Introdução ao K-Means Clustering

Tarefa de Leitura

Capítulo 10 do
Introduction to Statistical Learning
por Gareth James, et al.

K-Means Clustering

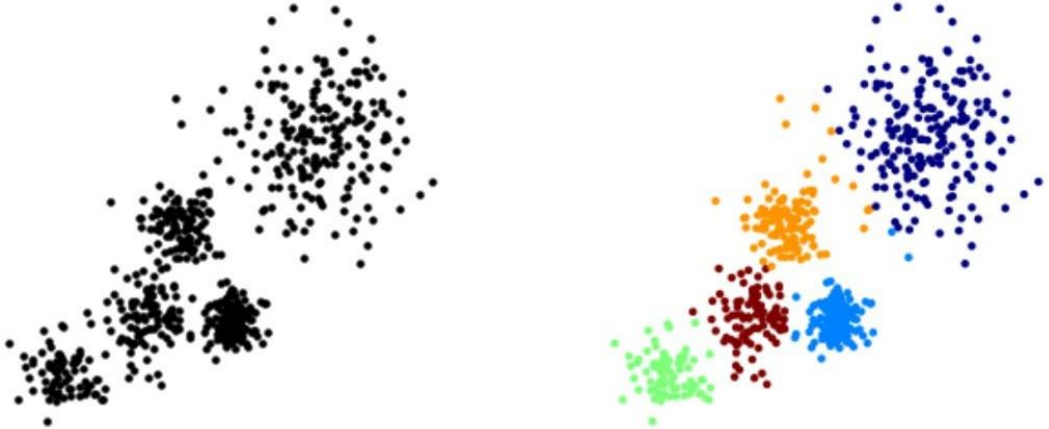
K Means Clustering é um algoritmo de aprendizado não supervisionado que tentará agrupar clusters semelhantes em seus dados.

Como é um problema típico de agrupamento?

- Agrupar documentos semelhantes
- Agrupar clientes com base em características
- Segmentação de mercado
- Identifique grupos físicos semelhantes

K-Means Clustering

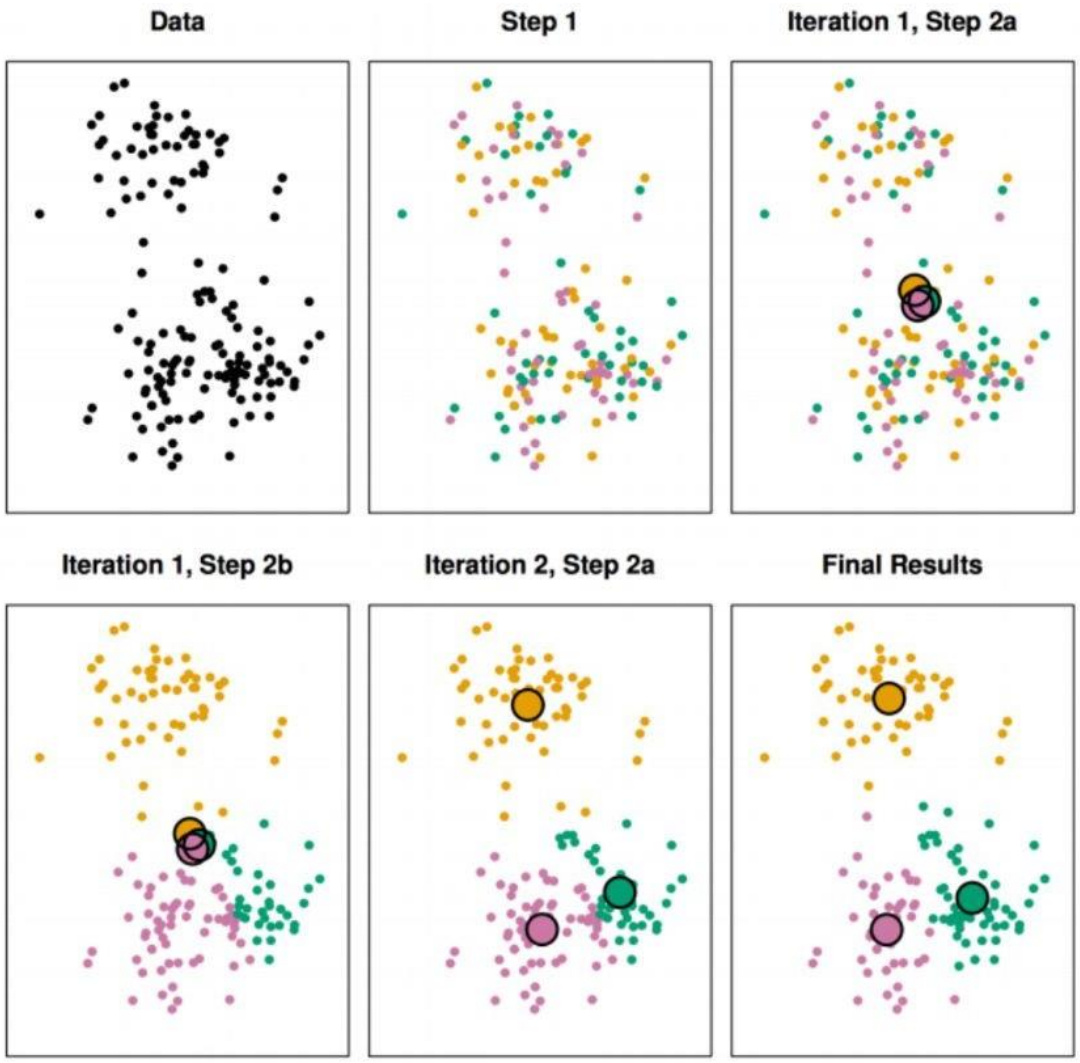
- O objetivo geral é dividir os dados em grupos distintos, de modo que as observações dentro de cada grupo sejam semelhantes.



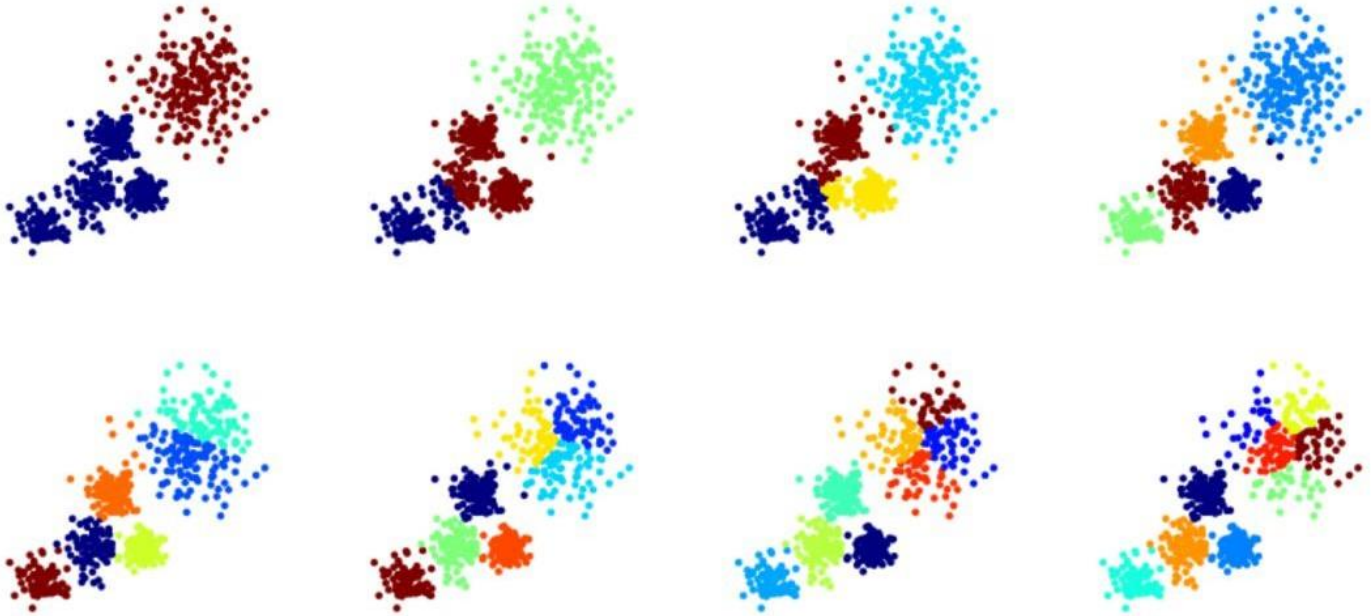
K-Means Clustering

O algoritmo K Means

- Escolha um número de clusters “K”
- Atribuir aleatoriamente cada ponto a um cluster
- Até que os clusters parem de mudar, repita o seguinte:
  - Para cada cluster, calcule o centroide do cluster tomando o vetor médio de pontos no cluster
  - Atribua cada ponto de dados ao cluster para o qual o centroide é o mais próximo


Data
Step 1
Iteration 1, Step 2a
Iteration 1, Step 2b
Iteration 2, Step 2a
Final Results

Escolhendo um valor de K



Escolhendo um valor de K

- Não há resposta fácil para escolher um “melhor” valor de K
- Uma maneira é o método do cotovelo (elbow method)

Em primeiro lugar, calcule a Soma do Erro Quadrático (Sum of Squared Error - SSE) para alguns valores de K (por exemplo 2, 4, 6, 8, etc.).

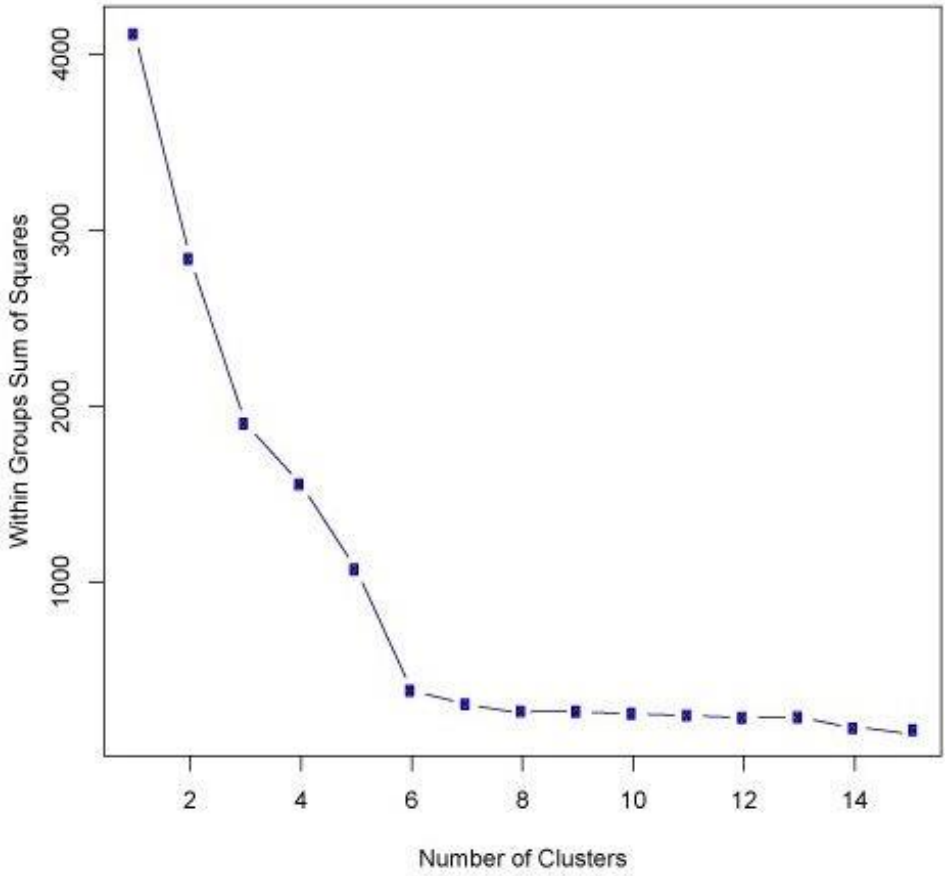
O SSE é definido como a soma do quadrado da distância entre cada membro do cluster e seu centroide.

Escolhendo um valor de K

Se você plotar K em relação ao SSE, verá que o erro diminui à medida que k aumenta; isso porque quando o número de clusters aumenta, eles devem ser menores, então a distorção também é menor.

A ideia do método do cotovelo (elbow method) é escolher o K no qual o SSE diminui abruptamente.

Isso produz um "efeito cotovelo (elbow effect)" no gráfico, como pode ser visto na imagem a seguir:


Within Groups Sum of Squares
1000
2000
3000
4000
2
4
6
8
10
12
14
Number of Clusters