CEINE.

Centro de Investigación en
Inteligencia de Negocios

Support Vector Machines

Ricardo Muñoz

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

fcfm
FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE

Contenido

- 1. Introducción
- 2. Caso Separable
- 3. Caso No-Separable
- 4. Extensiones

Ejemplo introductorio

Caso Retención de Clientes: “detección de fuga”.

- Dada ciertas características del cliente (edad, ingreso, crédito, saldo promedio, comportamiento en general) (atributos)
- Determinar si el cliente cerrará su cuenta corriente en los próximos meses.

Aprender de información de otros clientes, generar alguna “Regla” y aplicar esta regla a casos nuevos.

Minimización del riesgo empírico

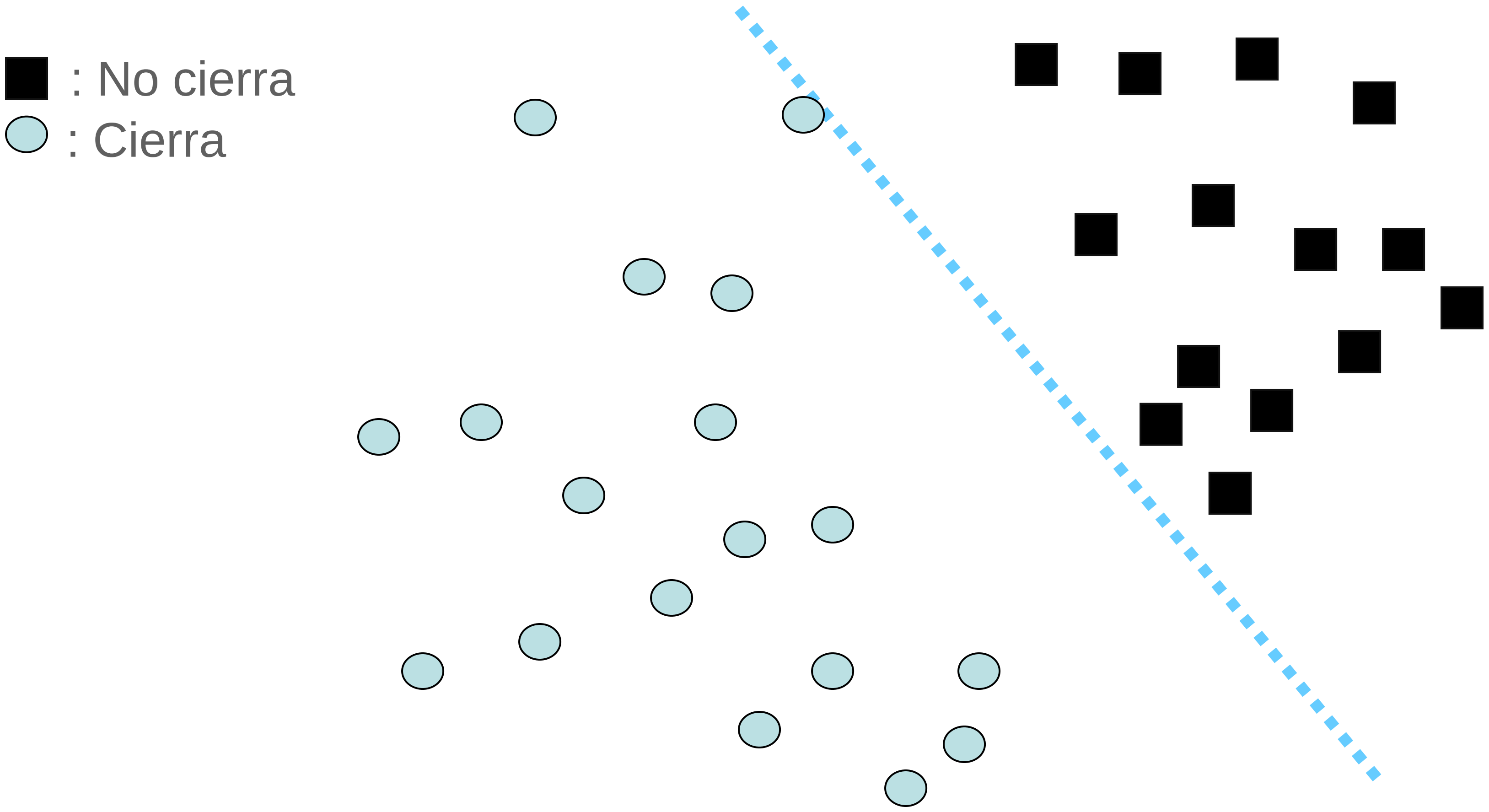
Queremos encontrar una función f que minimice:

$$R_{emp}[f] = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - f(x_i)\right)^2$$

Donde y es el valor conocido del objeto x, f(x) es la función de inducción y n es el número de objetos

Motivación SVM

Caso particular de dos conjuntos linealmente disjuntos en R2

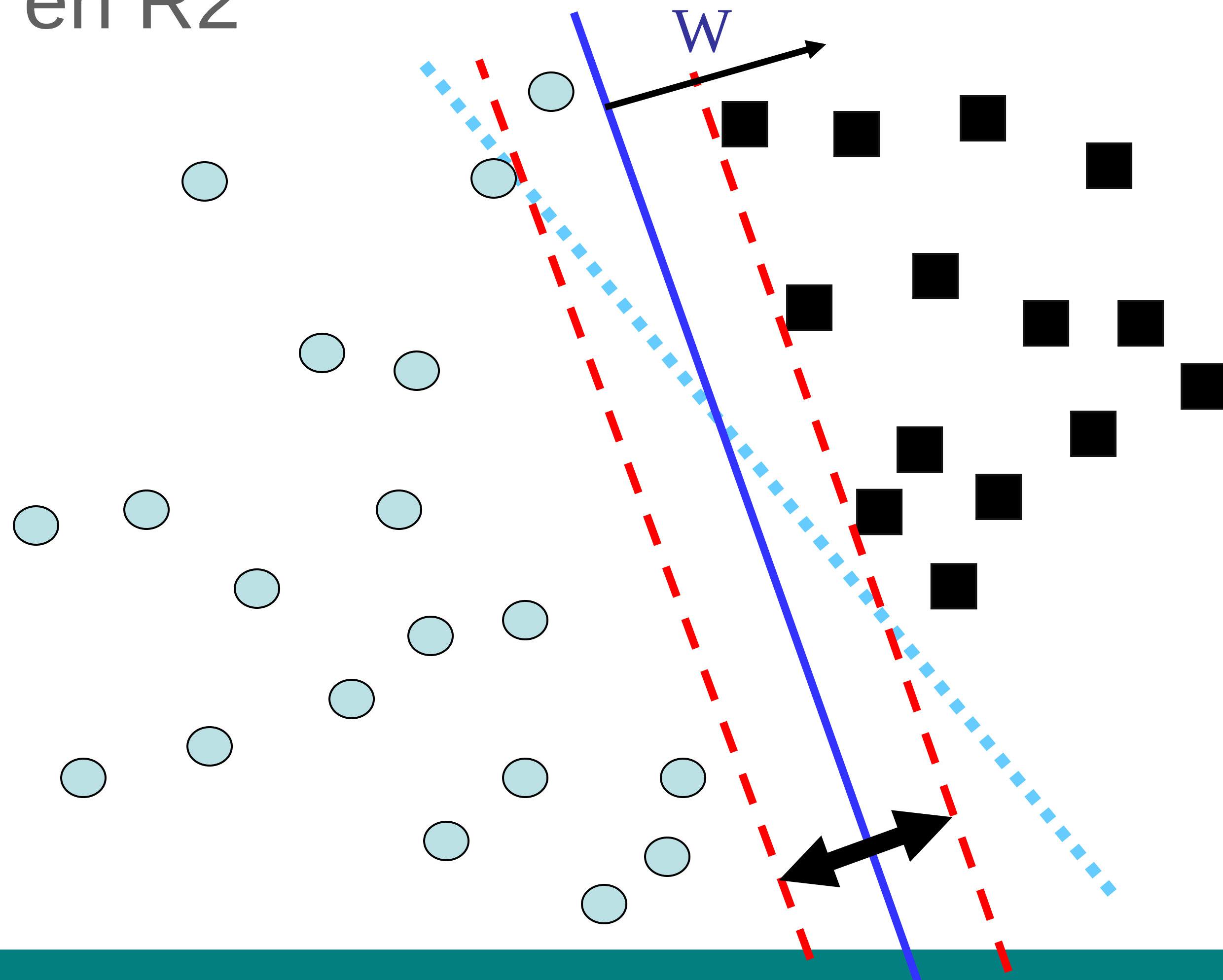


Motivación SVM

Caso particular de dos conjuntos linealmente disjuntos en R2

■ : No cierra

○ : Cierra



Support Vector Machines
(Para Clasificación)

IDEA:

Construir una función clasificadora que:

- Minimice el error en la separación de los objetos dados (del conjunto de entrenamiento)
- Maximice el margen de separación (mejora la generalización del clasificador en conjunto de test)

Dos objetivos:

Minimizar Error (ajuste del modelo)	Maximizar Margen (generalización)

SVM Lineal – Caso Separable

N objetos que consistenten del par : $\boldsymbol{x}_i \in \mathrm{R}^m$, i=1,…,n y de su "etiqueta" asociada $y_i \in \{-1,1\}$

Supongamos que $\exists$ un hiperplano separador $\boldsymbol{w}.\boldsymbol{x}$+b=0 que separa los ejemplos positivos de los ejemplos negativos. Esto es, Todos los objetos del conjunto de entrenamiento satisfacen:

$$x_i \cdot w + b \geq +1 \text{ cuando } y_i = +1$$
$$x_i \cdot w + b \leq -1 \text{ cuando } y_i = -1$$

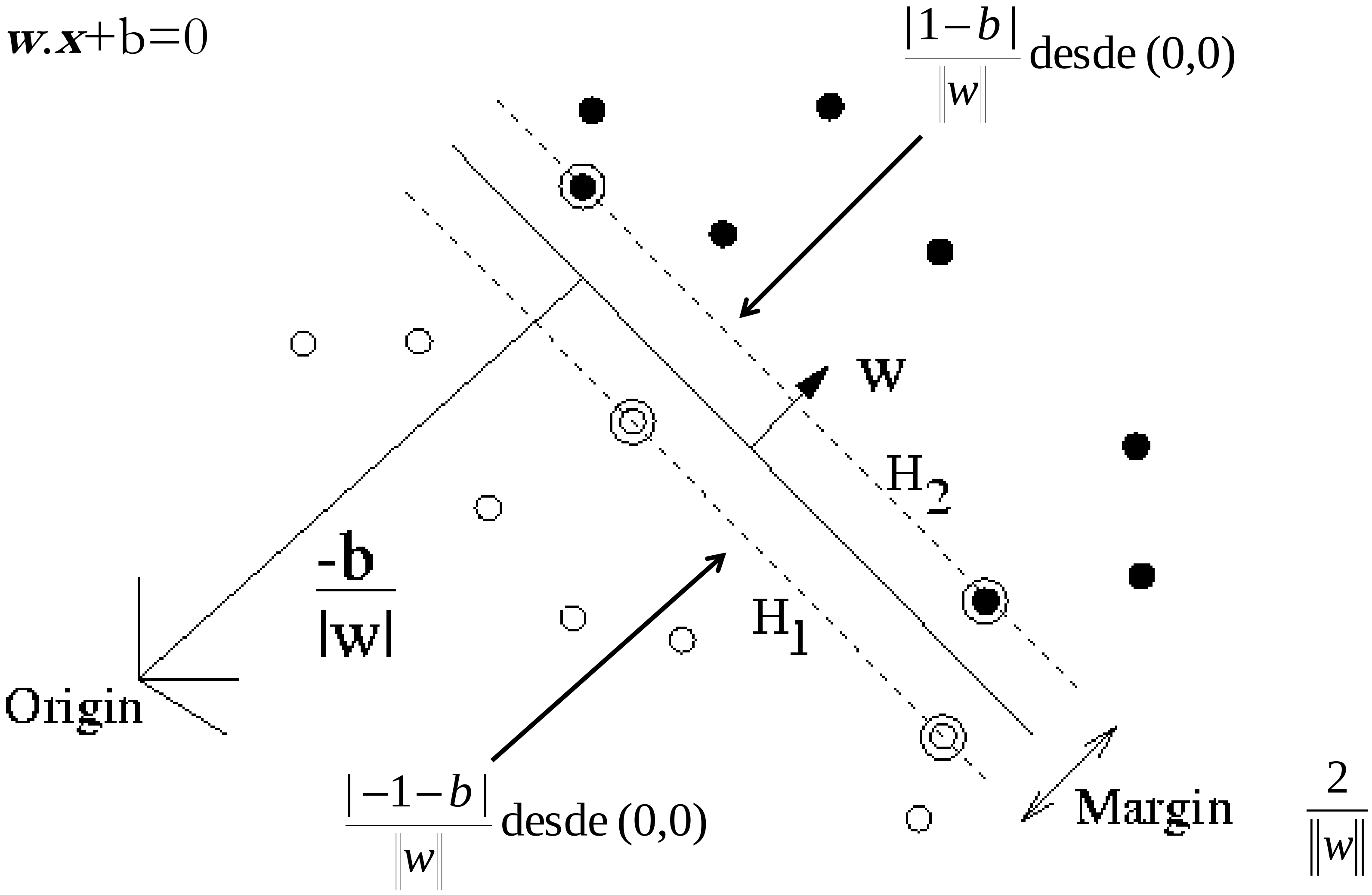
equivalentemente:

$$y_i(x_i \cdot w + b) - 1 \geq 0 \forall i$$

Sean d_+ (d_-) las distancias más cercanas desde el hiperplano separador al ejemplo positivo (negativo) más cercano. El *margen* del hiperplano separador se define como $d_+ + d_-$

SVM Lineal – Caso Separable

$$\boldsymbol{w}.\boldsymbol{x}+\mathrm{b}=0$$



Formulación matemática (SVM primal)

1/Margen

$$\text{Minimizar } \frac{1}{2}\|W\|^2$$

$$\text{sujeto a :}$$

$$y_i\left(x_i \cdot w + b\right) - 1 \geq 0$$

W: Normal al hiperplano separador.

b : Posición del hiperplano

X_i: Objetos de entrenamiento

Y_i : Clase del objeto i.

Formulación matemática (SVM L dual)

KERNELS!!!

Luego...

$$\text{Maximizar} \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_s y_i y_s x_i x_s$$

sujeto a :

$$\alpha_i \geq 0 \qquad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

SVM Lineal – Caso No Separable

N objetos que consistenten del par : $\boldsymbol{x}_i \in \mathrm{R}^m$, i=1,…,n y de su "etiqueta" asociada $y_i \in \{-1,1\}$

Se introducen variables de holgura positivas ξ_i:

$$x_i \cdot w + b \geq +1 - \xi_i \text{ cuando } y_i = +1$$
$$x_i \cdot w + b \leq -1 + \xi_i \text{ cuando } y_i = -1$$

Y se modifica la función objetivo a:

$$\|w\|^2 / 2 + C(\sum \xi_i)$$

Formulación matemática (SVM primal)

1/Margen

Error en clasificación

$$\text{Minimizar } \frac{1}{2}\|W\|^2 + C\sum \xi_i$$

$$\text{sujeto a :}$$

$$y_i(x_i \cdot w + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$

W: Normal al hiperplano separador.

b : Posición del hiperplano

X_i: Objetos de entrenamiento

Y_i : Clase del objeto i.

ξ_i : Error en la separación

Formulación matemática (SVM dual)

KERNELS!!!

Luego...

$$\text{Maximizar} \sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_s y_i y_s x_i x_s$$

sujeto a :

$$C \geq \alpha_i \geq 0 \qquad i = 1,\ldots,n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

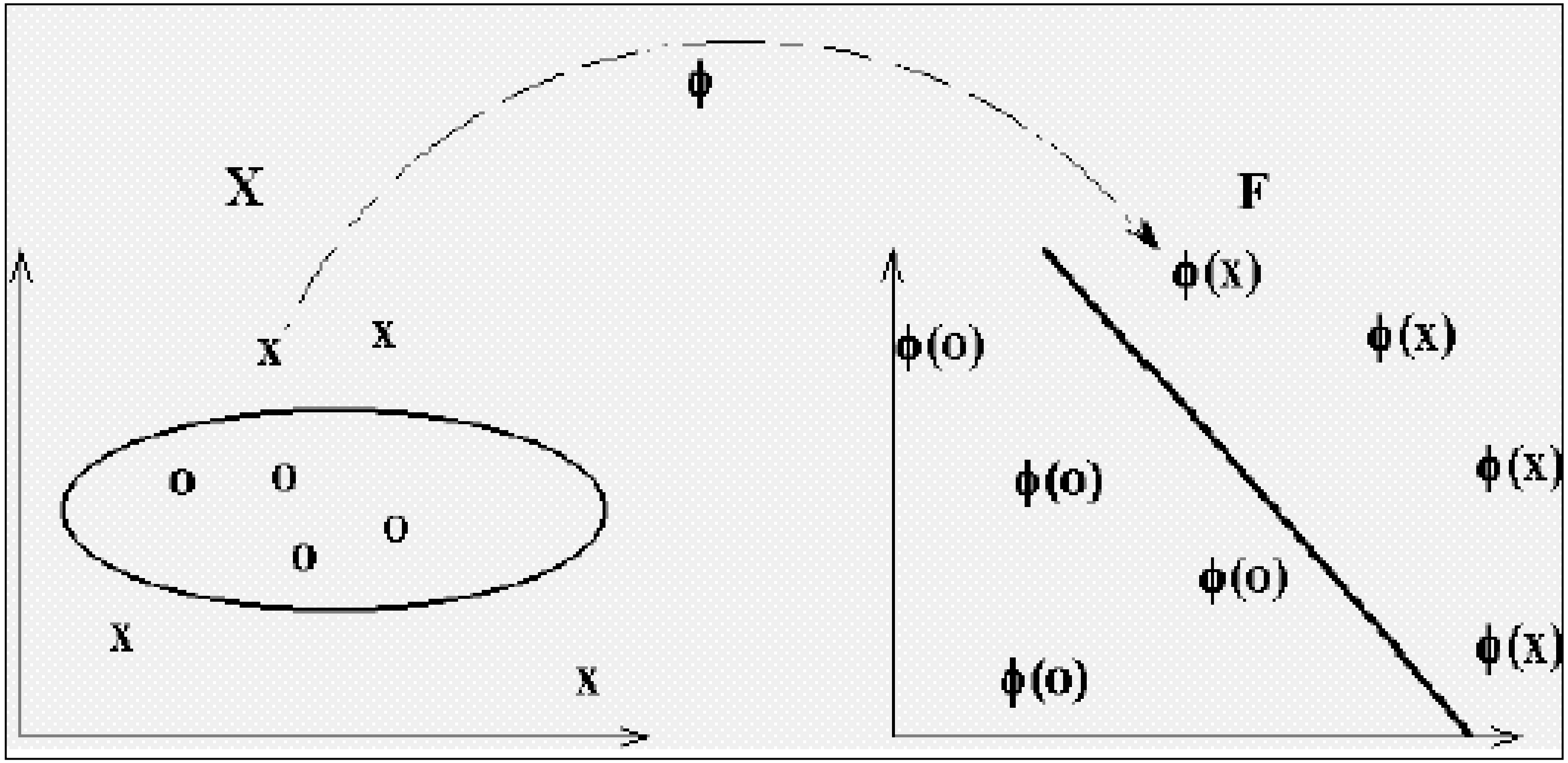
Clasificador

- El clasificador lineal de los SVM es:

$$f(x) = \mathrm{W} \cdot x + \mathrm{b} = \sum_i \alpha_i y_i \cdot x \; + b$$

- Se determina el signo de la función *f(x)*
  - Si signo(f(x)) = +1 $\Rightarrow$ pertenece a clase +1
  - Si signo(f(x)) = -1 $\Rightarrow$ pertenece a clase -1

SVM no lineal

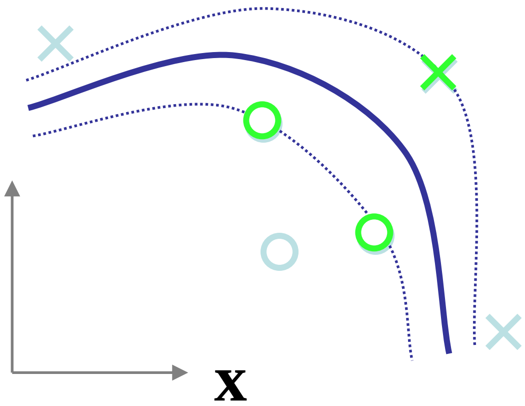


Objetos linealmente no separables en R^2, pueden serlo otro espacio

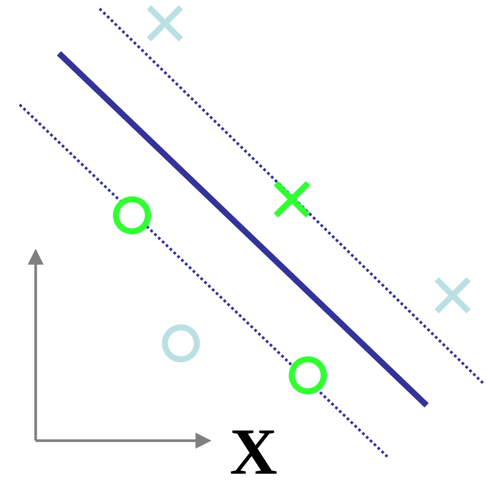
SVM no lineal

- Idea:
  - Proyectar los objetos a un espacio de mayor dimensión y realizar una clasificación lineal en este nuevo espacio.
  - Función de transformación
  - $K(x_i, x_s) = \Phi(x_i) \bullet \Phi(x_s)$
  - Basta reemplazar $x_i \cdot x_s$ por K(x_i , x_s)

Kernel Machines



$$\Phi(\cdot)$$



$$\mathbf{X}_i = \Phi(\mathbf{x}_i)$$

$$\mathbf{X} = \Phi(\mathbf{x})$$

$$\mathbf{X}_i \cdot \mathbf{X} = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x})$$

$$y = \operatorname{sign}\Big(\sum_{i \in S} \alpha_i y_i \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) + b\Big)$$

$$y = \operatorname{sign}\Big(\sum_{i \in S} \alpha_i y_i \mathbf{X}_i \cdot \mathbf{X} + b\Big)$$

$$K(\cdot,\cdot)$$

$$\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}) = K(\mathbf{x}_i, \mathbf{x})$$

Condición de Mercer

$$y = \operatorname{sign}\Big(\sum_{i \in S} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\Big)$$

Funciones de Kernel

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$$

$$K(\mathbf{x}, \mathbf{y}) = e^{-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2}$$

$$K(\mathbf{x}, \mathbf{y}) = \tanh(\kappa \mathbf{x} \cdot \mathbf{y} - \delta)$$

SVM para selección de atributos

- Idea:

  Penalizar en la función objetivo por cada atributo utilizado.

- Función de penalización:

  Penalizar si el coeficiente asociado al atributo es mayor que cero.

$$f(x) = e - \exp(-\alpha x)$$

$\alpha > 0$

e :vector de unos

CEINE.

SVM para selección de atributos

$$\text{Minimizar } \frac{1}{2}\left\|W\right\|^2 + C_1 \sum \xi_i + \boxed{C_2\, e^T (e - \exp(-\alpha v))}$$

$$\text{sujeto a:}$$

$$y_i (x_i \cdot w + b) - 1 + \xi_i \geq 0$$

$$\xi_i \geq 0$$

$$-v \leq w \leq v$$

$$v \geq 0$$

Características de SVM

- Herramienta matemática
- No tiene mínimos locales (árboles de decisión)
- No tiene el problema de Overfitting (Redes Neuronales)
- Solución no depende de estructura del planteamiento del problema.
- Aplicabilidad en distintos tipos de problemas (Clasificación, Regresión, descubrimiento de patrones en general)

Referencias

- Apuntes Curso “Introducción a la minería de datos”, Richard Weber.
- C. Burges, A tutorial on support vector machines for pattern recognition, Data Mining and Knowledge Discovery 2 (1998), no. 2, 121–167

CEINE.

Centro de Investigación en
Inteligencia de Negocios

INGENIERIA INDUSTRIAL
UNIVERSIDAD DE CHILE

fcfm

FACULTAD DE CIENCIAS
FÍSICAS Y MATEMÁTICAS
UNIVERSIDAD DE CHILE