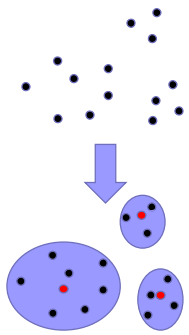
Clustering (agrupamento)

Victor Lobo

1

Clustering

- Objectivo fundamental
  - Definir agrupamentos de dados que têm **algo em comum ou parecido**
    - Sumarizar
    - Detectar grandes grupos / detectar outliers
    - Facilitar a gestão de múltiplas entidades



- Também conhecido como:
  - Técnicas de Agrupamento
  - Na comunidade de estatística: Classificação (dividir em classes)

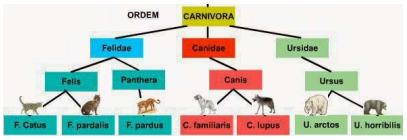
2

Medidas de qualidade

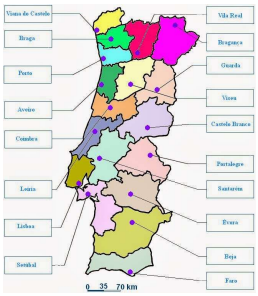
- Clustering, no seu conceito mais geral é um problema “ill-posed”
  - Não há métricas universais para clustering
- Medidas comuns
  - **Minimizar** o somatório das distâncias **intra-cluster**, e **maximizar** o somatório das distâncias **inter-clusters**
  - Definir um número fixo de **CENTROIDES**, e minimizar o somatório da distância entre os dados e o centroide mais próximo

3

Exemplos



- Agrupar os animais por raça/espécie/família/ordem classe



- Dividir o país em distritos
- Dividir os cadetes em turmas de inglês

401	402	405	…
412	403	409	…
413	415	417	…
423	420	418	…
431	432	428	…

4

Tipos de técnicas

- **Hierarquicas**
  - Sub-divisões cada vez mais detalhadas
    - "Divisivas" – Começa no todo, e vai dividindo
    - "Aglomerativas" – Vai juntando grupos cada vez maiores
  - *Dendogramas*, métodos de Ward, etc
- **Partições**
  - Divide o espaço em blocos sem relações hierárquicas
  - *K-Médias*, Fuzzy C-mean, DBSCAN,SOM, etc
- Outras (por densidade, por grelha, etc)

5

Algoritmo "k-médias" de Lloyd

- Algoritmo de Lloyd
  1. Escolher aleatoriamente *k* centroides μ_k
  2. Para cada exemplo x_j, encontrar o centroide μ_k mais próximo, e atribuir-lhe a classe C_k.
  3. Recalcular os centroides de cada classe
     $\mu_k = \Sigma x_i / n_k : x_i \in C_k$
  4. Voltar a 2, até que não haja alterações em μ_k
- Exemplo

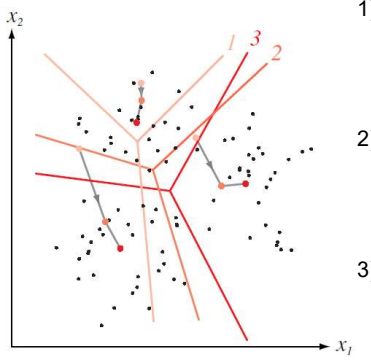
6

Variantes

- Algoritmo de McQueen
  - actualizações incrementais
- K-medoids
  - Os centroids têm que ser pontos que existam nos dados originais
- Fuzzy c-Means
  - As pertenças aos centroids são funções fuzzy

7

Exemplo



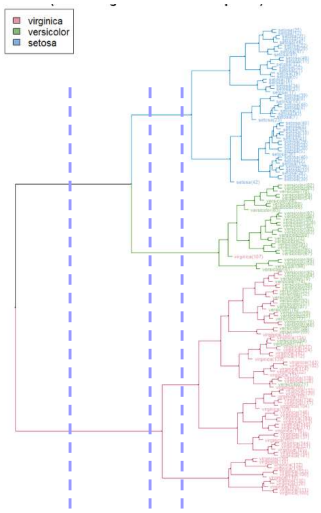
1) Para os centroides originais (a vermelho claro), as fronteiras são apresentadas em 1. Se calcularmos a media desses pontos, obtemos os centroides de 2.
2) Para o 2º conjunto de centroides (a vermelho médio), as fronteiras são apresentadas em 2. Se calcularmos a media desses pontos, obtemos os centroides de 3.
3) Para o 3º conjunto de centroides (a vermelho vivo), as fronteiras são apresentadas em 3. Ao calcularmos os novos centroides, as fronteiras não mudam.

8

8

Classificadores hierarquicos

- Critérios para associar dados
  - Distância a um do grupo
  - Distância à média do grupo
  - Juntar sempre 2 a 2 grupos equivalentes
- Para definir os clusters
  - Escolher um nível de corte
- Exemplo para o Iris-Dataset



9