9 - CLUSTERING I

Aprendizagem 2024/2025

CLUSTERING

- **Agrupar observações** em clusters com base nas suas similaridades
- Queremos clusters coesos e separados

ALGORITMOS

- **k-means** (algoritmo baseado em partição, clusters pontos):
  - (a) Atribuir pontos a clusters
  - (b) Ajustar centroides
  - (c) Verificar convergência
- **EM** (algoritmo baseado em modelo, clusters distribuições):
  - (a) E-step, atribuir pontos ao cluster respetivo
  - (b) M-step, recalcular os parâmetros

AVALIAÇÃO

- **Silhueta** (critério interno; avalia a separação + coesão; pode ser feito ao nível do ponto, cluster ou solução): $S(x1) = 1 - \frac{a(x1)}{b(x1)}$, onde $a(x1)$ representa a distância média aos pontos do cluster, e b$(x1)$ a distância média aos pontos do cluster mais próximo
- **Puridade** (critério externo; quão bem a solução agrupa observações com a mesma classe no mesmo cluster): $\frac{1}{P}\sum_i \phi(c_i)$, onde P é o número de pontos

AVALIAÇÃO COESÃO E SEPARAÇÃO

- **Distância intra-cluster** (avalia a coesão): soma da distância de todos os pontos relativamente ao seu cluster
- **Distância inter-cluster** (avalia a separação):

$$\frac{1}{K^2}\sum_i \sum_j d(u_i, u_j), \text{ para } K \text{ clusters}$$

SUMÁRIO

- Ficha 10: 1, 2, 3