X}) = \text{signo}\left(\sum_{i=1}^n \alpha_i Y_i \mathbf{X}_i^T \mathbf{X} + b\right)$$

porque en el óptimo  $\mathbf{W} = \sum_{i=1}^n \alpha_i Y_i \mathbf{X}_i$  (KKT - estacionaridad)

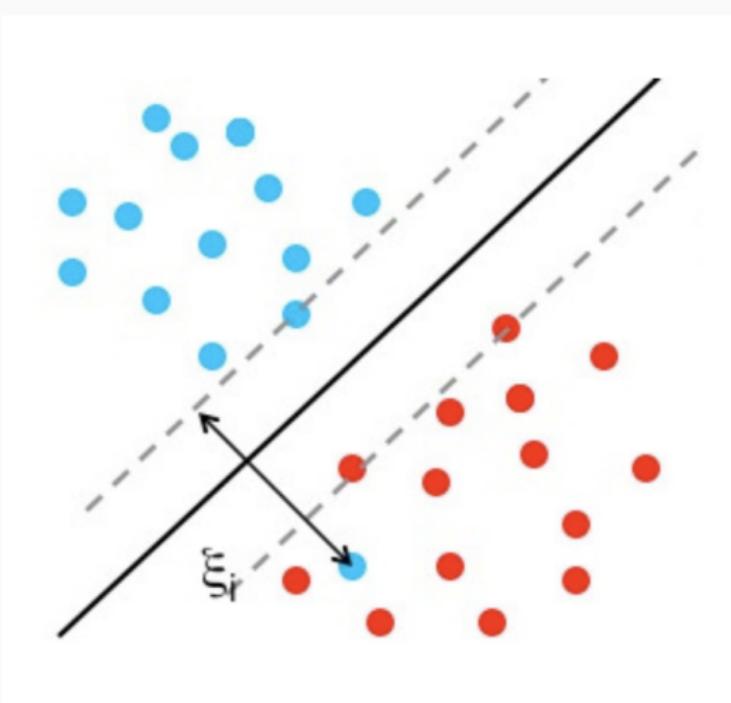


## Vectores de soporte

Dado que o bien  $\alpha_i$  es nulo, o bien el vector  $\mathbf{X}_i$  está en uno de los planos, llamamos a estos vectores los soportes.

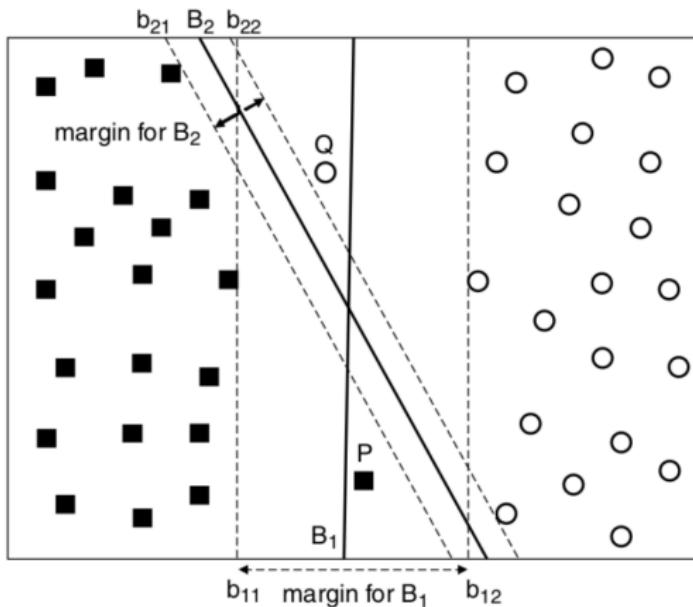
## SVM lineal: datos no separables

Funciona también cuando es ruidoso, pq la primera formulación no sirve:



## SVM lineal: datos no separables o ruidosos

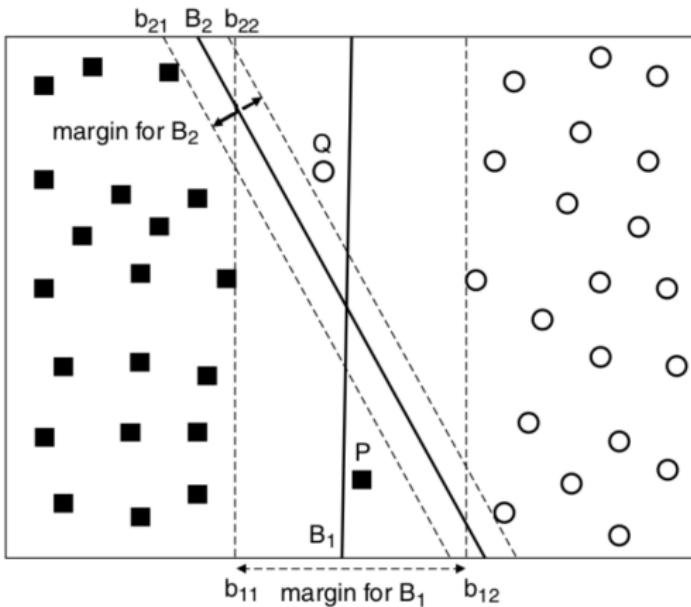
Funciona también cuando es ruidoso, pq la primera formulación no sirve:



Acá tenemos dos hiperplanos separadores posibles  $B_1$  y  $B_2$

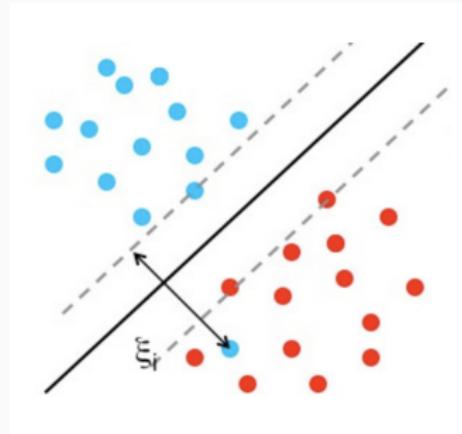
## SVM lineal: datos no separables o ruidosos

Funciona también cuando es ruidoso, pq la primera formulación no sirve:



Acá tenemos dos hiperplanos separadores posibles  $B_1$  y  $B_2 \rightarrow B_1$  es mejor!

# SVM lineal: soft margin



## Introducción de $\xi_i$ – Problema en el primal

minimizar  
 $\mathbf{W}, b$

sujeto a

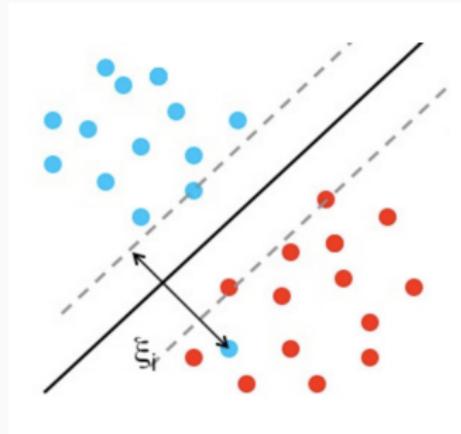
$$\frac{1}{2} \|\mathbf{W}\|^2 + C \sum_i \xi_i$$

$$\forall i, Y_i(\mathbf{W}^T \mathbf{X}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Rq:  $C$  equivale a un parámetro de regularización

# SVM lineal: soft margin



## Introducción de $\xi_i$ – Problema en el primal

minimizar  
 $\mathbf{W}, b$

sujeto a

$$\frac{1}{2} \|\mathbf{W}\|^2 + C \sum_i \xi_i$$

$$\forall i, Y_i(\mathbf{W}^T \mathbf{X}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

Rq:  $C$  equivale a un parámetro de regularización

# SVM lineal: soft margin y dual

Después de las condiciones de KKT, obtenemos:

## Solución del problema dual

Maximizar <sup>$\alpha$</sup>

$$\sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_j y_i \mathbf{x}_i^T \mathbf{x}_j$$

sujeto a

$$\forall i, 0 \leq \alpha_i \leq C, \text{ ademas } \sum_i \alpha_i y_i = 0$$

Notas:

- Algunos datos de soporte pueden estar en el otro lado de los planos  $H_1$  o  $H_{-1}$
- $C$  es un hiperparámetro que controla el compromiso entre la complejidad del modelo y el número de errores de clasificación del modelo.

# SVM lineal: enfoque con regularización

## Optimización en el espacio primal

$$\min_{\mathbf{W}, b} \sum_{i=1}^n \max(1 - Y_i(\mathbf{W}^T \mathbf{X}_i + b), 0) + \lambda \|\mathbf{W}\|^2$$

- $f(\mathbf{X}) = \text{signo}(\mathbf{W}^T \mathbf{X} + b)$
- Pérdida (bisagra) :  $\ell(\mathbf{X}, Y) = \max(1 - Y_i(\mathbf{W}^T \mathbf{X} + b), 0)$
- $\lambda$  es el parámetro de regularización
- $Y_i(\mathbf{W}^T \mathbf{X} + b)$  se llama margen del clasificador

# SVM lineal: enfoque con regularización

## Optimización en el espacio primal

$$\min_{\mathbf{W}, b} \sum_{i=1}^n \max(1 - Y_i(\mathbf{W}^T \mathbf{X}_i + b), 0) + \lambda \|\mathbf{W}\|^2$$

- $f(\mathbf{X}) = \text{signo}(\mathbf{W}^T \mathbf{X} + b)$
- Pérdida (bisagra) :  $\ell(\mathbf{X}, Y) = \max(1 - Y_i(\mathbf{W}^T \mathbf{X} + b), 0)$
- $\lambda$  es el parámetro de regularización
- $Y_i(\mathbf{W}^T \mathbf{X} + b)$  se llama margen del clasificador

**Intuición:** La función de perdida va a poner un costo cada vez que un ejemplo sea a -1 de distancia del hiperplano de separación, forzando los ejemplos a estar (i) del buen lado, (ii) mas lejos del HP posible, (iii) minimizando la norma del HP.

# SVM lineal: enfoque con regularización

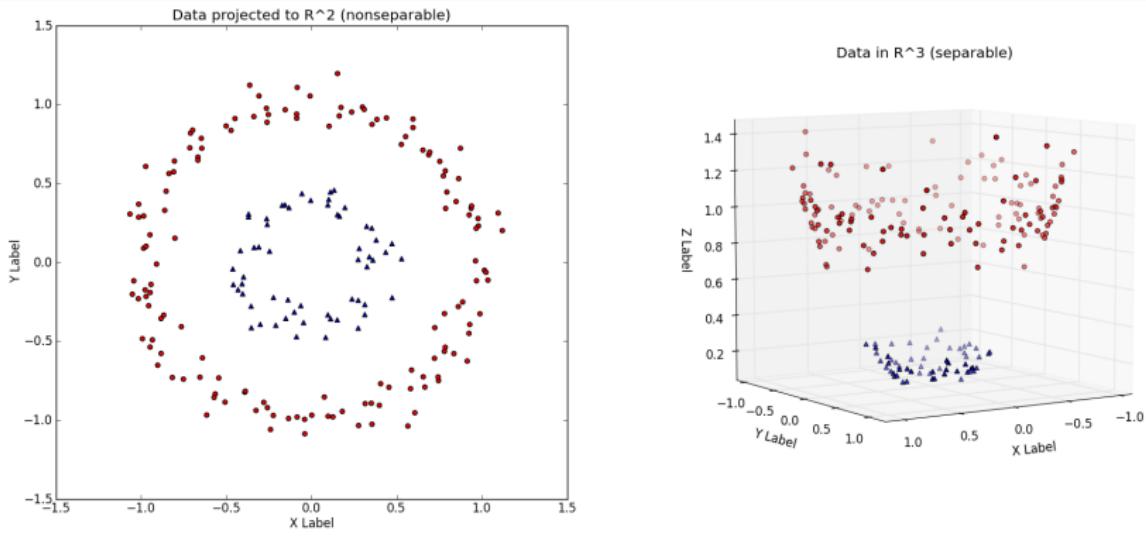
## Optimización en el espacio primal

$$\min_{\theta} \sum_{i=1}^n \ell(Y_i, f_{\theta}(\mathbf{X}_i)) + pen(\theta)$$

- $f(\mathbf{X}) = \text{signo}(\mathbf{W}^T \mathbf{X} + b)$
- Pérdida (bisagra) :  $\ell(\mathbf{X}, Y) = \max(1 - Y_i(\mathbf{W}^T \mathbf{X} + b), 0)$
- $\lambda$  es el parámetro de regularización
- $Y_i(\mathbf{W}^T \mathbf{X} + b)$  se llama margen del clasificador

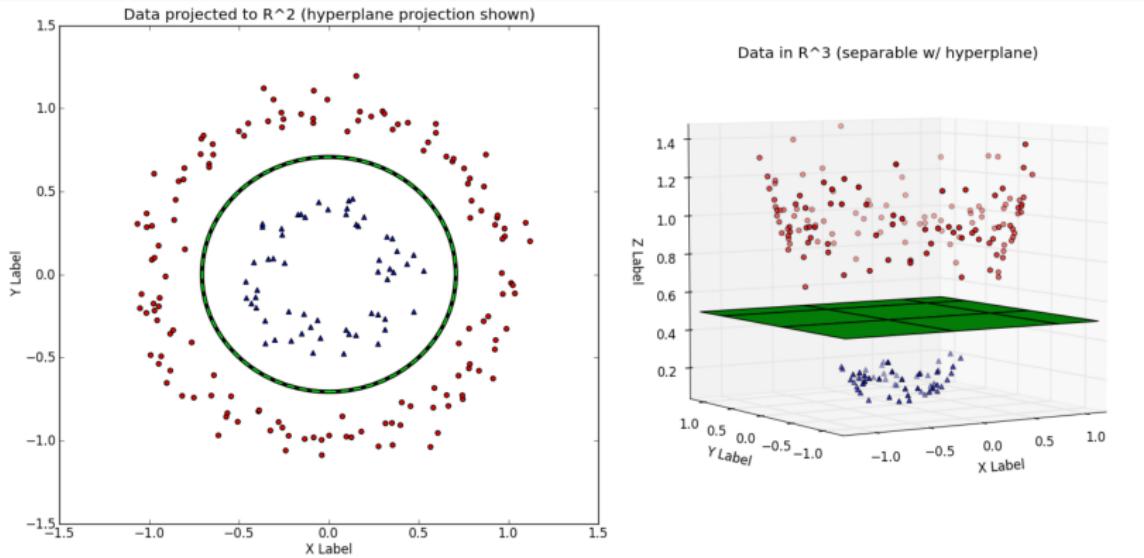
**Intuición:** La función de perdida va a poner un costo cada vez que un ejemplo sea a -1 de distancia del hiperplano de separación, forzando los ejemplos a estar (i) del buen lado, (ii) mas lejos del HP posible, (iii) minimizando la norma del HP.

# SVM con kernel: aumento del espacio



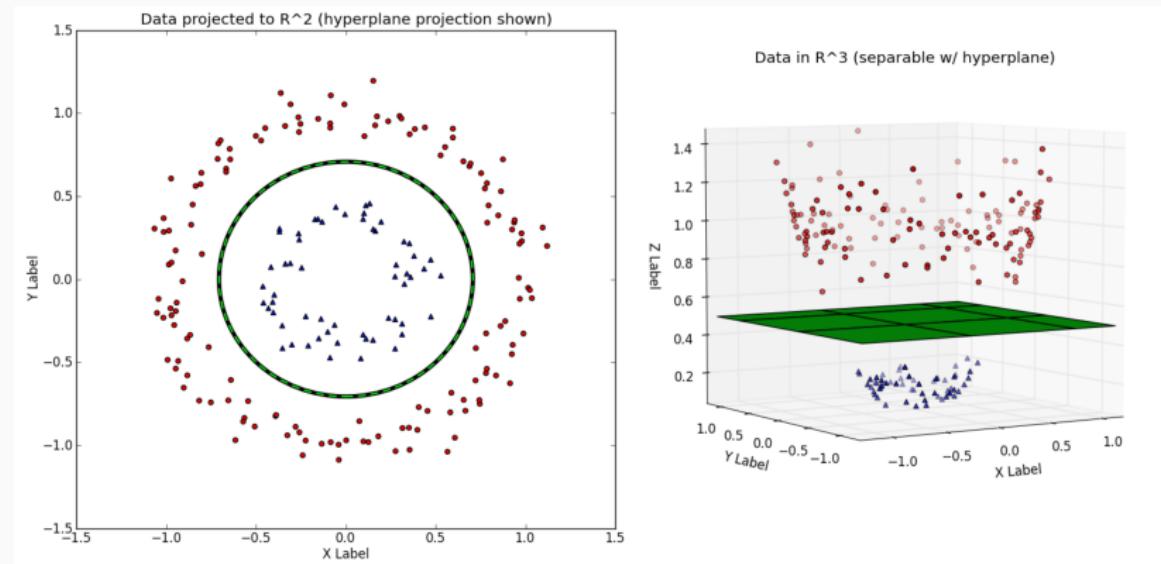
Recordatorio: Si pasamos a una dimensión mas alta, podemos resolver un problema non lineal como lineal

# SVM con kernel: aumento del espacio



Recordatorio: Si pasamos a una dimensión mas alta, podemos resolver un problema non lineal como lineal

# SVM con kernel: aumento del espacio



Recordatorio: Si pasamos a una dimensión mas alta, podemos resolver un problema non lineal como lineal

**Peligro:** Curse of dimensionality !!  $W$  es mas grande

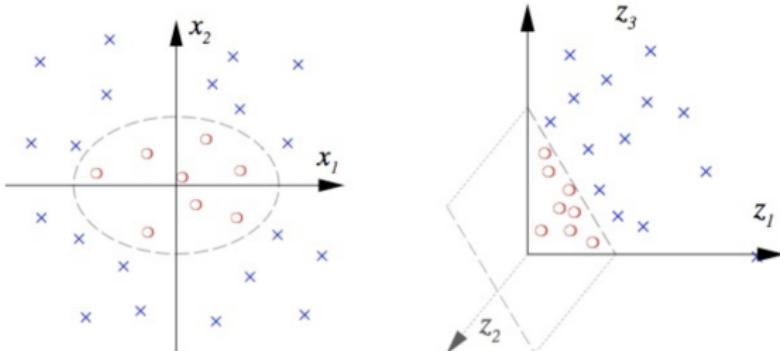
# SVM con kernel: Kernel Trick

- I. El problema inicial solo implica el cálculo de productos escalares  $\mathbf{X}_i^T \mathbf{X}_j$
- II. Si transformamos los datos usando una función no lineal  $\varphi$ , solo necesito saber calcular  $\varphi(\mathbf{X}_i)^T \varphi(\mathbf{X}_j)$
- III. Para un nuevo dato  $\mathbf{X}$ , necesitamos saber calcular  $\varphi(\mathbf{X})^T \varphi(\mathbf{X}_j)$
- IV. Podemos reemplazar  $\mathbf{X}_i^T \mathbf{X}_j$  por una función  $k$  llamada **núcleo**, tal que  $k(\mathbf{X}_i, \mathbf{X}_j) = \varphi(\mathbf{X}_i)^T \varphi(\mathbf{X}_j)$
- V . Teorema de Aronzajn: La matriz  $K = (k(\mathbf{X}_i, \mathbf{X}_j))_{i,j}$  es definida positiva  $\Leftrightarrow \exists \varphi$  tal que  $k(\mathbf{X}_i, \mathbf{X}_j) = \varphi(\mathbf{X}_i)^T \varphi(\mathbf{X}_j)$
- VI .  $\mathbf{W}$  es ahora de la misma dimensión que el espacio de llegada de  $\varphi$

# SVM con kernel: Kernel Trick

$$\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^3$$

$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



Reemplazo de  $\mathbf{X}$  por una transformación **no lineal**  $\Phi(\mathbf{X})$ :  
el clasificador esta  $f(\mathbf{X}) = \text{signo}(\sum_{i=1}^n \alpha_i Y_i (\Phi(\mathbf{X}_i)^T \Phi(\mathbf{X}) + b))$

## Truco del kernel: no necesitas conocer la función $\Phi$

- El cálculo de  $k(\mathbf{X}, \mathbf{X}') = \langle \Phi(\mathbf{X}) | \Phi(\mathbf{X}') \rangle$  es más simple que el de  $\Phi(\mathbf{X})$  y  $\Phi(\mathbf{X}')$  luego del producto escalar
- $\Phi$  puede ser especificada a través de su núcleo positivo  $k$

## Ejemplo de Kernel Trick

**Tomamos:**  $\Phi : (x_1, x_2) \rightarrow (x_1^2, \sqrt{2}x_1x_2, \sqrt{2}x_1, \sqrt{2}x_2, x_2^2, 1)$

$$\Phi(u, v) = u_1^2v_1^2 + 2u_1u_2v_1v_2 + u_2^2v_2^2 + 2u_1v_1 + 2u_2v_2 + 1$$

$$\Phi(u, v) = (u_1v_1 + u_2v_2 + 1)^2$$

$$\Phi(u, v) = (\langle u, v \rangle + 1)^2$$

El producto punto en el espacio transformado se puede expresar como una función de similitud en el espacio original:

$$K(u, v) = \Phi(u, v) = (\langle u, v \rangle + 1)^2$$

Una SVM clasificaría un ejemplo nuevo  $\mathbf{X}$  de la siguiente forma:

$$f(\mathbf{X}) = \text{signo}\left(\sum_{i=1}^n \alpha_i Y_i (\Phi(\mathbf{X}_i)^T \Phi(\mathbf{X}) + b)\right)$$

$$f(\mathbf{X}) = \text{signo}\left(\sum_{i=1}^n \alpha_i Y_i (K(\mathbf{X}_i^T \mathbf{X}) + b)\right)$$

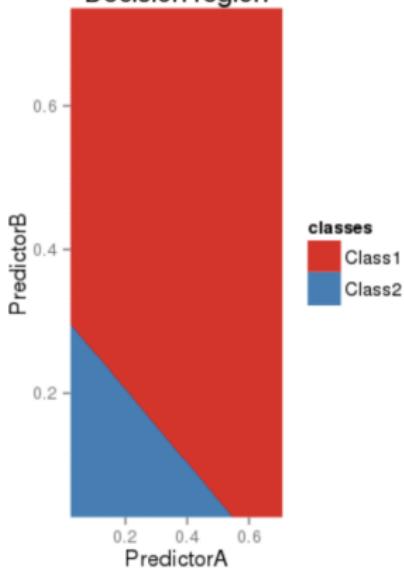
$$f(\mathbf{X}) = \text{signo}\left(\sum_{i=1}^n \alpha_i Y_i ((\langle \mathbf{X}_i, \mathbf{X} \rangle + 1)^2 + b)\right)$$

# SVM lineal

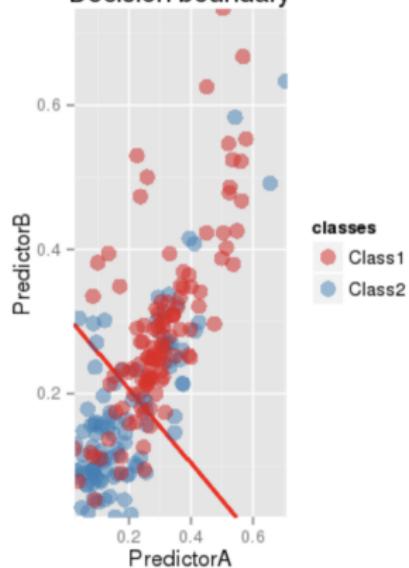
$$\text{Núcleo lineal : } k(\mathbf{X}, \mathbf{X}') = \langle \mathbf{X} | \mathbf{X}' \rangle$$

## Support Vector Machine

Decision region



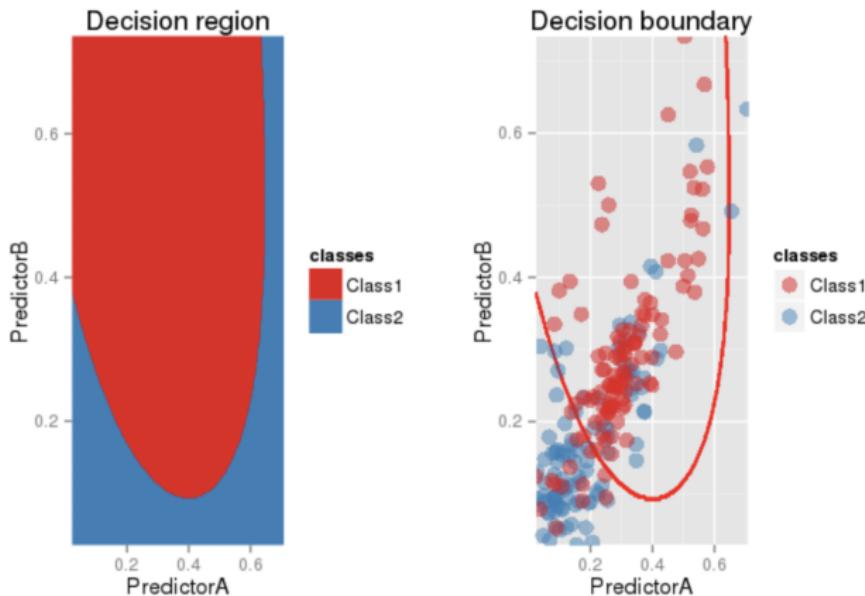
Decision boundary



# SVM: núcleo polinomial

$$\text{Núcleo polinomial : } k(\mathbf{X}, \mathbf{X}') = (1 + \langle \Phi(\mathbf{X}) | \Phi(\mathbf{X}') \rangle)^d$$

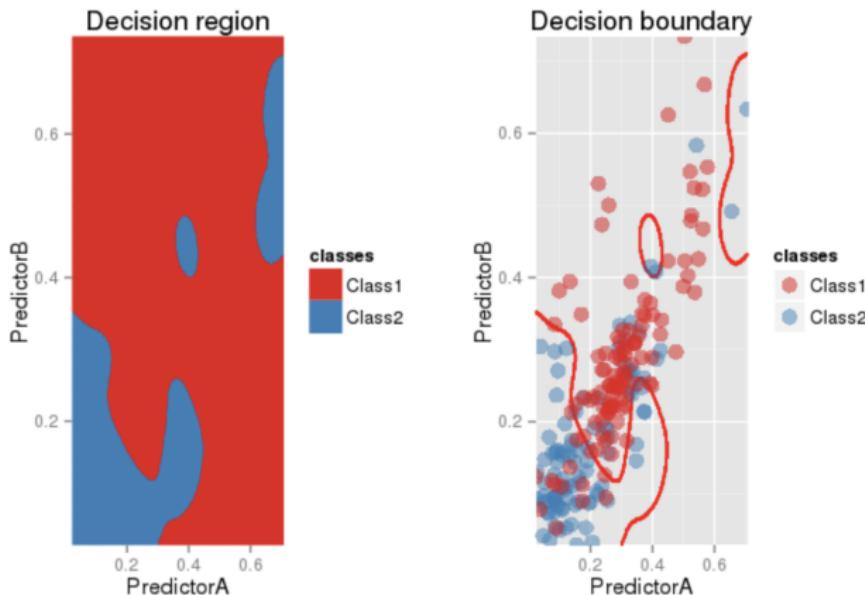
Support Vector Machine with polynomial kernel



# SVM: núcleo gaussiano

$$\text{Núcleo Gaussiano : } k(\mathbf{X}, \mathbf{X}') = \exp^{-\frac{-||\mathbf{x}-\mathbf{x}'||^2}{2}}$$

Support Vector Machine with Gaussian kernel



# SVM: Conclusion

---

- El problema de aprendizaje de una SVM se formula como un problema de **optimización convexa** en donde hay algoritmos eficientes para encontrar el óptimo global.
- Otros métodos de clasificación como los árboles de decisión y las redes neuronales tienden a encontrar **óptimos locales**.
- La SVM optimiza explícitamente la **capacidad de generalización** al maximizar el margen del límite de decisión.
- En una SVM el usuario debe **ajustar** hiper-parámetros, como el tipo de función de Kernel y el costo C para las variables de holgura (esto puede ser caro).
- La SVM puede aplicarse a los datos categóricos creando **variables dummy** binarias por cada categoría.
- La formulación de SVM presentada en esta clase se limita a problemas de **clasificación binaria**.

# Outline : SVR

---

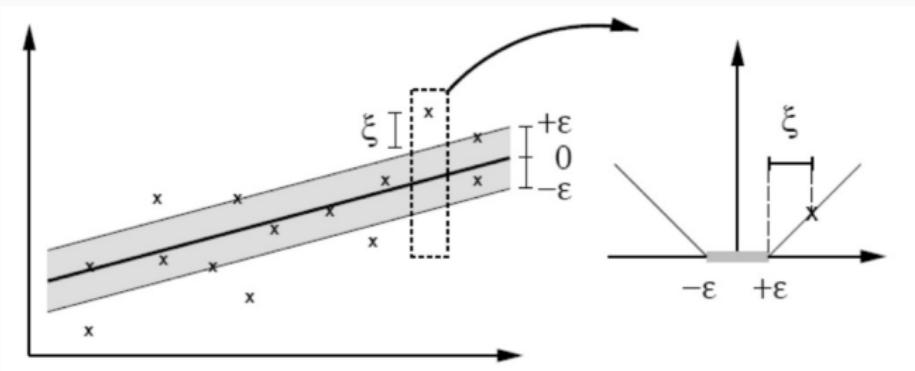
Máquinas de Vectores de Soporte

SVR

# SVR: Principio

- Extendemos la idea del margen máximo a la regresión.
- Imponemos un  $\epsilon$ -tubo: pérdida  $\epsilon$ -sensible:

$$|\hat{y} - y|_\epsilon = \max(0, |\hat{y} - y| - \epsilon)$$



En este caso, vamos a poner un costo si el punto esta a mas de  $\epsilon$  de la predicción.

## SVR en el espacio primal

Sea  $C$  y  $\epsilon$ :

$$\min_{b, \mathbf{W}, \xi} \frac{1}{2} \|\mathbf{W}\|^2 + \sum_i \xi_i + \xi_i^*$$

s.c.

$$Y_i - f(\mathbf{X}_i) \leq \epsilon + \xi_i$$

$$f(\mathbf{X}_i) - Y_i \leq \epsilon + \xi_i^*$$

$$\xi_i, \xi_i^* \geq 0$$

Donde  $f_{\mathbf{W}, b, \Phi} = \mathbf{W}^T \Phi(\mathbf{X}) + b$

## SVR: Dual

### SVR en el espacio dual

$$\min_{\alpha, \alpha^*} \sum_{i,j} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) k(\mathbf{X}_i, \mathbf{X}_j) + \epsilon \sum_i i(\alpha_i + \alpha_i^*) - \sum_i y_i(\alpha_i - \alpha_i^*)$$

S.C.

$$\sum_i (\alpha_i - \alpha_i^*) = 0$$

$$0 \geq \alpha_i \geq C$$

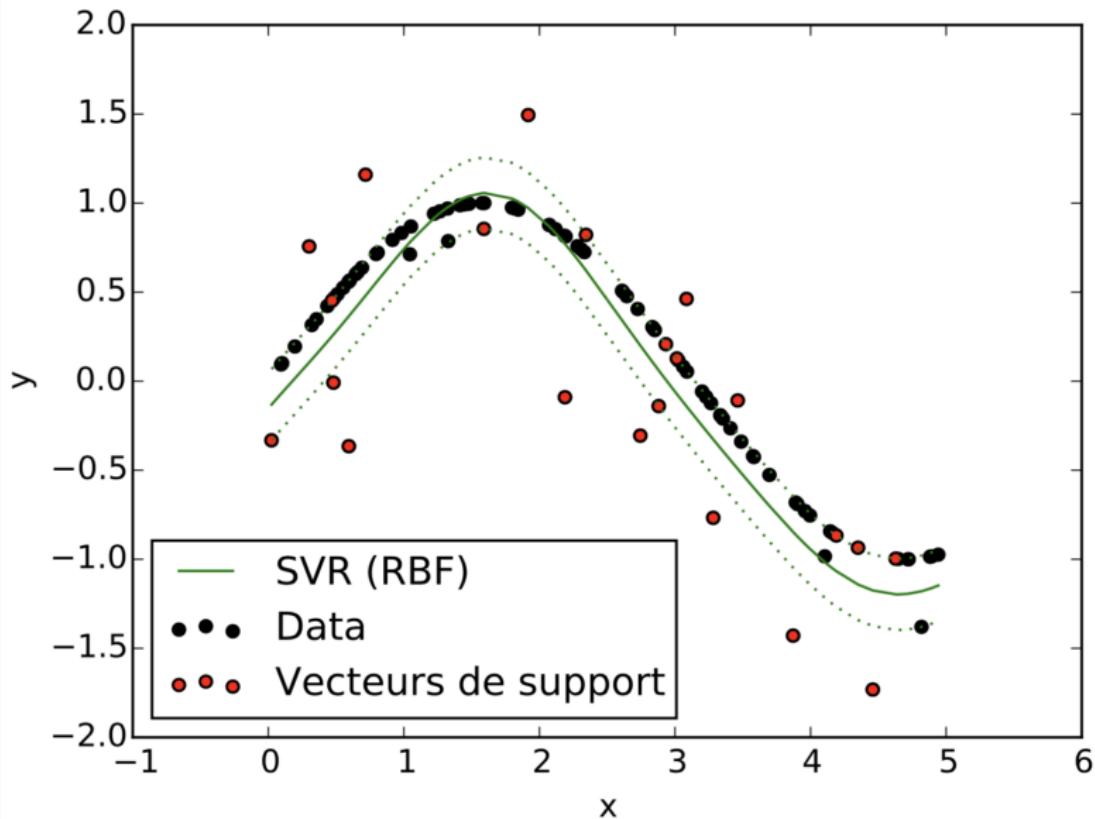
$$0 \geq \alpha_i^* \geq C$$

$$\mathbf{W} = \sum_{i=1}^n (\alpha_i - \alpha_i^*) \Phi(\mathbf{X}_i)$$

### Solución

$$f(\mathbf{X}) = \sum_i i = 1^n (\alpha_i - \alpha_i^*) k(\mathbf{X}_i, \mathbf{X}) + b$$

## SVR: Ejemplo



**Questions?**

## References i