Agrupamento

Carlos Diego Rodrigues

18 de janeiro de 2022

Universidade Federal do Ceará

Agrupamento

- Definição: um grupo (*cluster*) é um subconjunto de dados que possui similaridades.
- Agrupamento, também chamado aprendizado não-supervisionado, é o processo de dividir um conjunto de dados em grupos cujos elementos em cada grupo sejam tão **similares** quanto possível e cujos elementos entre grupos sejam tão **diferentes** quanto possível.
- Este processo pode revelar relações imprevistas que podem auxiliar no aprendizado supervisionado.
- Porém diversos problemas reais são caracterizados como problemas de agrupamento.

Metodologias

Dois grupos principais de algoritmos de agrupamento podem ser identificados:

- Agrupamento hierárquico
  - Aglomerativo
  - Divisivo
- Agrupamento particional
  - *K-means*
  - EM
  - Mapa auto-organizado

Características de um bom agrupamento

Um bom algoritmo de agrupamento deve possuir as seguintes características:

- Descobrir alguns ou todos os grupos escondidos
- Similaridades dentro do grupo, dissimilaridades fora do grupo.
- Lidar com vários tipos de atributos.
- Lidar com ruído, falta de dados e *outliers*.
- Lidar com alta dimensionalidade.
- Escalabilidade, interpretabilidade e usabilidade.

Agrupamento hierárquico

- Dois tipos:
  - Aglomerativo
  - Divisivo
- O agrupamento divisivo é mais raro e de difícil aplicação.
- Em geral envolve métodos próprios para determinados tipos de problema.
- Deve ser capaz de traçar as características fundamentais que separam os grupos.

Agrupamento hierárquico aglomerativo

- Baseia-se na ideia de aglutinação de grupos.
- A qualquer momento do algoritmo vamos avaliar quais são os dois grupos mais **próximos** entre si e realizar uma operação de fusão de grupos.
- Com a fusão de dois grupos, o número de grupos é reduzido em uma unidade e o algoritmo continua até que reste apenas um grupo.
- Gera como saída n agrupamentos, cada um com $i = 1, \cdots, n$ grupos.

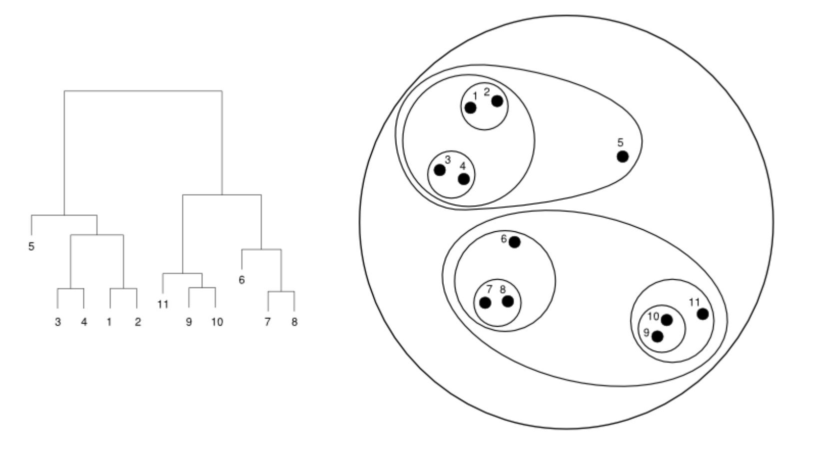
Funções de distância

- Atributos numéricos:
  - Linear: $\sum_{i=1}^{k} |x_i - y_i|$
  - Euclidiana: $\sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}$
  - Minkowski: $\left(\sum_{i=1}^{k} (|x_i - y_i|)^q\right)^{(\frac{1}{q})}$
- Atributos nominais:
  - Contagem de coincidências.
  - Distância entre categorias (ordinais).

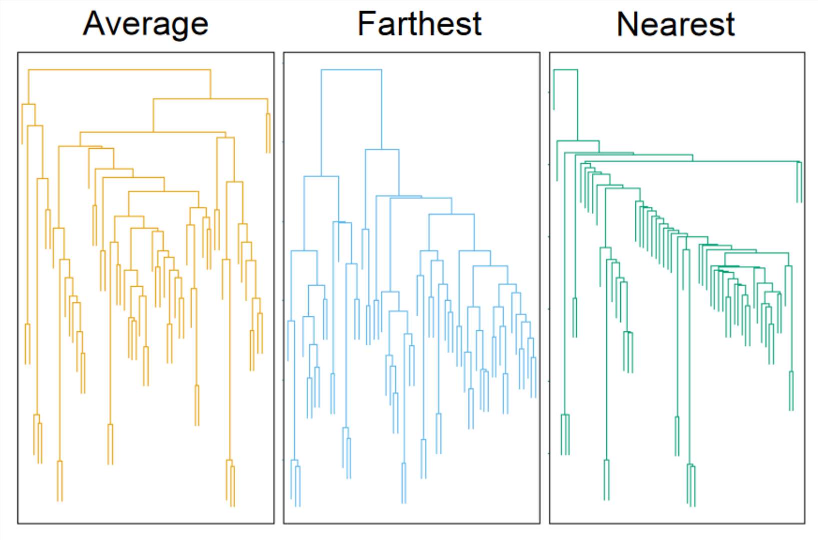
Distância entre grupos

- Porém precisamos também medir a distância não apenas entre elemento, mas também a distância entre grupos.
- Algumas medidas típicas são:
  - Menor distância entre dois elementos (ligação simples).
  - Maior distância entre dois elementos (ligação completa).
  - Média das distâncias entre os elementos (ligação média).
  - Distância entre os **centroides** de cada grupo.

Exemplo



Exemplo



Agrupamento particional: K-means

- O algoritmo *K-means* divide um conjunto de n instâncias passadas como parâmetro em k grupos, onde este valor é uma das entradas do algoritmo.
- Uma questão que acompanha este algoritmo é: qual o melhor valor de k de forma a minimizar a distância entre elementos de grupos diferentes?
- Não é conhecida uma resposta a priori, mas pode ser avaliada a posteriori conforme a execução do algoritmo para diferentes valores k.
- O algoritmo funciona através da criação de centroides artificiais para os grupos, que se deslocam de forma a minimizar a distância dos elementos associados àquele grupo.

K-means: definições

- Vamos chamar de c_r o centroide associado ao grupo r.
- Vamos também utilizar uma função de distância, assim como no agrupamento hierárquico, chamemos $D(x_i, x_j)$.
- Considere a variável α_{ri} indicando que a instância i foi associada ao grupo r. A função objetivo do algoritmo K-means é:

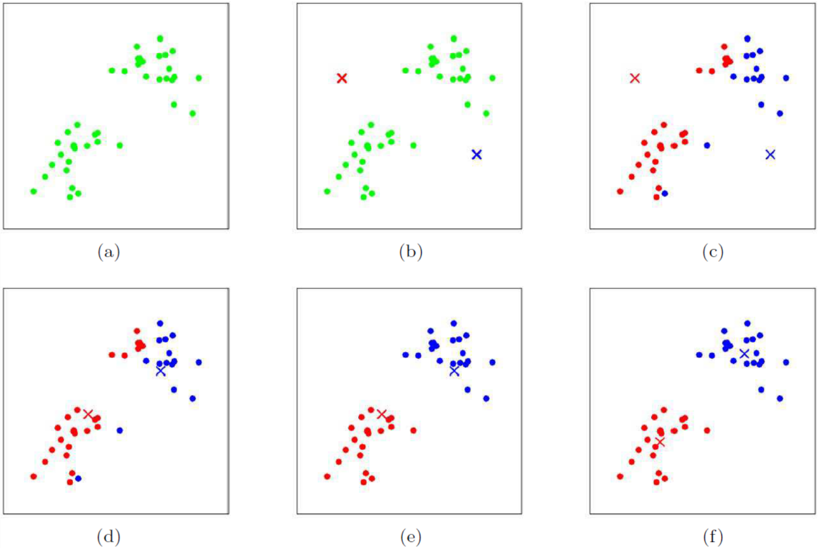
$$J(\alpha, c) = \sum_{r=1}^{k} \sum_{i=1}^{n} \alpha_{ri} \cdot D(x_i, c_r)$$

- Este problema é difícil de ser resolvido por um processo de otimização e o algoritmo K-means faz uma aproximação convergente a um mínimmo local deste problema.

K-means: funcionamento

1. Partindo de um conjunto de centroides iniciais $c_r, \; r = 1, \cdots, k$ (possivelmente aleatórios).
2. Calcula-se qual é a melhor associação α com c fixo, de maneira a minimizar a função $J(\alpha, c)$. Ou seja, para cada instância, qual centroide mais próximo dela.
3. Calcula-se então um valor de c com α fixo, de maneira a minimizar a função $J(\alpha, c)$. Ou seja, um reposicionamento dos centroides com um agrupamento fixo.
4. Repete-se os passos 2 e 3 até que não sejam mais observadas mudanças no valor de α (o que acarreta o fim das mudanças em c igualmente e portanto a convergência do processo).

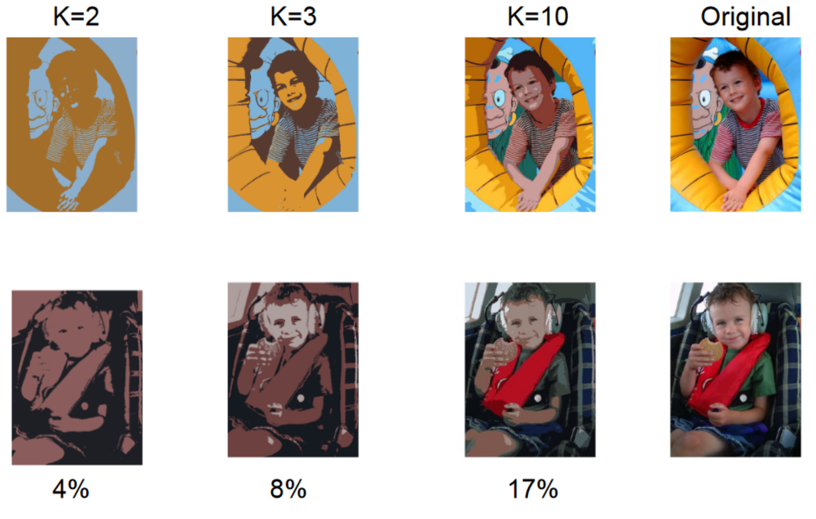
Exemplo K-means



(a) (b) (c)

(d) (e) (f)

Aplicação compressão de imagens



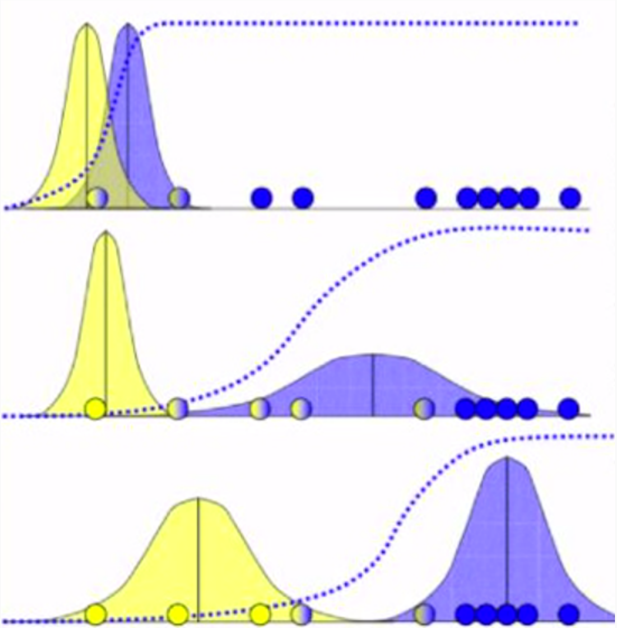
EM: Expectation Minimization

- Similarmente ao método K-means, o método EM consiste em descobrir modelos de dispersão para cada um dos grupos que desejamos encontrar.
- Também recebe como entrada um número fixo de grupos a serem formados.
- No entanto no método EM parte da premissa que os grupos possuem uma distribuição gaussiana dentro do espaço de características.
- Queremos portanto encontrar uma decomposição de máxima verossimilhança das instâncias em k distribuições gaussianas.

K-means: funcionamento

1. Partimos de distribuições gaussianas iniciais com parâmetros de média, variância e matriz de covariância possivelmente aleatórios.
2. Calcula-se qual é a probabilidade de cada instância para cada uma das distribuições consideradas (a priori).
3. A partir de uma redução bayesiana, calcula-se então qual é a probabilidade de cada ponto pertencer a cada uma das distribuições consideradas, dado que os pontos são explicados por alguma delas (a posteriori).
4. Recalcula-se os parâmetros de cada uma das distribuições consideradas, ponderando-se a média, a variância e matriz de covariância de acordo com as probabilidades (a posteriori) encontradas no passo anterior.
5. Repete-se os passos 2, 3 e 4 até que não sejam mais observadas mudanças significativas no valores das médias,

Exemplo EM unidimensional



Exemplo EM bidimensional