

# 4 - KNN AND EVALUATION

Aprendizagem 2024/2025

## KNN

- Classificar uma nova observação com base nos  $k$  vizinhos mais próximos
- Vizinhos podem ser calculados com base em **distâncias** (queremos o mais perto) ou **semelhanças** (queremos o mais semelhante)

## DISTÂNCIAS E SEMELHANÇAS

- **Minkowski:**  $d(x_a, x_b) = \sqrt[q]{|x_{a1} - x_{b1}|^q + \dots + |x_{am} - x_{bm}|^q}$
- **Manhattan:** Minkowski com  $q = 1$
- **Euclidian:** Minkowski com  $q = 2$
- **Chebyshev:**  $d(x_a, x_b) = \max(|x_{a1} - x_{b1}|, \dots, |x_{am} - x_{bm}|)$
- **Cosine similarity:**  $d(x_a, x_b) = \frac{x_a \cdot x_b}{\|x_a\| \|x_b\|}$

## MEDIDAS DE AGREGAÇÃO

- Variáveis **categóricas**: moda, moda ponderada
- Variáveis **numéricas**: média, média ponderada

## ERROS

- **Mean absolute error:**  $MAE = \frac{1}{n} \sum |y - y'|$
- **Mean squared error:**  $MSE = \frac{1}{n} \sum (y - y')^2$
- **Root mean squared error:**  $RMSE = \sqrt{MSE}$

## ROC E AUC

- **ROC**: representa a performance de um classificador binário para cada valor de threshold (TPR no eixo do y, FPR no eixo do x)

$$TPR = \frac{TP}{P}, \quad FPR = \frac{FP}{N}$$

- **AUC** (area under curve): representa a probabilidade do modelo distinguir entre as duas classes

$$AUC = 1 \text{ (perfeita)}, \quad AUC = 0.5 \text{ (random)}$$

# SUMÁRIO

- Ficha 4