

# Support Vector Machines

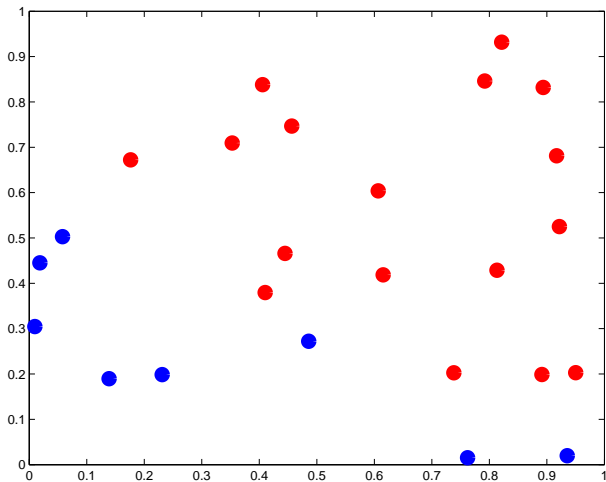
Fernando Lozano

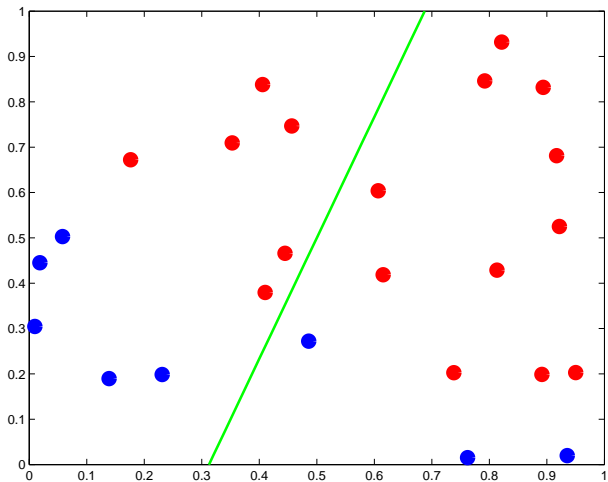
Universidad de los Andes

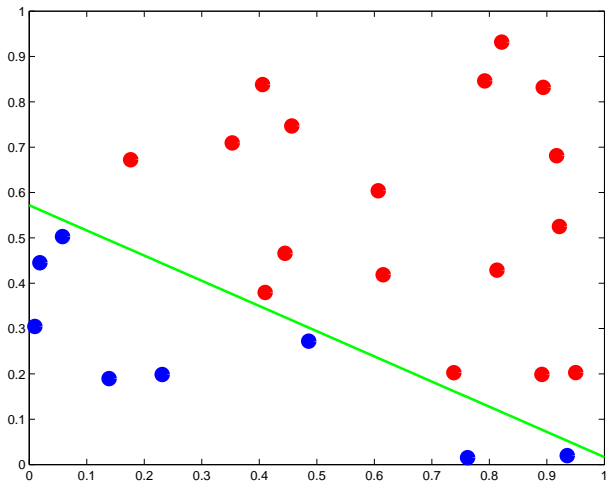
30 de septiembre de 2014



# El perceptrón







# Convergencia del Perceptrón

Teorema

*Suponga:*

# Convergencia del Perceptrón

## Teorema

*Suponga:*

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$

# Convergencia del Perceptrón

## Teorema

*Suponga:*

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

# Convergencia del Perceptrón

## Teorema

*Suponga:*

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

*Entonces el algoritmo del perceptrón ejecuta el paso de actualización a lo sumo  $\left(\frac{K\|\hat{\mathbf{w}}\|}{\delta}\right)^2$  veces.*



# Convergencia del Perceptrón

## Teorema

*Suponga:*

- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

*Entonces el algoritmo del perceptrón ejecuta el paso de actualización a lo sumo  $\left(\frac{K\|\hat{\mathbf{w}}\|}{\delta}\right)^2$  veces.*

- Deseable tener margen  $\delta$  grande.

# Convergencia del Perceptrón

## Teorema

*Suponga:*

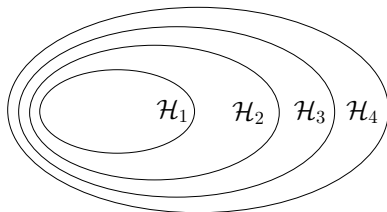
- $\|\mathbf{x}_i\| \leq K \in \mathbb{R}, \quad i = 1 \dots, n.$
- $\exists \hat{\mathbf{w}} \in \mathbb{R}^{d+1}, \delta > 0 \quad \text{tal que} \quad \hat{\mathbf{w}}^T \mathbf{x}_i \geq \delta \quad i = 1, \dots, n.$

*Entonces el algoritmo del perceptrón ejecuta el paso de actualización a lo sumo  $\left(\frac{K\|\hat{\mathbf{w}}\|}{\delta}\right)^2$  veces.*

- Deseable tener margen  $\delta$  grande.
- Algoritmo del perceptrón no tiene en cuenta el margen.

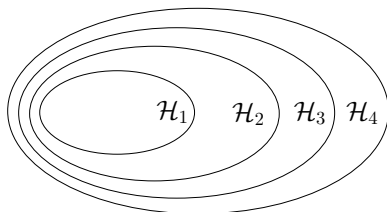
# Motivación desde teoría de aprendizaje

- Structural Risk Minimization:



# Motivación desde teoría de aprendizaje

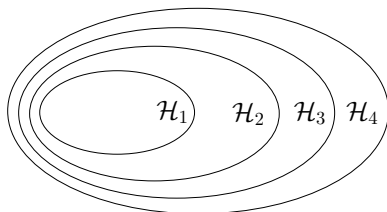
- Structural Risk Minimization:



- Algoritmo de aprendizaje determina la complejidad apropiada de la clase de hipótesis.

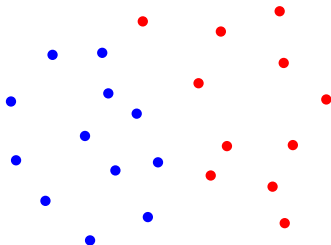
# Motivación desde teoría de aprendizaje

- Structural Risk Minimization:

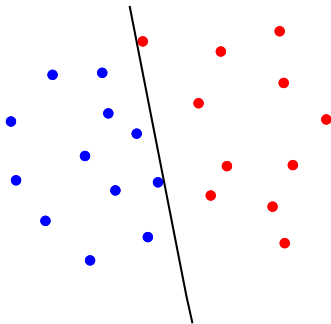


- Algoritmo de aprendizaje determina la complejidad apropiada de la clase de hipótesis.
- Cómo variar **suavemente** la complejidad?

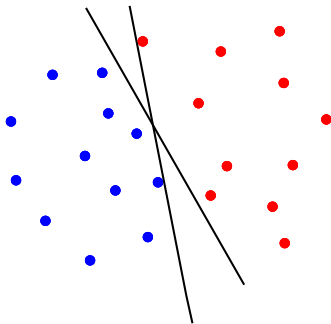
# Clasificador con margen (caso linealmente separable)



# Clasificador con margen (caso linealmente separable)



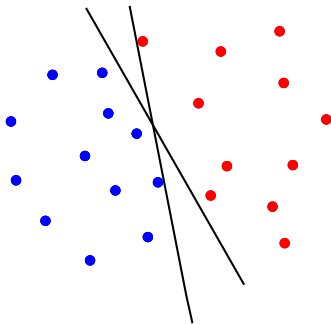
# Clasificador con margen (caso linealmente separable)





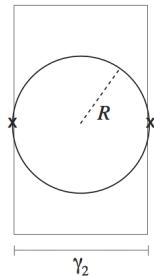
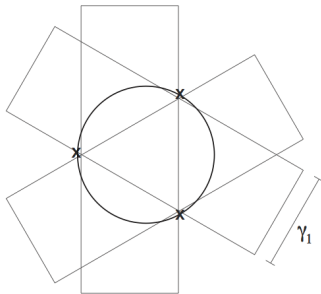


# Clasificador con márgen (caso linealmente separable)



- **Márgen**: Distancia de un punto a la superficie de separación.
- Es deseable tener **márgenes grandes**.

# Margen grande vs. Complejidad



# Cómo encontrar un separador lineal con margen grande?

# Cómo encontrar un separador lineal con margen grande?

- Hipótesis  $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$

# Cómo encontrar un separador lineal con margen grande?

- Hipótesis  $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$
- Separación:

# Cómo encontrar un separador lineal con margen grande?

- Hipótesis  $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$
- Separación:

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\geq 1 && \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\leq -1 && \text{Si } y_i = -1 \end{aligned}$$

# Cómo encontrar un separador lineal con margen grande?

- Hipótesis  $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$
- Separación:

$$\left. \begin{array}{ll} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 & \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 & \text{Si } y_i = -1 \end{array} \right\} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad \forall i$$

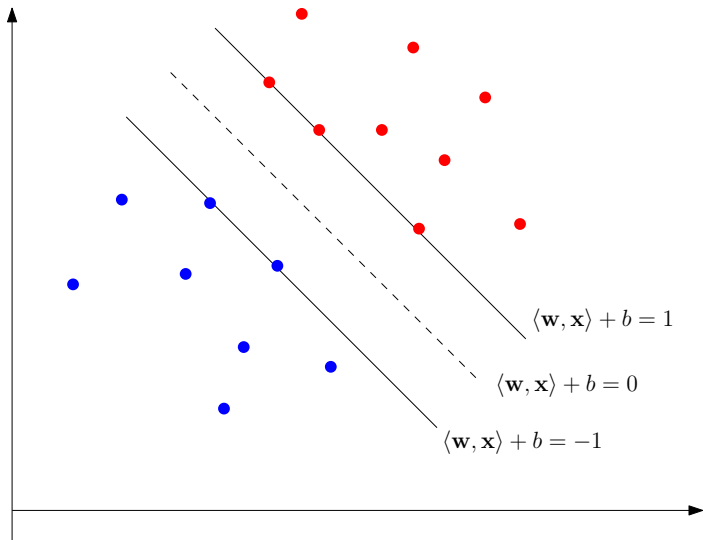


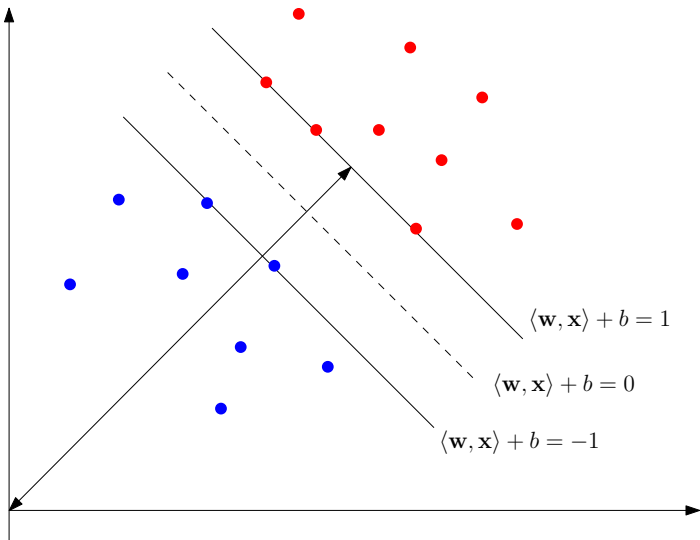
# Cómo encontrar un separador lineal con margen grande?

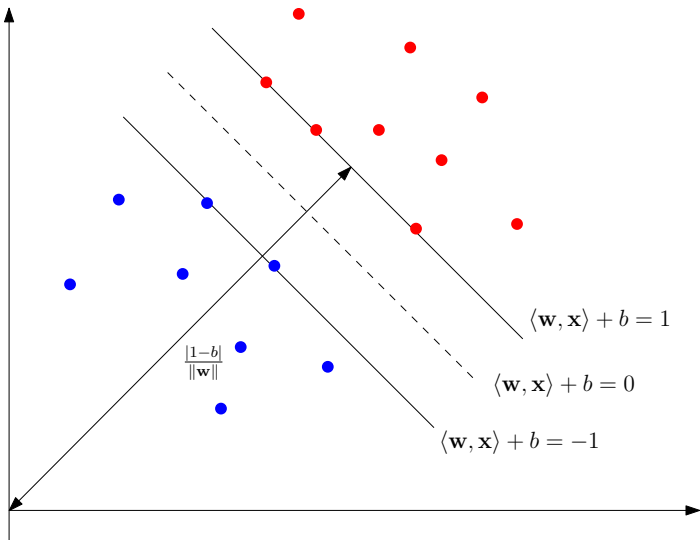
- Hipótesis  $h(\mathbf{x}) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$
- Separación:

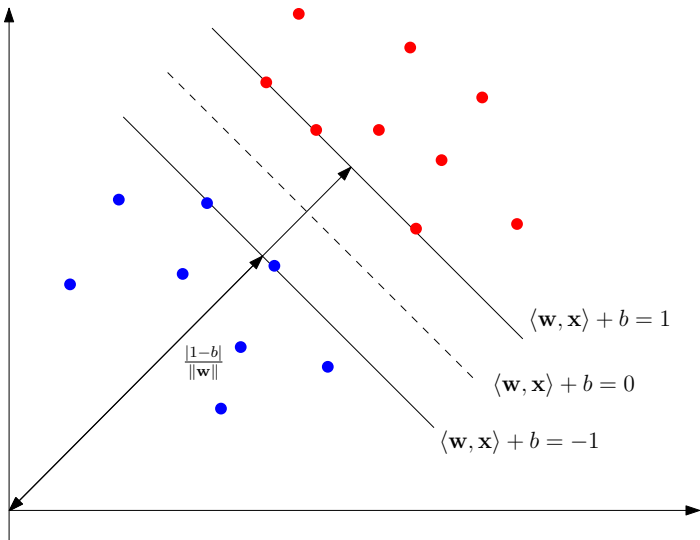
$$\left. \begin{array}{ll} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 & \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 & \text{Si } y_i = -1 \end{array} \right\} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad \forall i$$

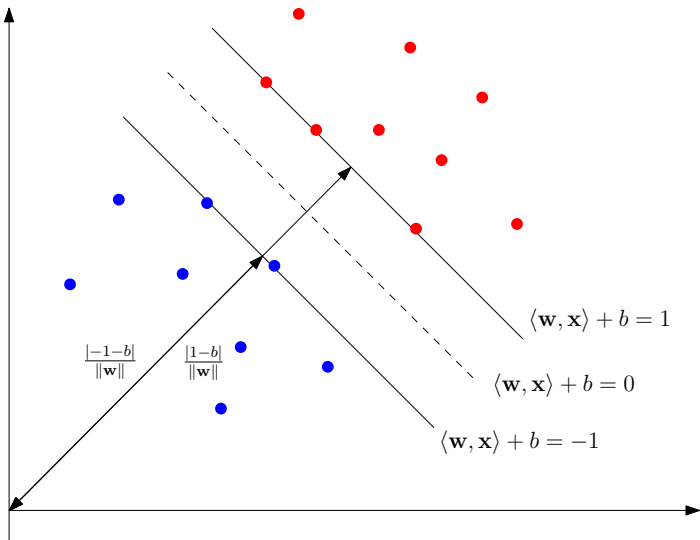
- **Márgen?**

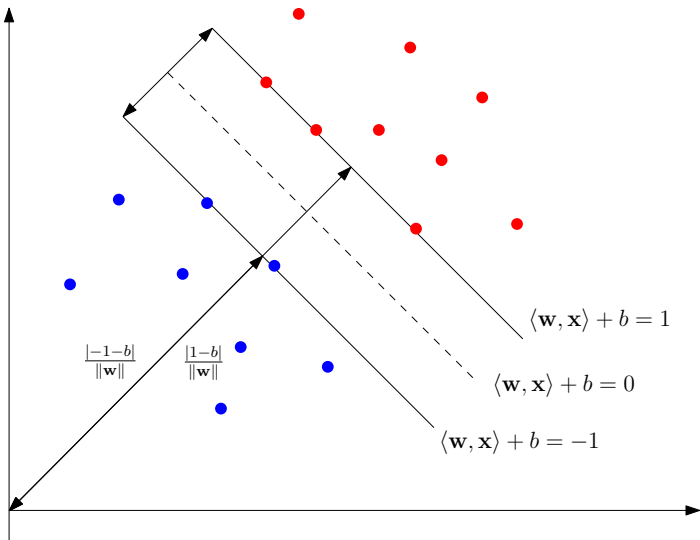


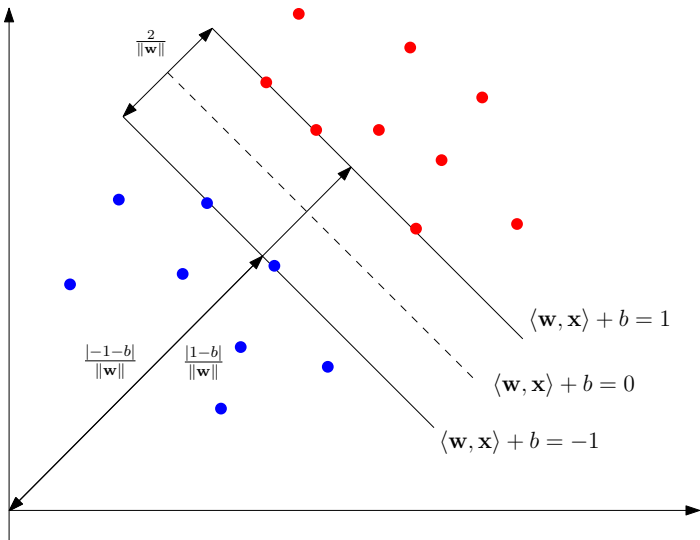














- Problema de optimización:

$$\begin{array}{ll} \text{mín} & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{array}$$

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de programación cuadrática.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de programación cuadrática.
- Problema convexo:

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de programación cuadrática.
- Problema convexo:
  - ▶ Único mínimo global.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de programación cuadrática.
- Problema convexo:
  - ▶ Único mínimo global.
  - ▶ Condiciones de Karush-Kuhn-Tucker (KKT) son suficientes y necesarias para mínimo global y máximo global del problema dual.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de programación cuadrática.
- Problema convexo:
  - ▶ Único mínimo global.
  - ▶ Condiciones de Karush-Kuhn-Tucker (KKT) son suficientes y necesarias para mínimo global y máximo global del problema dual.
  - ▶ Solución eficiente.

- Problema de optimización:

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n \end{aligned}$$

- Problema de **programación cuadrática**.
- Problema **convexo**:
  - ▶ Único mínimo global.
  - ▶ Condiciones de Karush-Kuhn-Tucker (KKT) son suficientes y necesarias para **mínimo global y máximo global del problema dual**.
  - ▶ Solución eficiente.
  - ▶ Tamaño puede ser grande.

# Problema dual



# Problema dual

- Introducimos multiplicadores de Lagrange

$$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0.$$

## Problema dual

- Introducimos multiplicadores de Lagrange  $\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

# Problema dual

- Introducimos multiplicadores de Lagrange

$\boldsymbol{\alpha} = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1)$$

## Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

## Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

- Minimizamos  $L(\mathbf{w}, b, \alpha)$  con respecto a  $\mathbf{w}, b$  para obtener la **función dual**:

## Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

- Minimizamos  $L(\mathbf{w}, b, \alpha)$  con respecto a  $\mathbf{w}, b$  para obtener la **función dual**:

$$\frac{\partial L}{\partial b} = 0$$

## Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

- Minimizamos  $L(\mathbf{w}, b, \alpha)$  con respecto a  $\mathbf{w}, b$  para obtener la **función dual**:

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

## Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

- Minimizamos  $L(\mathbf{w}, b, \alpha)$  con respecto a  $\mathbf{w}, b$  para obtener la **función dual**:

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0$$



# Problema dual

- Introducimos multiplicadores de Lagrange

$\alpha = [\alpha_1, \alpha_2, \dots, \alpha_n] \geq 0$ . El **Lagrangiano**:

$$\begin{aligned} L(\mathbf{w}, b, \alpha) &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) \\ &= \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) + \sum_{i=1}^n \alpha_i \end{aligned}$$

- Minimizamos  $L(\mathbf{w}, b, \alpha)$  con respecto a  $\mathbf{w}, b$  para obtener la **función dual**:

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^n \alpha_i y_i = 0$$

$$\frac{\partial L}{\partial \mathbf{w}} = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

- Reemplazando en el Lagrangiano:

- Reemplazando en el Lagrangiano:

$$L = \frac{1}{2} \left\langle \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i, \sum_{j=1}^n (\alpha_j y_j) \mathbf{x}_j \right\rangle \\ - \sum_{i=1}^n \alpha_i y_i \left( \left\langle \sum_{j=1}^n (\alpha_j y_j) \mathbf{x}_j, \mathbf{x}_i \right\rangle + b \right) + \sum_{i=1}^n \alpha_i$$

- Reemplazando en el Lagrangiano:

$$L = \frac{1}{2} \left\langle \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i, \sum_{j=1}^n (\alpha_j y_j) \mathbf{x}_j \right\rangle - \sum_{i=1}^n \alpha_i y_i \left( \left\langle \sum_{j=1}^n (\alpha_j y_j) \mathbf{x}_j, \mathbf{x}_i \right\rangle + b \right) + \sum_{i=1}^n \alpha_i$$

- Obtenemos la **función dual**:

$$g(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$

# Problema Dual

$$\begin{aligned} & \text{máx} && \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & \text{sujeto a} && \sum_{i=1}^n \alpha_i y_i = 0 \\ & && \boldsymbol{\alpha} \geq 0 \end{aligned}$$

# Problema Dual

$$\begin{aligned} & \text{máx} && \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & \text{sujeto a} && \sum_{i=1}^n \alpha_i y_i = 0 \\ & && \boldsymbol{\alpha} \geq 0 \end{aligned}$$

- Problema de programación cuadrática **cóncavo**.

# Problema Dual

$$\begin{aligned} & \text{máx} && \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & \text{sujeto a} && \sum_{i=1}^n \alpha_i y_i = 0 \\ & && \boldsymbol{\alpha} \geq 0 \end{aligned}$$

- Problema de programación cuadrática **cóncavo**.
- Los datos  $\mathbf{x}_i$  sólo aparecen en productos punto.

# Condiciones de Karush-Kuhn-Tucker (KKT)



# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$
- ❷  $\alpha_i \geq 0 \quad i = 1, \dots, n$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$
- ❷  $\alpha_i \geq 0 \quad i = 1, \dots, n$
- ❸  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{2} \quad \alpha_i \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{3} \quad \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

$$\textcircled{4} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{2} \quad \alpha_i \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{3} \quad \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

$$\textcircled{4} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$\textcircled{5} \quad \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) = 0$$

# Condiciones de Karush-Kuhn-Tucker (KKT)

①  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$

②  $\alpha_i \geq 0 \quad i = 1, \dots, n$

③  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$

④  $\sum_{i=1}^n \alpha_i y_i = 0$

⑤  $\alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) = 0$

- Las condiciones 3 y 5 implican que la suma  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$  sólo involucra vectores  $\mathbf{x}_i$  para los cuales la restricción es **activa**.

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{2} \quad \alpha_i \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{3} \quad \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

$$\textcircled{4} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$\textcircled{5} \quad \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) = 0$$

- Las condiciones 3 y 5 implican que la suma  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$  sólo involucra vectores  $\mathbf{x}_i$  para los cuales la restricción es **activa**.
- Estos vectores se llaman **vectores de soporte**.

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{2} \quad \alpha_i \geq 0 \quad i = 1, \dots, n$$

$$\textcircled{3} \quad \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

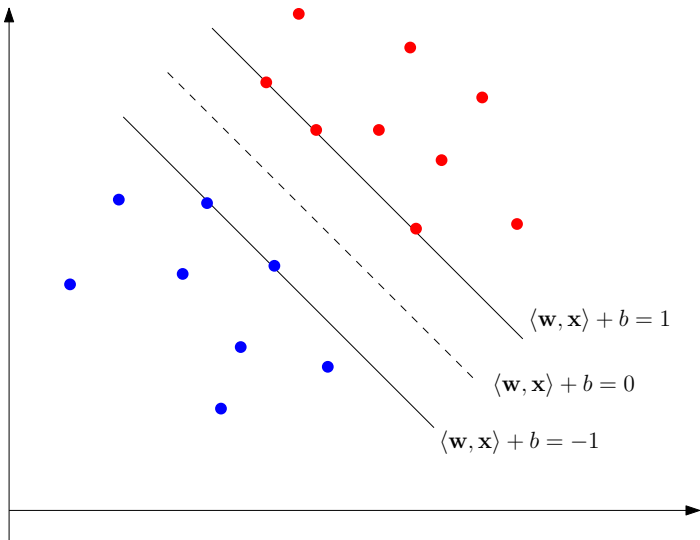
$$\textcircled{4} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

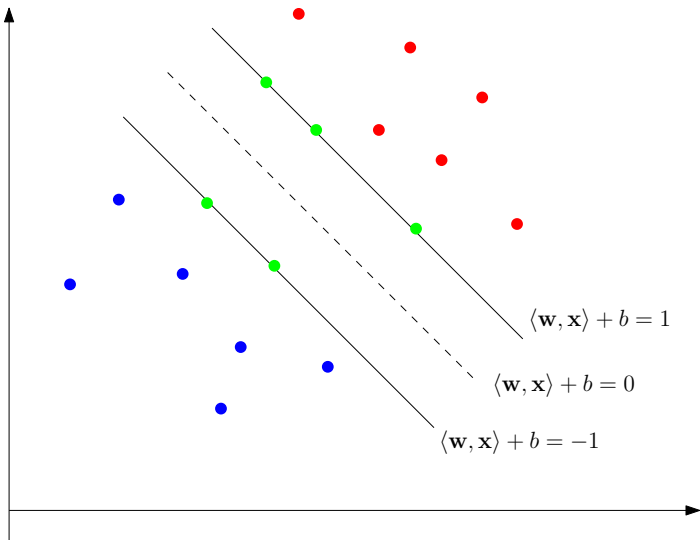
$$\textcircled{5} \quad \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1) = 0$$

- Las condiciones 3 y 5 implican que la suma  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$  sólo involucra vectores  $\mathbf{x}_i$  para los cuales la restricción es **activa**.
- Estos vectores se llaman **vectores de soporte**.
- Si  $S$  es el conjunto de vectores de soporte tenemos:

$$\mathbf{w} = \sum_{i: \mathbf{x}_i \in S} (\alpha_i y_i) \mathbf{x}_i$$







# Caso no separable

- En general, puede no existir una solución con error cero en los datos.

## Caso no separable

- En general, puede no existir una solución con error cero en los datos.
- Se introducen **variables de holgura**  $\zeta_1, \zeta_2, \dots, \zeta_n \geq 0$ :

## Caso no separable

- En general, puede no existir una solución con error cero en los datos.
- Se introducen **variables de holgura**  $\zeta_1, \zeta_2, \dots, \zeta_n \geq 0$ :

$$\begin{aligned} \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\geq 1 - \zeta_i && \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b &\leq -1 + \zeta_i && \text{Si } y_i = -1 \end{aligned}$$

## Caso no separable

- En general, puede no existir una solución con error cero en los datos.
- Se introducen **variables de holgura**  $\zeta_1, \zeta_2, \dots, \zeta_n \geq 0$ :

$$\left. \begin{array}{ll} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 - \zeta_i & \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 + \zeta_i & \text{Si } y_i = -1 \end{array} \right\} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad \forall i$$

- ▶ Si hay error en  $\mathbf{x}_i$

## Caso no separable

- En general, puede no existir una solución con error cero en los datos.
- Se introducen **variables de holgura**  $\zeta_1, \zeta_2, \dots, \zeta_n \geq 0$ :

$$\left. \begin{array}{ll} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 - \zeta_i & \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 + \zeta_i & \text{Si } y_i = -1 \end{array} \right\} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad \forall i$$

- ▶ Si hay error en  $\mathbf{x}_i \Rightarrow \zeta_i > 1$ .

# Caso no separable

- En general, puede no existir una solución con error cero en los datos.
- Se introducen **variables de holgura**  $\zeta_1, \zeta_2, \dots, \zeta_n \geq 0$ :

$$\left. \begin{array}{ll} \langle \mathbf{w}, \mathbf{x}_i \rangle + b \geq 1 - \zeta_i & \text{Si } y_i = 1 \\ \langle \mathbf{w}, \mathbf{x}_i \rangle + b \leq -1 + \zeta_i & \text{Si } y_i = -1 \end{array} \right\} y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad \forall i$$

- ▶ Si hay error en  $\mathbf{x}_i \Rightarrow \zeta_i > 1$ .
- ▶ Luego  $\sum_{i=1}^n \zeta_i$  es una **cota superior** del número de errores.



# Problema de optimización

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_i \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n \\ & \zeta_i \geq 0 \end{aligned}$$

# Problema de optimización

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_i \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n \\ & \zeta_i \geq 0 \end{aligned}$$

- $C$  es un parámetro que indica el balance deseado entre margen y error.

# Problema de optimización

$$\begin{aligned} \text{mín} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^n \zeta_i \\ \text{sujeto a} \quad & y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n \\ & \zeta_i \geq 0 \end{aligned}$$

- $C$  es un parámetro que indica el balance deseado entre margen y error.

# Problema Dual


El Lagrangiano:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 + \color{red}{C} \sum_{i=1}^n \zeta_i - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) - \sum_{i=1}^n \mu_i \zeta_i$$

# Problema Dual

El Lagrangiano:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^n \zeta_i - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) - \sum_{i=1}^n \mu_i \zeta_i$$


resulta en el problema dual: 

$$\begin{aligned} & \text{máx} && \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ & \text{sujeto a} && \sum_{i=1}^n \alpha_i y_i = 0 \\ & && 0 \leq \alpha_i \leq C \end{aligned}$$

# Problema Dual

El Lagrangiano:

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 + \sum_{i=1}^n \zeta_i - \sum_{i=1}^n \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) - \sum_{i=1}^n \mu_i \zeta_i$$

resulta en el problema dual: 

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Unico cambio en el dual es cota superior en los multiplicadores  $\alpha_i$ .

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$\textcircled{1} \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$



# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$
- ❹  $\alpha_i + \mu_i = C$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$
- ❹  $\alpha_i + \mu_i = C$
- ❺  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$
- ❹  $\alpha_i + \mu_i = C$
- ❺  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$
- ❻  $\sum_{i=1}^n \alpha_i y_i = 0$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$
- ❹  $\alpha_i + \mu_i = C$
- ❺  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$
- ❻  $\sum_{i=1}^n \alpha_i y_i = 0$
- ❼  $\alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) = 0$

# Condiciones de Karush-Kuhn-Tucker (KKT)

- ❶  $y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$
- ❷  $\zeta_i \geq 0$
- ❸  $\mu_i \geq 0$
- ❹  $\alpha_i + \mu_i = C$
- ❺  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$
- ❻  $\sum_{i=1}^n \alpha_i y_i = 0$
- ❼  $\alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) = 0$
- ❽  $\zeta_i \mu_i = 0$

# Condiciones de Karush-Kuhn-Tucker (KKT)

$$① \quad y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i \geq 0 \quad i = 1, \dots, n$$

$$② \quad \zeta_i \geq 0$$

$$③ \quad \mu_i \geq 0$$

$$④ \quad \alpha_i + \mu_i = C$$

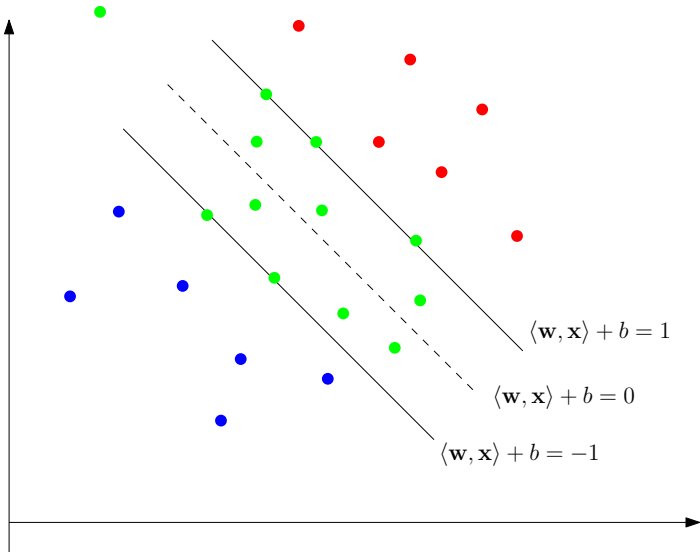
$$⑤ \quad \mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \mathbf{x}_i$$

$$⑥ \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$⑦ \quad \alpha_i (y_i (\langle \mathbf{w}, \mathbf{x}_i \rangle + b) - 1 + \zeta_i) = 0$$

$$⑧ \quad \zeta_i \mu_i = 0$$

- Ahora los vectores de soporte incluyen vectores para los cuales  $\alpha_i = C$





# Caso no lineal

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\mathcal{X} \rightarrow \mathcal{H}$$

$$\mathbf{x} \mapsto \phi(\mathbf{x})$$

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\begin{aligned}\mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x})\end{aligned}$$

Donde  $\mathcal{H}$  es un espacio de **Hilbert**:

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\begin{aligned}\mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x})\end{aligned}$$

Donde  $\mathcal{H}$  es un espacio de **Hilbert**:

- ▶ Producto punto.

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\begin{aligned}\mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x})\end{aligned}$$

Donde  $\mathcal{H}$  es un espacio de **Hilbert**:

- ▶ Producto punto.
- ▶ Completo.

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\begin{aligned}\mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x})\end{aligned}$$

Donde  $\mathcal{H}$  es un espacio de **Hilbert**:

- ▶ Producto punto.
- ▶ Completo.
- Operamos en el **espacio de características**  $\mathcal{H}$ :

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \longrightarrow \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$$

## Caso no lineal

- **Idea:** Proyectar datos a un espacio donde sean más fácilmente separables.

$$\begin{aligned}\mathcal{X} &\rightarrow \mathcal{H} \\ \mathbf{x} &\mapsto \phi(\mathbf{x})\end{aligned}$$

Donde  $\mathcal{H}$  es un espacio de **Hilbert**:

- ▶ Producto punto.
- ▶ Completo.
- Operamos en el **espacio de características**  $\mathcal{H}$ :

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle \longrightarrow \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$$

- $\mathcal{X}$  puede ser un conjunto arbitrario.



# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b)$$

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b) = \text{sign} \left( \sum_{i=1}^n (\alpha_i y_i) \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \right)$$

# Problema de optimización

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b) = \text{sign} \left( \sum_{i=1}^n (\alpha_i y_i) \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \right)$$

- Datos sólo aparecen en términos de **productos internos**.

## Ejemplo

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

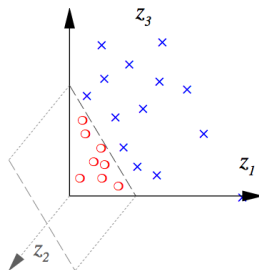
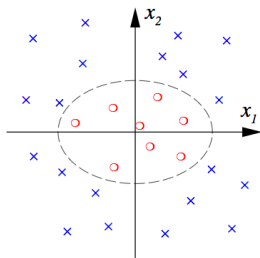
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



# Ejemplo

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

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



# Kernels

- Suponga que existe un **kernel**:

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$

$$(\mathbf{x}_1, \mathbf{x}_2) \mapsto k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$$

# Kernels

- Suponga que existe un **kernel**:

$$k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$$
$$(\mathbf{x}_1, \mathbf{x}_2) \mapsto k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}}$$

- Podemos calcular el producto punto  $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$  operando en el **conjunto original**  $\mathcal{X}$ .

# Kernels

- Suponga que existe un **kernel**:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R} \\ (\mathbf{x}_1, \mathbf{x}_2) &\mapsto k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}} \end{aligned}$$

- Podemos calcular el producto punto  $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$  operando en el **conjunto original**  $\mathcal{X}$ .
- Más aún, no requerimos conocer el mapeo  $\phi$ , o el espacio  $\mathcal{H}$ .

# Kernels

- Suponga que existe un **kernel**:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R} \\ (\mathbf{x}_1, \mathbf{x}_2) &\mapsto k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}} \end{aligned}$$

- Podemos calcular el producto punto  $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$  operando en el **conjunto original**  $\mathcal{X}$ .
- Más aún, no requerimos conocer el mapeo  $\phi$ , o el espacio  $\mathcal{H}$ .
- De hecho  $\mathcal{H}$  puede tener dimensión infinita.

# Kernels

- Suponga que existe un **kernel**:

$$\begin{aligned} k : \mathcal{X} \times \mathcal{X} &\rightarrow \mathbb{R} \\ (\mathbf{x}_1, \mathbf{x}_2) &\mapsto k(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x}_1), \phi(\mathbf{x}_2) \rangle_{\mathcal{H}} \end{aligned}$$

- Podemos calcular el producto punto  $\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}}$  operando en el **conjunto original**  $\mathcal{X}$ .
- Más aún, no requerimos conocer el mapeo  $\phi$ , o el espacio  $\mathcal{H}$ .
- De hecho  $\mathcal{H}$  puede tener dimensión infinita.
- Cómo sabemos si una función  $k(.,.)$  es un **kernel**?

# El truco del Kernel

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

# El truco del Kernel

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$



# El truco del Kernel

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

# El truco del Kernel

$$\text{máx} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{sujeto a} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$

# El truco del Kernel

$$\text{máx} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{sujeto a} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

# El truco del Kernel

$$\text{máx} \quad \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j)$$

$$\text{sujeto a} \quad \sum_{i=1}^n \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b)$$

# El truco del Kernel

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b) = \text{sign} \left( \sum_{i=1}^n (\alpha_i y_i) \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \right)$$

# El truco del Kernel

$$\begin{aligned} \text{máx} \quad & \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{sujeto a} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C \end{aligned}$$

- Solución:  $\mathbf{w} = \sum_{i=1}^n (\alpha_i y_i) \phi(\mathbf{x}_i)$
- Evaluación:

$$\begin{aligned} \text{sign}(\langle w, \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b) &= \text{sign} \left( \sum_{i=1}^n (\alpha_i y_i) \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}) \rangle_{\mathcal{H}} + b \right) \\ &= \text{sign} \left( \sum_{i=1}^n (\alpha_i y_i) k(\mathbf{x}_i, \mathbf{x}) + b \right) \end{aligned}$$

## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

Comprobando:



## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

Comprobando:

$$\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} = \left\langle (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2), (x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2) \right\rangle$$

## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

Comprobando:

$$\begin{aligned}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} &= \left\langle (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2), (x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2) \right\rangle \\ &= x_{i1}^2x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2x_{j2}^2\end{aligned}$$

## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

Comprobando:

$$\begin{aligned}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} &= \left\langle (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2), (x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2) \right\rangle \\ &= x_{i1}^2x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2x_{j2}^2 \\ &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2\end{aligned}$$

## Ejemplo

- Si  $\mathcal{X} = \mathbb{R}^2$ ,  $\mathcal{H} = \mathbb{R}^3$ ,  $\phi(\mathbf{x}) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)$ , tenemos el kernel:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$$

Comprobando:

$$\begin{aligned}\langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle_{\mathcal{H}} &= \left\langle (x_{i1}^2, \sqrt{2}x_{i1}x_{i2}, x_{i2}^2), (x_{j1}^2, \sqrt{2}x_{j1}x_{j2}, x_{j2}^2) \right\rangle \\ &= x_{i1}^2x_{j1}^2 + 2x_{i1}x_{i2}x_{j1}x_{j2} + x_{i2}^2x_{j2}^2 \\ &= (x_{i1}x_{j1} + x_{i2}x_{j2})^2 \\ &= \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2\end{aligned}$$

- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.

- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.
- Por ejemplo, para  $\mathcal{X} = \mathbb{R}^2$  y el kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$ , podemos tener:

- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.
- Por ejemplo, para  $\mathcal{X} = \mathbb{R}^2$  y el kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$ , podemos tener:
  - ▶  $\mathcal{H} = \mathbb{R}^4$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.
- Por ejemplo, para  $\mathcal{X} = \mathbb{R}^2$  y el kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$ , podemos tener:
  - ▶  $\mathcal{H} = \mathbb{R}^4$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$





- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.
- Por ejemplo, para  $\mathcal{X} = \mathbb{R}^2$  y el kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$ , podemos tener:
  - ▶  $\mathcal{H} = \mathbb{R}^4$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$



- ▶  $\mathcal{H} = \mathbb{R}^2$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} \frac{x_1^2 - x_2^2}{2x_1 x_2} \\ x_1^2 + x_2^2 \end{pmatrix}$$

- El mapeo  $\phi$  y el espacio  $\mathcal{H}$  correspondientes a un kernel **no son únicos**.
- Por ejemplo, para  $\mathcal{X} = \mathbb{R}^2$  y el kernel  $k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^2$ , podemos tener:
  - ▶  $\mathcal{H} = \mathbb{R}^4$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$



- ▶  $\mathcal{H} = \mathbb{R}^2$  y

$$\phi(\mathbf{x}) = \begin{pmatrix} \frac{x_1^2 - x_2^2}{2x_1 x_2} \\ x_1^2 + x_2^2 \end{pmatrix}$$



# Ejemplos de kernels

# Ejemplos de kernels

- Polinomial:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d$$

# Ejemplos de kernels

- Polinomial:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d$$

- Polinomial no homogéneo:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + c)^d$$

# Ejemplos de kernels

- Polinomial:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d$$

- Polinomial no homogéneo:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + c)^d$$

- Gaussiano (RBF):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp \left( -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right)$$

# Ejemplos de kernels

- Polinomial:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle^d$$

- Polinomial no homogéneo:

$$k(\mathbf{x}_i, \mathbf{x}_j) = (\langle \mathbf{x}_i, \mathbf{x}_j \rangle + c)^d$$

- Gaussiano (RBF):

$$k(\mathbf{x}_i, \mathbf{x}_j) = \exp \left( -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2\sigma^2} \right)$$

- Sigmoidal:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \tanh(a \langle \mathbf{x}_i, \mathbf{x}_j \rangle + b)$$

Cómo sabemos si una función  $k(.,.)$  es un **kernel**?



Cómo sabemos si una función  $k(.,.)$  es un **kernel**?

Dos aproximaciones:

# Cómo sabemos si una función $k(.,.)$ es un **kernel**?

Dos aproximaciones:

- Reproducing Kernel Hilbert Spaces (RKHS)

# Cómo sabemos si una función $k(.,.)$ es un **kernel**?

Dos aproximaciones:

- Reproducing Kernel Hilbert Spaces (RKHS) (Aronszajn, 1950)

# Cómo sabemos si una función $k(.,.)$ es un **kernel**?

Dos aproximaciones:

- Reproducing Kernel Hilbert Spaces (RKHS) (Aronszajn, 1950)
- Teorema de Mercer

# Cómo sabemos si una función $k(.,.)$ es un **kernel**?

Dos aproximaciones:

- Reproducing Kernel Hilbert Spaces (RKHS) (Aronszajn, 1950)
- Teorema de Mercer (Mercer, 1911).

## Definición

Dada una función  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  y  $x_1, x_2, \dots, x_n \in \mathcal{X}$ , la matriz  $n \times n$  con entradas

$$K_{ij} = k(x_i, x_j)$$

se llama la *matriz de Gram* de  $k$  con respecto a  $x_1, x_2, \dots, x_n$ .

## Definición

Dada una función  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  y  $x_1, x_2, \dots, x_n \in \mathcal{X}$ , la matriz  $n \times n$  con entradas

$$K_{ij} = k(x_i, x_j)$$

se llama la *matriz de Gram* de  $k$  con respecto a  $x_1, x_2, \dots, x_n$ .

## Definición

Una función  $k : \mathcal{X}^2 \rightarrow \mathbb{R}$  para la cual para todo  $n \in \mathbb{N}$ , y todo  $x_1, x_2, \dots, x_n \in \mathcal{X}$  resulta en una matriz de Gram *positiva semidefinida* es un *kernel positivo definido* (o simplemente un kernel).

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .



# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\phi : \mathcal{X} \rightarrow \mathbb{R}^{\mathcal{X}}$$

.

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

- Receta:

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

- Receta:
  - 1 Imagen de  $\phi \longrightarrow$  espacio vectorial.

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

- Receta:
  - 1 Imagen de  $\phi \longrightarrow$  espacio vectorial.
  - 2 Producto punto.

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

- Receta:
  - 1 Imagen de  $\phi \rightarrow$  espacio vectorial.
  - 2 Producto punto.
  - 3  $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$

# El mapa del kernel reproductor

- Kernel positivo definido  $k$ .
- Mapeo:

$$\begin{aligned}\phi : \mathcal{X} &\rightarrow \mathbb{R}^{\mathcal{X}} \\ x &\mapsto k(., x).\end{aligned}$$

- Receta:
  - 1 Imagen de  $\phi \rightarrow$  espacio vectorial.
  - 2 Producto punto.
  - 3  $k(x_i, x_j) = \langle \phi(x_i), \phi(x_j) \rangle_{\mathcal{H}}$
  - 4 Completar espacio.

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$



- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto:

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto: Si  $g(.) = \sum_{j=1}^m \beta_j k(., x'_j)$  definimos:

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto: Si  $g(.) = \sum_{j=1}^m \beta_j k(., x'_j)$  definimos:

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j)$$

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto: Si  $g(.) = \sum_{j=1}^m \beta_j k(., x'_j)$  definimos:

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j)$$

- Note que: t

$$\langle f, g \rangle = \sum_{j=1}^m \beta_j f(x'_j)$$

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto: Si  $g(.) = \sum_{j=1}^m \beta_j k(., x'_j)$  definimos:

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j)$$

- Note que: t

$$\langle f, g \rangle = \sum_{j=1}^m \beta_j f(x'_j) = \sum_{i=1}^n \alpha_i g(x_i)$$

- Vectores:

$$f(.) = \sum_{i=1}^n \alpha_i k(., x_i) \quad n \in \mathbb{N}, \alpha_i \in \mathbb{R}, x_1, \dots \in \mathcal{X}$$

- Producto punto: Si  $g(.) = \sum_{j=1}^m \beta_j k(., x'_j)$  definimos:

$$\langle f, g \rangle = \sum_{i=1}^n \sum_{j=1}^m \alpha_i \beta_j k(x_i, x'_j)$$

- Note que: t

$$\langle f, g \rangle = \sum_{j=1}^m \beta_j f(x'_j) = \sum_{i=1}^n \alpha_i g(x_i)$$