Agrupamento (*Clustering*)

Introdução

Classificar coisas semelhantes em classes ou grupos é uma atividade básica do processo de conhecimento humano.

Classificação supervisionada

x

Classificação não supervisionada

Classificação supervisionada: Distribuem ou alocam os objetos em classes preestabelecidas (Análise Discriminante, modelos Logit e Probit).

Classificação não supervisionada: Fracionam um conjunto de objetos em subconjuntos homogêneos (Análise de Agrupamentos).

Introdução

- A necessidade de agrupar elementos por suas características está presente em várias áreas do conhecimento, como nas ciências biológicas, ciências sociais e comportamentais, ciências da terra, medicina, informática, entre outras.
- Tendo em vista a dificuldade de se examinar todas as combinações de grupos possíveis em um grande volume de dados, foram desenvolvidas diversas técnicas capazes de auxiliar na formação dos agrupamentos.
- A análise de agrupamentos (cluster) busca agrupar elementos de dados baseando-se na similaridade entre eles. Os grupos são determinados de forma a obter-se homogeneidade dentro dos grupos e heterogeneidade entre eles.

3

O que é Análise de Cluster?

Análise de Cluster é uma técnica multivariada cuja finalidade é agregar objetos com base nas características que eles possuem. O resultado são grupos que exibem máxima homogeneidade de objetos ***dentro*** de grupos e, ao mesmo tempo, máxima heterogeneidade ***entre*** os grupos.

Finalidade primária é agrupar *n* objetos caracterizados por *p* atributos, em grupos, denominados "clusters" de tal forma que os objetos em um mesmo grupo sejam semelhantes entre si, mas ao mesmo tempo diferentes dos objetos pertencentes aos outros grupos.

Padrões em um mesmo cluster devem ser similares enquanto que em diferentes clusters devem ser dissimilares.

4

Análise de Cluster – Pontos relevantes

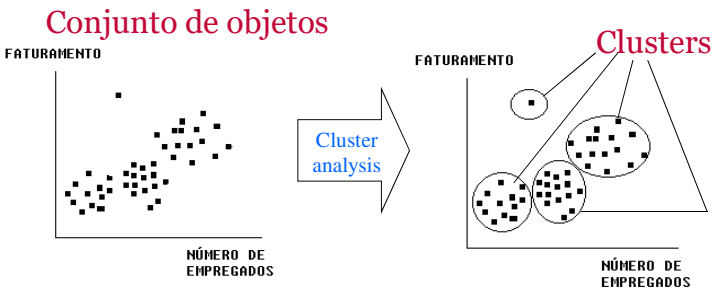
Na análise de clusters ***não há*** qualquer tipo de ***dependência entre as variáveis***: os grupos definem-se por si mesmo sem que haja uma relação causal entre as variáveis utilizadas.

Os métodos são ***exploratórios***: a ideia é ***gerar hipóteses, em vez de testá-las***. É necessária uma validação posterior dos resultados através da aplicação de outros métodos estatísticos.

Uma dificuldade inicial é a de não existir uma única via de definição de grupos, ou seja, um critério de partição ou agrupamento de indivíduos.

Análise de Cluster

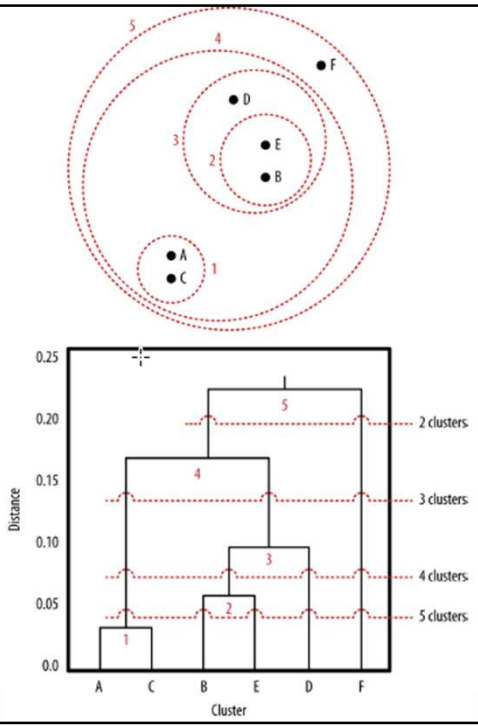
O primeiro passo da análise é definir um critério para a formação dos grupos. Um critério que parece ser razoável é considerar a proximidade entre os pontos. Pontos próximos, então, representariam regiões com comportamentos semelhantes no que se refere às variáveis do gráfico, ou seja, regiões que podem fazer parte de um mesmo grupo



grouping items by similarity

hierarchical clustering

hierarchical clustering

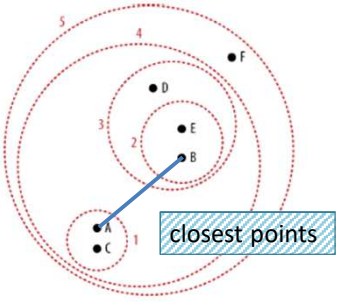


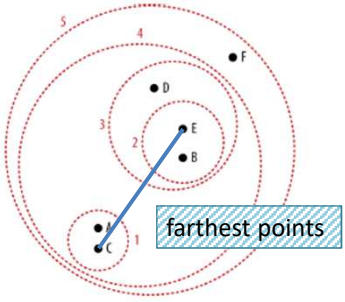
- (Provost & Fawcett, 2013:165)

how to calculate the distance between clusters?

- *linkage function*:
  - closest two points in each cluster
  - farthest two points in each cluster
  - average of distances between all points in each cluster
  - …

What's the distance between clusters 1 and 2?





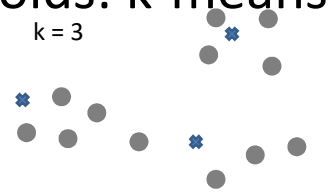
- (Provost & Fawcett, 2013:165)

k-means

clustering around centroids: k-means

1. choose a desired number of clusters *k*
2. select *k* "centers"
3. for each cluster, find the points closest to its center (assign them to the cluster)
4. recalculate center of each cluster
5. recalculate distance from each point to the cluster centers (elements may switch cluster)
6. repeat from 4 until max iterations reached OR no changes in cluster assignment

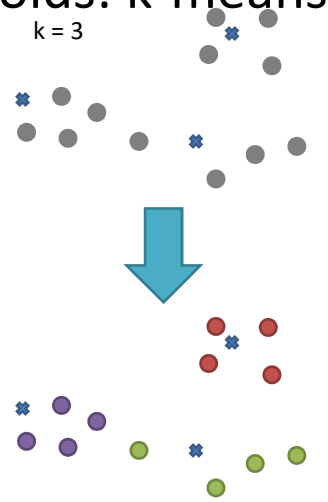
clustering around centroids: k-means



1. **choose a desired number of clusters *k***
2. **select *k* "centers"**
3. for each cluster, find the points closest to its center (assign them to the cluster)
4. recalculate center of each cluster
5. recalculate distance from each point to the cluster centers (elements may switch cluster)
6. repeat from 4 until max iterations reached OR no changes in cluster assignment

13

clustering around centroids: k-means

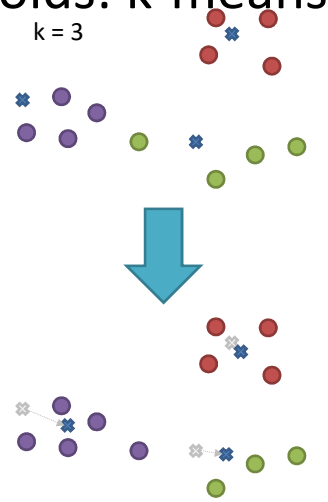


1. choose a desired number of clusters *k*
2. select *k* "centers"
3. **for each cluster, find the points closest to its center (assign them to the cluster)**
4. recalculate center of each cluster
5. recalculate distance from each point to the cluster centers (elements may switch cluster)
6. repeat from 4 until max iterations reached OR no changes in cluster assignment

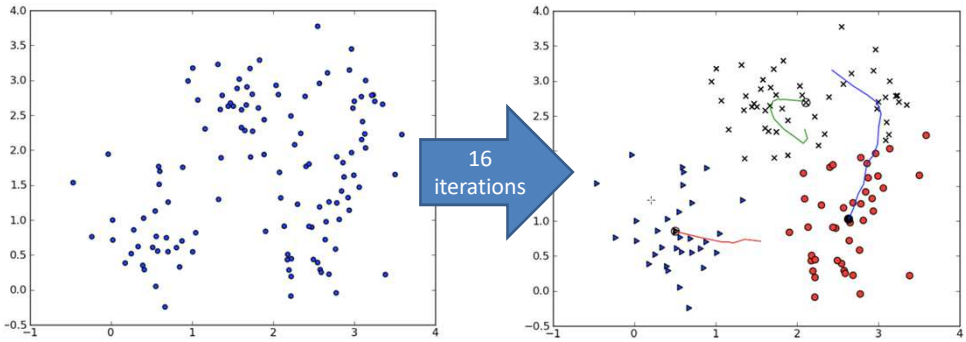
14

clustering around centroids: k-means

1. choose a desired number of clusters *k*
2. select *k* "centers"
3. for each cluster, find the points closest to its center (assign them to the cluster)
4. **recalculate center of each cluster**
5. **recalculate distance from each point to the cluster centers (elements may switch cluster)**
6. repeat from 4 until max iterations reached OR no changes in cluster assignment



k-means: example

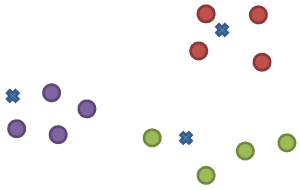


- (Provost & Fawcett, 2013:169-171)

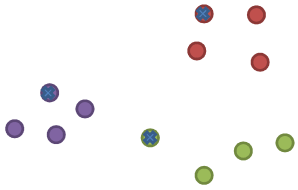
K-medoids

- Similar to k-means, but uses centroids (actual data points closest to the centers) instead of geometric centers.

k-means: centers



k-medoids: centroids

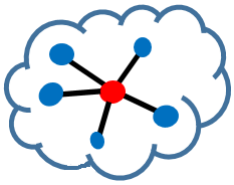


how to evaluate clustering results

- internal indices (based only on the clustering data)
- external indices (based on known values)

how to evaluate clustering results > internal indices

- **distortion**: sum of the squared distances between each data point and its corresponding centroid
  - lowest distortion → best clustering



- (Provost & Fawcett, 2013:169-172)

how to evaluate clustering results > internal indices

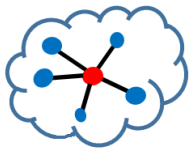
- **silhouette**:
- $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \rightarrow -1 \leq s(i) \leq 1$
  - *a(i)* = average distance between i and all other objects of the same cluster (smaller = better) *intracluster*
  - *b(i)* = lowest average distance between i to any other cluster (the "neighbouring cluster" of i)
  - average s(i): how adequate the clustering is *intercluster*

- (Provost & Fawcett, 2013:169-172)

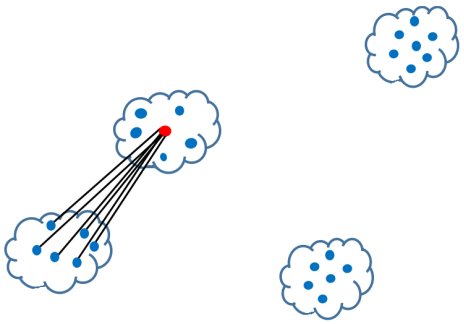
silhouette

- $s(i) = \frac{b(i)-a(i)}{\max\{a(i),b(i)\}} \rightarrow -1 \leq s(i) \leq 1$

a(i)



b(i)



- (Provost & Fawcett, 2013:169-172)

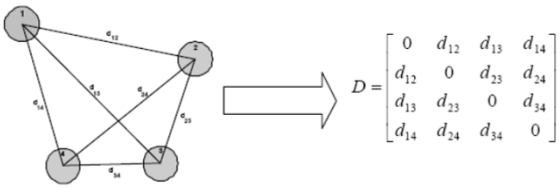
distance and similarity

Construção da matriz de similaridade ou matriz de distância

A distância entre dois pontos X_i e X_j é denotada por d_{ij}. O conjunto das distâncias duas a duas forma a matriz de distância de dimensão nxn com elementos d_{ij}

Em geral $d_{ij} = d_{ji}$ e $d_{ii} = 0$

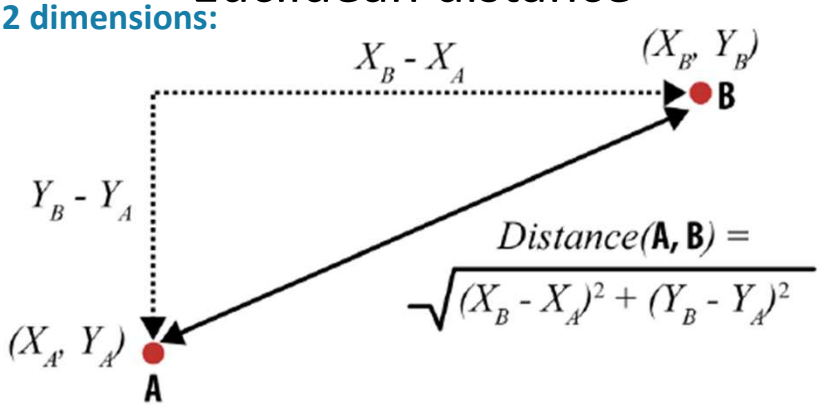
$$D = \begin{bmatrix} d_{11} & d_{12} & . & d_{1n} \\ d_{21} & d_{22} & . & d_{2n} \\ . & . & . & . \\ d_{n1} & d_{n2} & . & d_{nn} \end{bmatrix}$$



$$D = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} \\ d_{12} & 0 & d_{23} & d_{24} \\ d_{13} & d_{23} & 0 & d_{34} \\ d_{14} & d_{24} & d_{34} & 0 \end{bmatrix}$$

Euclidean distance

2 dimensions:



n dimensions:

$$d_{\text{Euclidean}}(\mathbf{X}, \mathbf{Y}) = \| \mathbf{X} - \mathbf{Y} \|_2 = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots}$$

- (Provost & Fawcett, 2013:142)

- Medidas para Dados Numéricos – Euclidiana, Manhattan
- Medidas para Dados Categóricos - Jaccard
- Medidas para Dados Mistos - Gower