INSTITUTO DO EMPREGO E FORMAÇÃO PROFISSIONAL, IP
DELEGAÇÃO REGIONAL DE LISBOA E VALE DO TEJO
CENTRO DE EMPREGO E FORMAÇÃO PROFISSIONAL DE SINTRA

UFCD – 10810

Fundamentos do desenvolvimento de modelos analíticos em Python

6 – K Means Clustering

Carga horária: 25 horas

Formador: Manuel Viana

1

Introdução

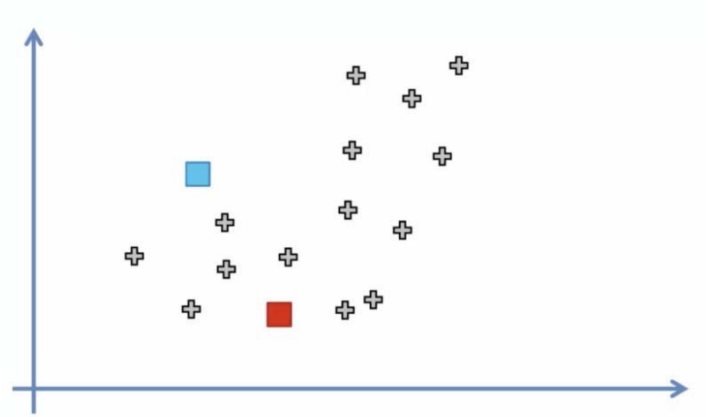
- O K-Means Clustering é um algoritmo iterativo e **não supervisionado** de Machine Learning.
- Normalmente os algoritmos não supervisionados, fazem inferências a partir de conjuntos de dados, usando apenas vetores de entrada, sem se referirem a resultados conhecidos ou rotulados.
- O objetivo passa por agrupar pontos de dados semelhantes e descobrir padrões subjacentes.
- Depois de executado e dos grupos definidos, quaisquer novos dados podem ser facilmente atribuídos ao grupo mais relevante.
- Aplicações a situações do mundo real:
  - perfis de cliente
  - segmentação do mercado
  - motores de pesquisa
  - astronomia

2

2

Como funciona?

1. Seleciona **K** pontos aleatórios (neste exemplo, 2) como centros de cluster (centroides).

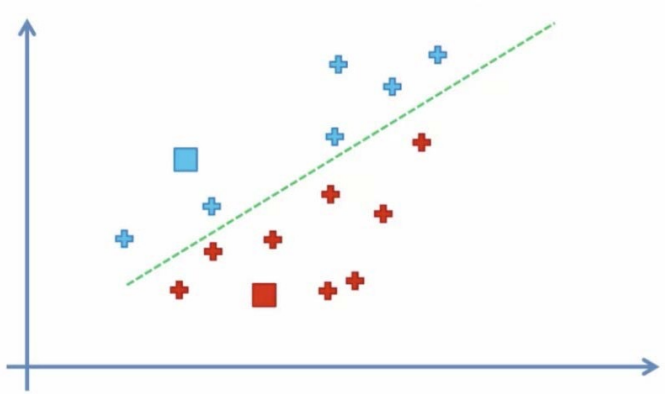


3

3

Como funciona?

2. Atribui cada ponto de dados ao cluster mais próximo, calculando a distância em relação a cada centroide.

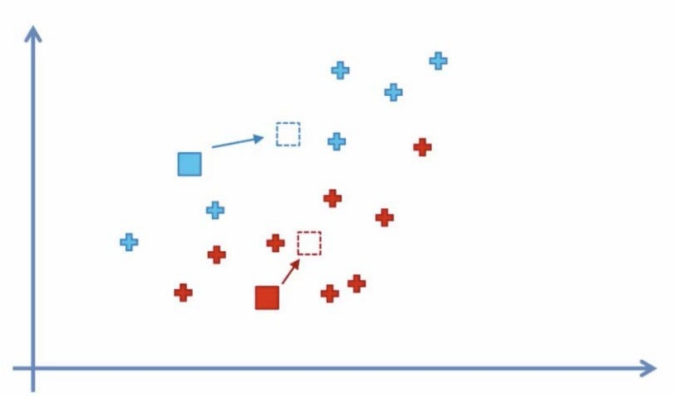


4

4

Como funciona?

3. Determina o novo centro do cluster, calculando a média dos pontos atribuídos.

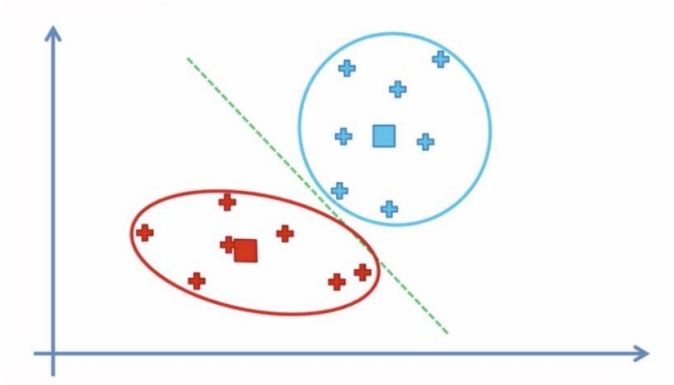


5

5

Como funciona?

4. Repete as etapas 2 e 3 até que nenhuma das atribuições do cluster mude.

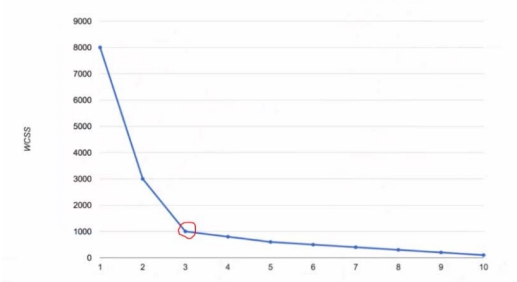


6

6

Qual o número ideal de clusters?

- Frequentemente, os dados terão várias dimensões, dificultando a visualização.
- Como consequência, o número ideal de clusters não é óbvio, mas pode ser calculado matematicamente.
  - **Método Elbow (cotovelo)**
  - Representa-se graficamente a relação entre o **número de clusters** e a **Soma dos Quadrados do Cluster** (WCSS) e, de seguida, seleciona-se o número de clusters onde a mudança no WCSS começa a estabilizar.

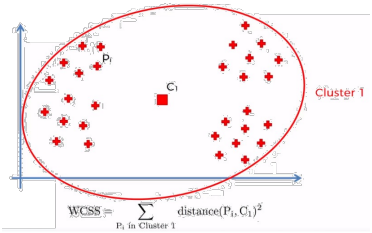


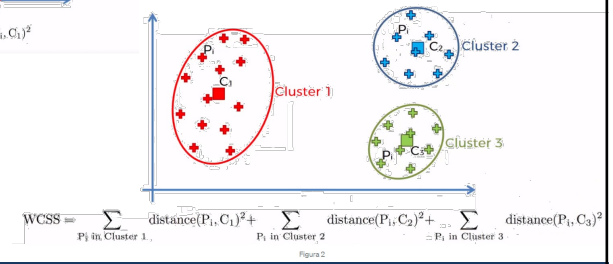
7

7

Qual o número ideal de clusters?

- Neste exemplo, o WCSS calculado para a figura 1 seria maior que o WCSS calculado para a figura 2.





$$WCSS = \sum_{P_i \text{ in Cluster 1}} distance(P_i, C_1)^2 + \sum_{P_i \text{ in Cluster 2}} distance(P_i, C_2)^2 + \sum_{P_i \text{ in Cluster 3}} distance(P_i, C_3)^2$$

Figura 2

8

8