



# El aprendizaje de métricas de distancia

Trabajo de Fin de Grado

Juan Luis Suárez Díaz

Tutores:

Francisco Herrera Triguero

Salvador García López

## 1 Introducción

- Objetivos
- Descripción del problema
- Aplicaciones

## 2 Matemáticas

# Objetivos

- Conocer la disciplina del aprendizaje de métricas de distancia.
- Estudiar los fundamentos matemáticos del aprendizaje de métricas de distancia.
- Analizar los principales algoritmos de aprendizaje de métricas de distancia.
- Desarrollar un software que integre los algoritmos de aprendizaje estudiados.

# El aprendizaje de métricas de distancia

## ¿Qué es?

Es una rama del aprendizaje automático cuya finalidad es aprender distancias a partir de los datos.

## Definición

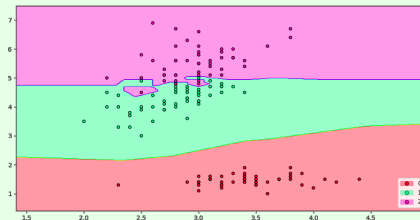
Sea  $X$  un conjunto no vacío. Una distancia sobre  $X$  es una aplicación  $d: X \times X \rightarrow \mathbb{R}$ , verificando:

- ①  $d(x, y) = 0 \iff x = y$  para cualesquiera  $x, y \in X$  (coincidencia)
  - ②  $d(x, y) = d(y, x)$  para cualesquiera  $x, y \in X$  (simetría)
  - ③  $d(x, z) \leq d(x, y) + d(y, z)$  para cualesquiera  $x, y, z \in X$  (desigualdad triangular).
- **Pseudodistancias:** Exigen solo  $d(x, x) = 0$  en ①.

# ¿Por qué distancias?

## Ejemplo

*Los clasificadores de vecinos cercanos*



## Problema

- Los algoritmos basados en distancias suelen utilizar distancias fijas.
- **Solución:** Aprender distancias.

## ¿Cómo aprender una distancia?

### Definición (*Distancias de Mahalanobis*)

Sea  $M \in \mathcal{M}_d(\mathbb{R})$  semidefinida positiva. Entonces,  $d_M: \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ , dada por

$$d(x, y) = \sqrt{(x - y)^T M (x - y)}$$

es una (pseudo-)distancia, denominada distancia de Mahalanobis.

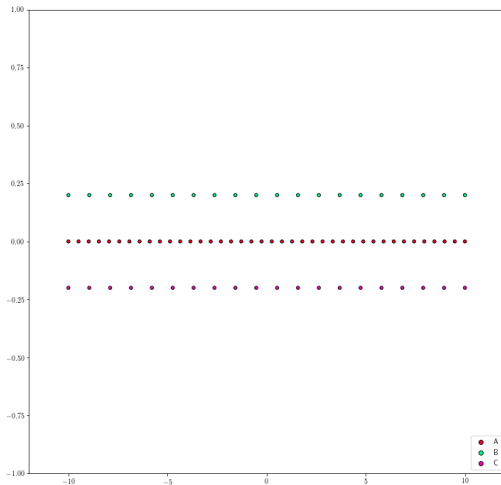
**Enfoque principal del aprendizaje de métricas de distancia**  
**Aprender distancias de Mahalanobis sobre espacios vectoriales  $d$ -dimensionales.**

### Dos opciones:

- 1 Aprender  $M$ .
- 2 Aprender una aplicación lineal  $L$ .  
Entonces,  $M = L^T L$  y  $d_M(x, y)^2 = \|L(x - y)\|_2^2$ .

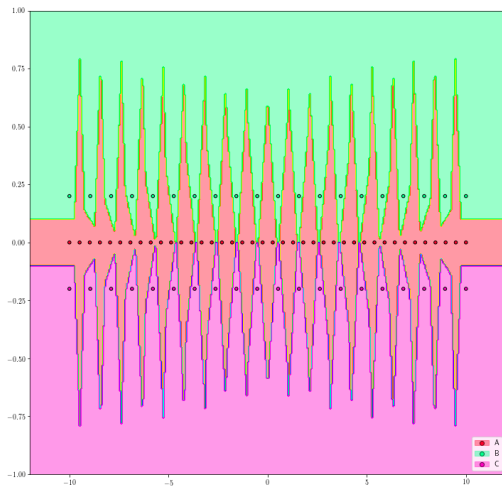
# Mejora de clasificadores basados en distancias

1-NN



# Mejora de clasificadores basados en distancias

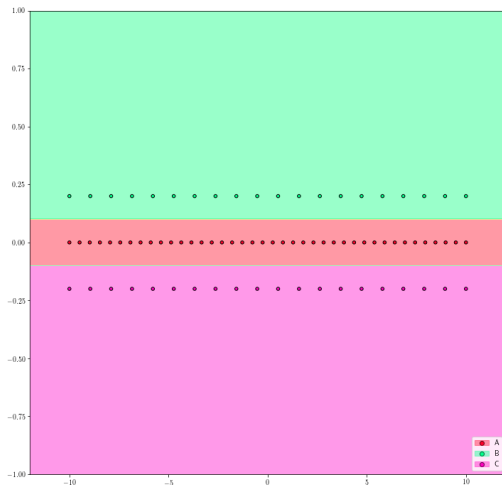
$$M = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$





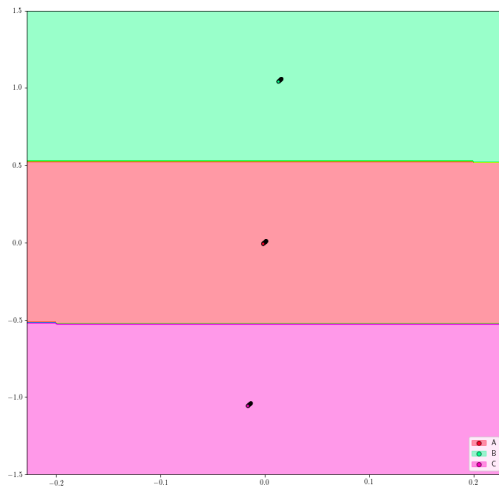
# Mejora de clasificadores basados en distancias

$$M \approx \begin{pmatrix} 0 & -0.004 \\ -0.004 & 27.5 \end{pmatrix}$$

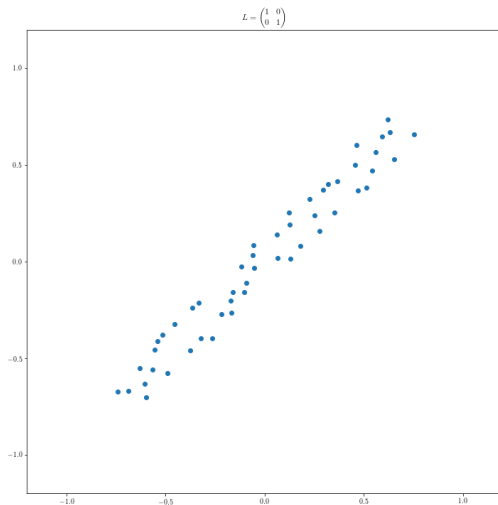


# Mejora de clasificadores basados en distancias

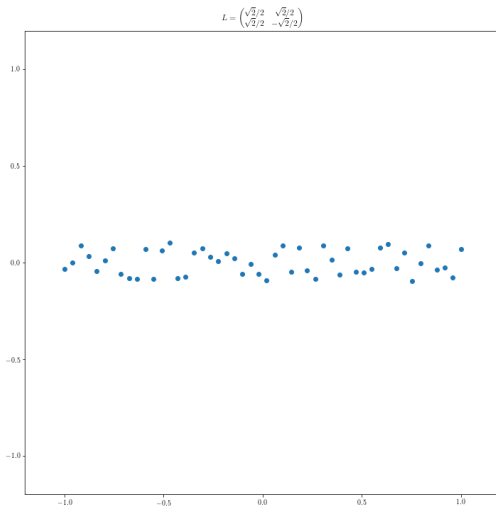
$$L \approx \begin{pmatrix} -0.0001 & 0.073 \\ -0.0008 & 5.24 \end{pmatrix}$$



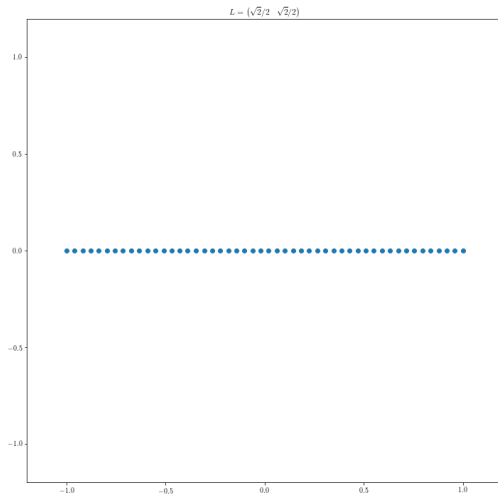
# Organización de datos y reducción de dimensionalidad



# Organización de datos y reducción de dimensionalidad



# Organización de datos y reducción de dimensionalidad



## 1 Introducción

## 2 Matemáticas

- Análisis convexo
- Análisis matricial
- Teoría de la información

### Las matemáticas bajo el aprendizaje de métricas de distancia

- ① **Análisis convexo.** De gran importancia en la mayoría de algoritmos de aprendizaje de métricas de distancia.
- ② **Análisis matricial.** Las matrices son la herramienta fundamental para modelar el problema.
- ③ **Teoría de la información.** Presente en algunos de los algoritmos.

# Hiperplanos soporte



A map of the Iberian Peninsula is shown with several colored regions: green in the northwest, yellow in the north-central area, light blue in the northeast, pink in the southwest, and red in the south. Numerous small, colored circles are scattered across the map, forming distinct clusters that correspond to the colored regions. A central white box with a black border contains text.

**¡Gracias por su atención!**

**Juan Luis Suárez Díaz**

[jlsuarezdiaz@correo.ugr.es](mailto:jlsuarezdiaz@correo.ugr.es)