K-means

Prof. Hugo de Paula

Análise de agrupamento: *Clustering*

Cluster

- Coleção de objetos que são similares uns aos outros (de acordo com algum critério de similaridade pré-fixado) e dissimilares a objetos pertencentes a outros clusters.

Análise de agrupamento (*clustering*)

- Separa os objetos em grupos com base na similaridade, e em seguida atribuir rótulos a cada grupo.

- Qualidade do resultado depende da medida da similaridade usada pelo método.

Aplicações de *Clustering*

- Distribuição e pré-processamento de dados.
- Proc. de imagens (segmentação); economia; marketing.
- WWW (Classificação de documentos, padrões de acesso)
- Agricultura (áreas de uso de terra); planejamento de cidades (agrupar casas de acordo com tipos, valores e localização).

Principais métodos de clusterização

- Métodos baseados em particionamento
- Métodos baseados em densidade
- Métodos hierárquicos

Algoritmo de Particionamento: K-means (MacQueen'67)

- Cada cluster é representado por um ponto central
- *K* é a quantidade de clusters desejada
- Variações: k-medóides, k-modas, k-medianas
- Requer uma medida de distância, e a possibilidade de se calcular médias entre os objetos
- Pode encontrar mínimos locais: solução é o *random restart*
- Pode entrar em loop infinito: solução é limitar número de iterações

K-means: exemplo

Base de dados = {2,4,10,12,3,20,30,11,25}, k=2

Centros iniciais, escolhidos aleatoriamente: m1 = 3, m2 = 4

1ª iteração: K1 = {2, 3}; m1 = 2.5; K2 = {4, 10, 12, 20, 30, 11, 25}; m2 = 16

2ª iteração: K1 = {2, 3, 4}; m1 = 3; K2 = {10, 12, 20, 30, 11, 25}; m2 = 18

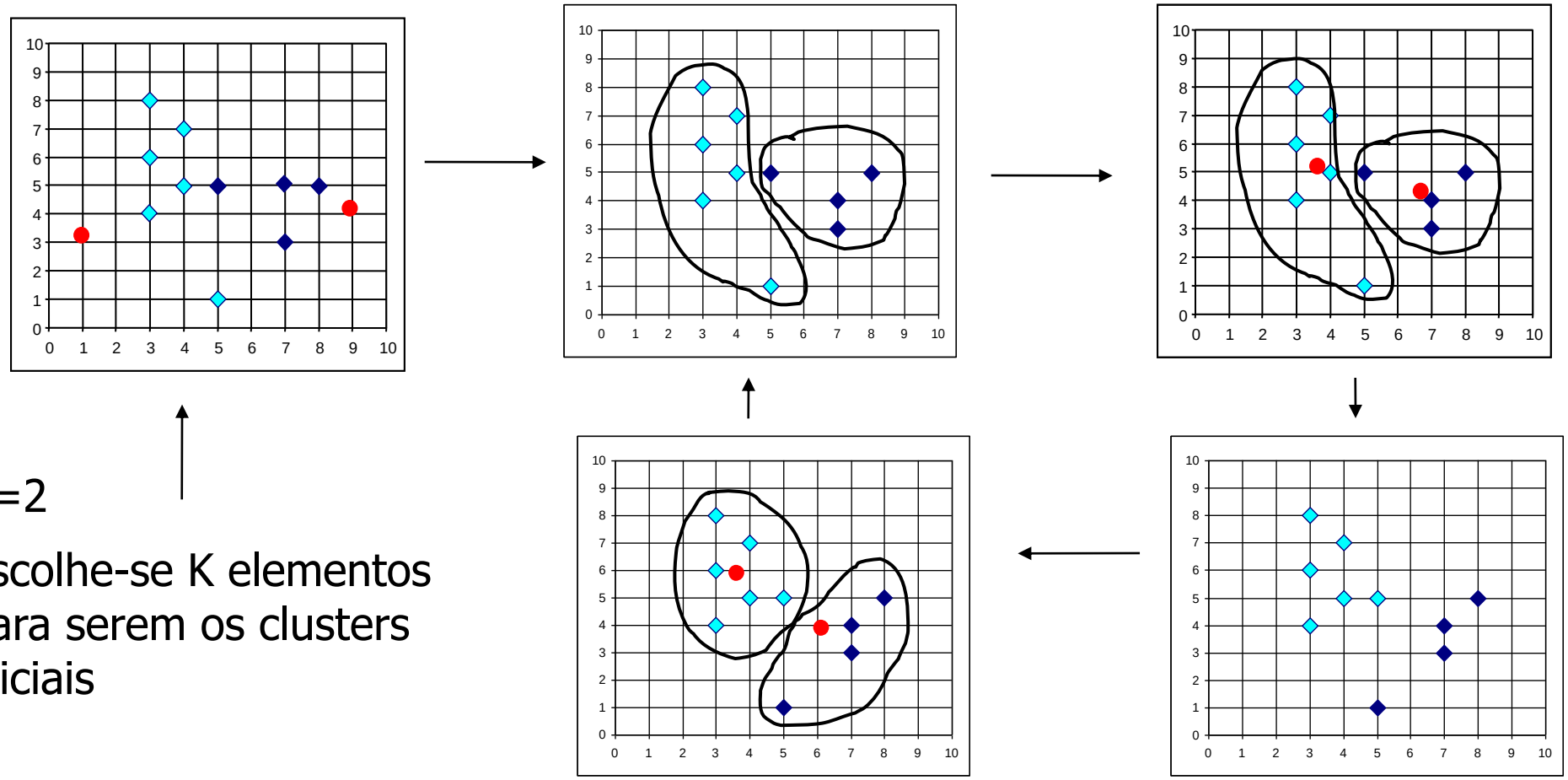
3ª iteração: K1 = {2, 3, 4, 10}; m1 = 4.75; K2 = {12, 20, 30, 11, 25}; m2 = 19.6

4ª iteração: K1 = {2, 3, 4, 10, 11, 12}; m1 = 7; K2 = {20, 30, 25}; m2 = 25

5ª iteração: K1 = {2, 3, 4, 10, 11, 12}; m1 = 7; K2 = {20, 30, 25}; m2 = 25

Sem alteração em relação à quarta iteração, fim do processamento

K-means (loop infinito)



K=2

Escolhe-se K elementos para serem os clusters iniciais

PUC Minas
Virtual