Clusterização

Aprendizado Não-supervisionado

- Como descobrir a estrutura fundamental de um dataset?
- Como resumir e agrupar esta estrutura de uma forma mais útil?
- Como representam-se efetivamente os dados em um formato compacto?
- Estes são objetivos do aprendizado não supervisionado, assim chamado porque lida-se com dados não etiquetados (não há um valor de Y).

Clusterização

- Exemplos em que métodos de aprendizado não-supervisionado podem ser úteis:
  - Uma plataforma de publicidade segmenta a população dos EUA em grupos menores com demografia e hábitos de consumo similares, de modo que os anunciantes possam atingir seu mercado-alvo com propagandas relevantes.
  - O Airbnb agrupa sua relação de casas por vizinhanças, de modo que os usuários possam pesquisar as listas mais facilmente.

Clusterização

O modo mais comum de preparar dados para agrupá-los é definir um conjunto de atributos numéricos para que possamos comparar os itens

K-means

- Informe antecipadamente a quantidade desejada de grupos
- O algoritmo irá determinar o tamanho do grupo baseando-se na estrutura dos dados
- Começará com k centróides dispostos aleatoriamente e vai designar cada item para o centróide mais próximo

K-means

- Após cada atribuição, cada centróide é movido para a posição média de todos os pontos

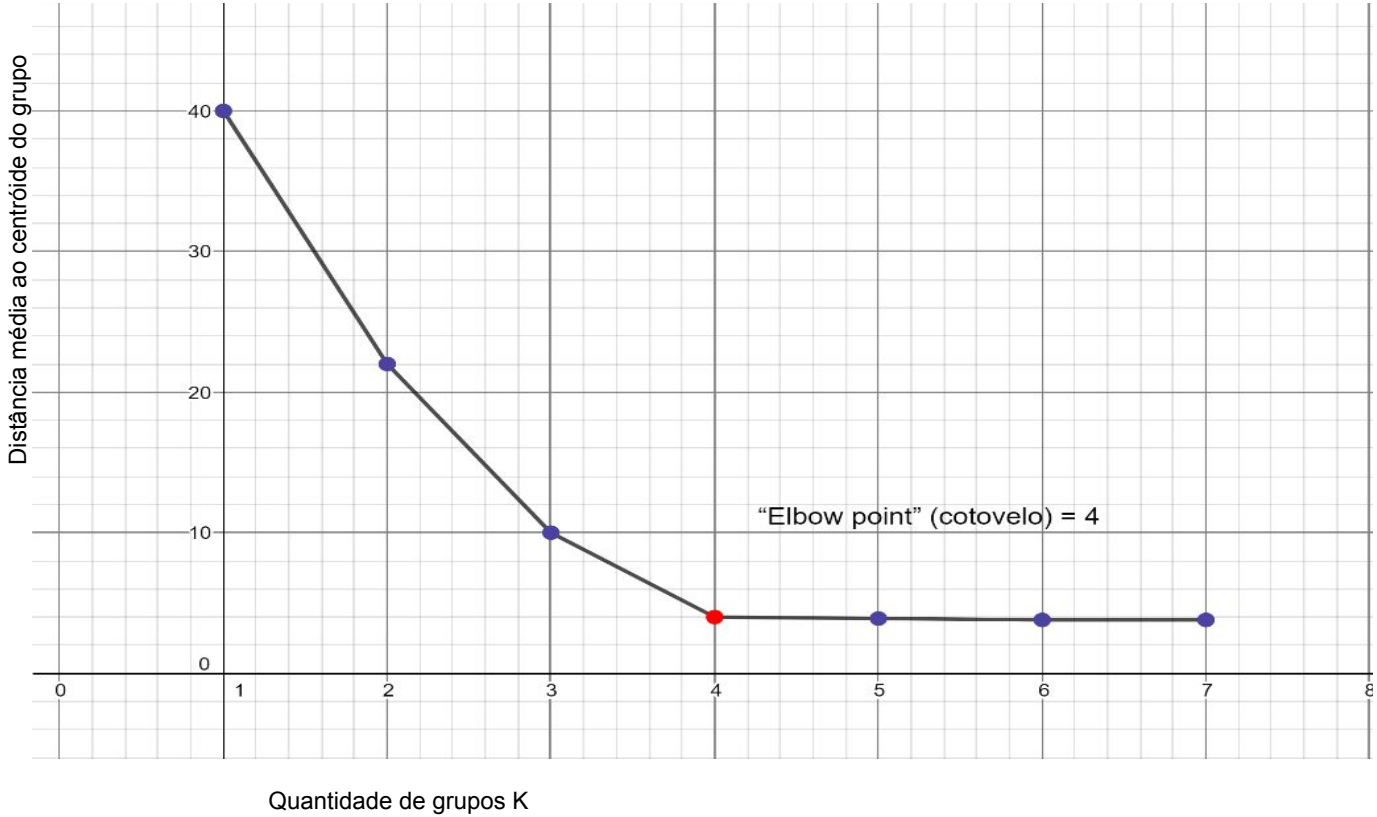
$$c_i = \frac{1}{|S_i|} \sum x_i \in S_i^{x_i}$$

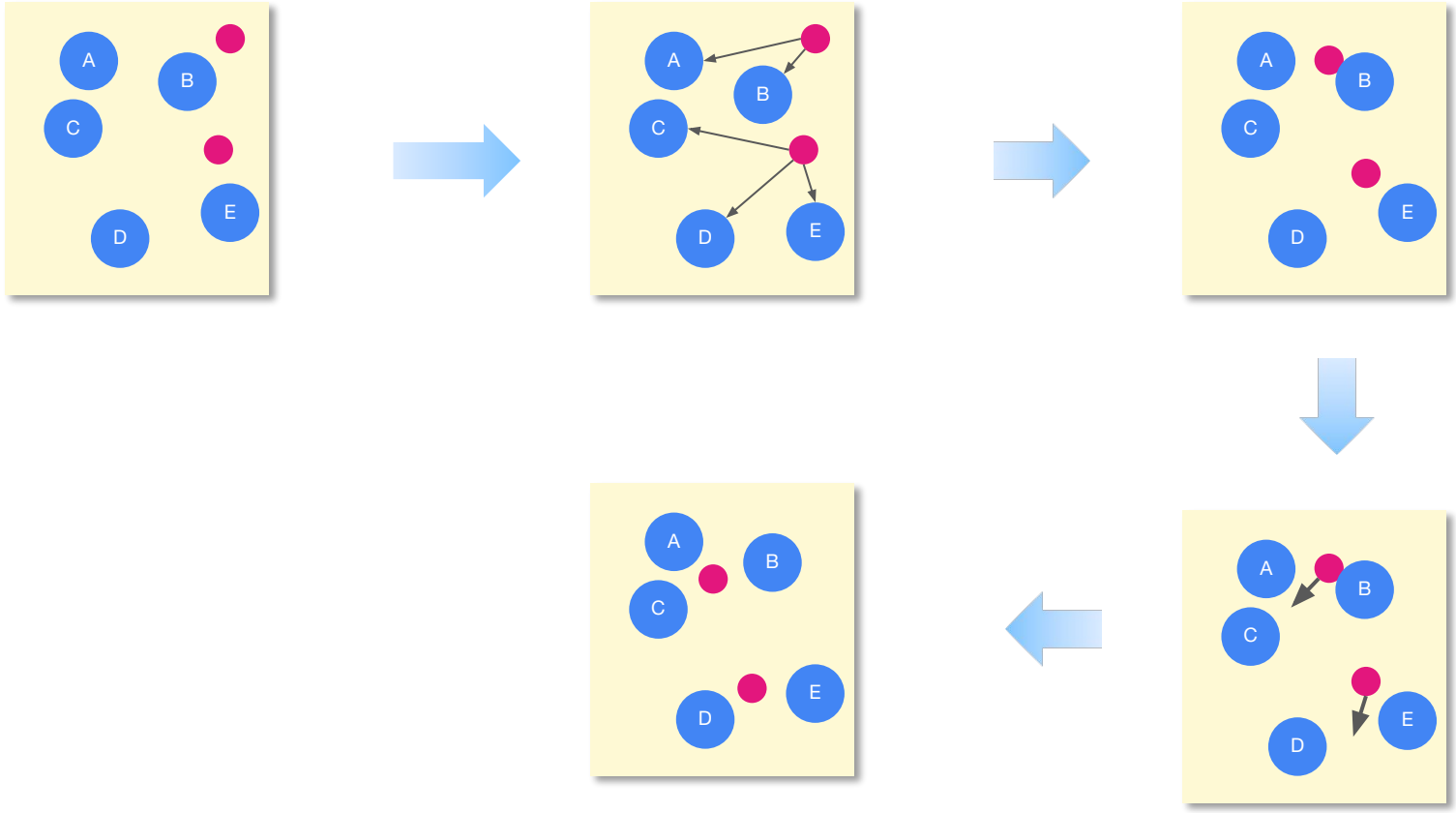
- As atribuições são realizadas novamente
- O algoritmo itera nos passos até que um critério de parada seja satisfeito (não muda mais, o valor da soma é mínimo, o número de iterações atingiu um máximo, por exemplo).

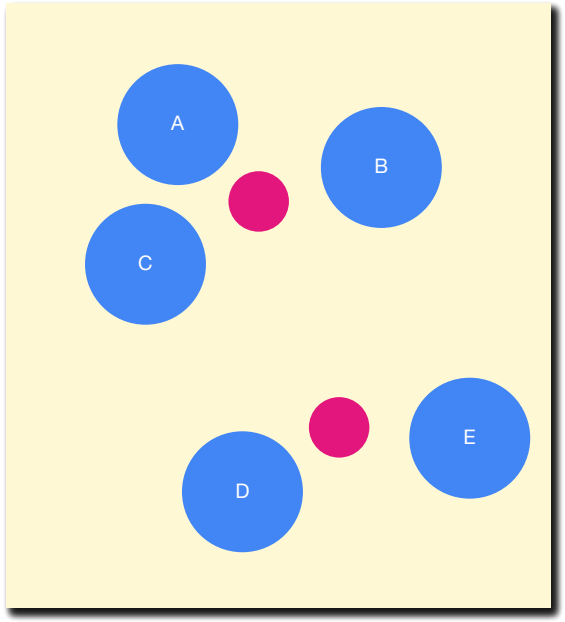
Escolhendo o valor de K

- Métrica usada:
  - distância média entre os pontos dos dados e o centróide de seu cluster (aumentar o valor de K irá sempre diminuir esta métrica)
  - assim, plota-se a distância média como uma função de K e observa-se o valor em que a taxa de queda varia bruscamente (cotovelo, *elbow point*)
- *cross-validation*, método de silhueta, algorito G-means

Escolhendo o valor de K




A
B
C
D
E
A
B
C
D
E
A
B
C
D
E
A
B
C
D
E
A
B
C
D
E


A
B
C
D
E

Exercício: