Clusterização

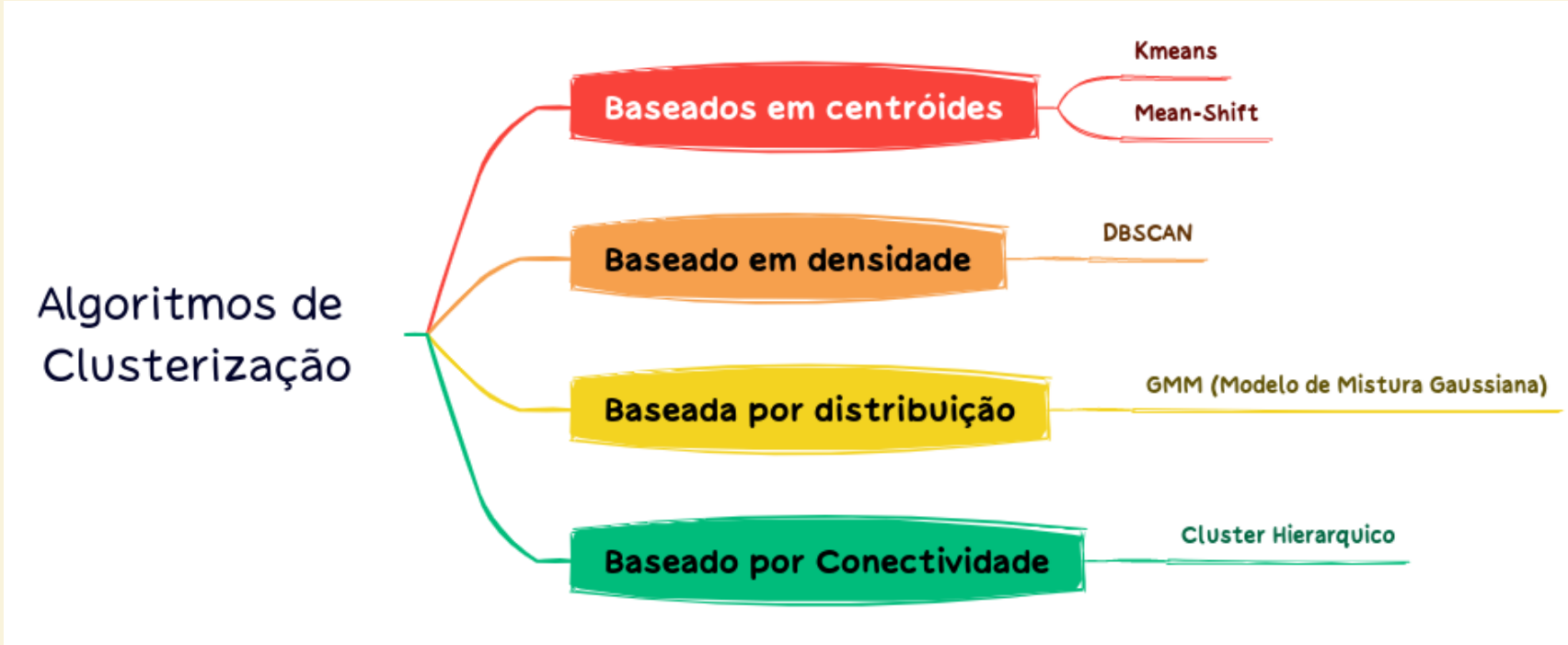
O que é?

- Métodos de análise de agrupamento (análise de cluster) têm por finalidade dividir a amostra em grupos. Essa divisão é feita de modo que os grupos difiram uns dos outros, mas os elementos pertencentes ao mesmo grupo sejam parecidos entre si.


Aprendizagem de Máquina
Supervisionado
Não Supervisionado
Classificação
Regressão
Clusterização

Aplicações

- Segmentação de clientes.
- Resenhas de usuários mostrados em sites de compras online quando um indivíduo busca por um produto.


Algoritmos de Clusterização
Baseados em centróides
Kmeans
Mean-Shift
Baseado em densidade
DBSCAN
Baseada por distribuição
GMM (Modelo de Mistura Gaussiana)
Baseado por Conectividade
Cluster Hierarquico

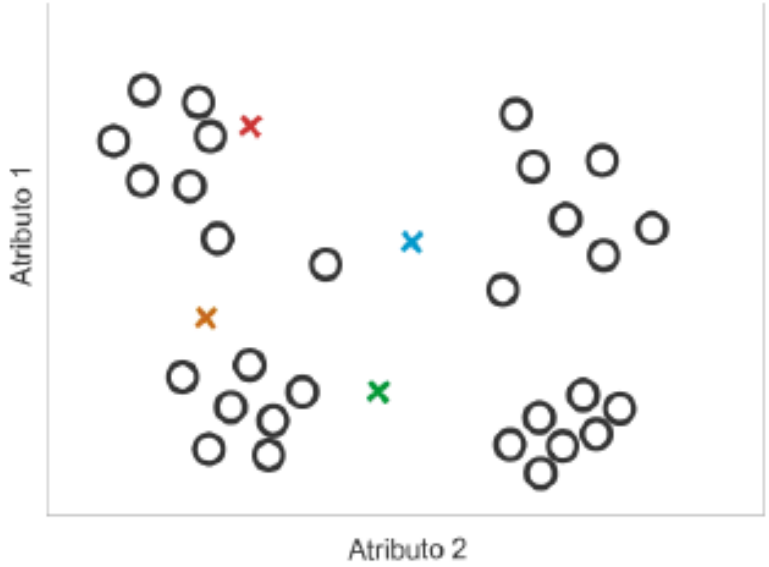
Kmeans

- Assume a medida de dissimilaridade é a distância euclidiana
- Para utilizá-lo é necessário especificar de antemão o valor de k, quantos clusters se deseja.

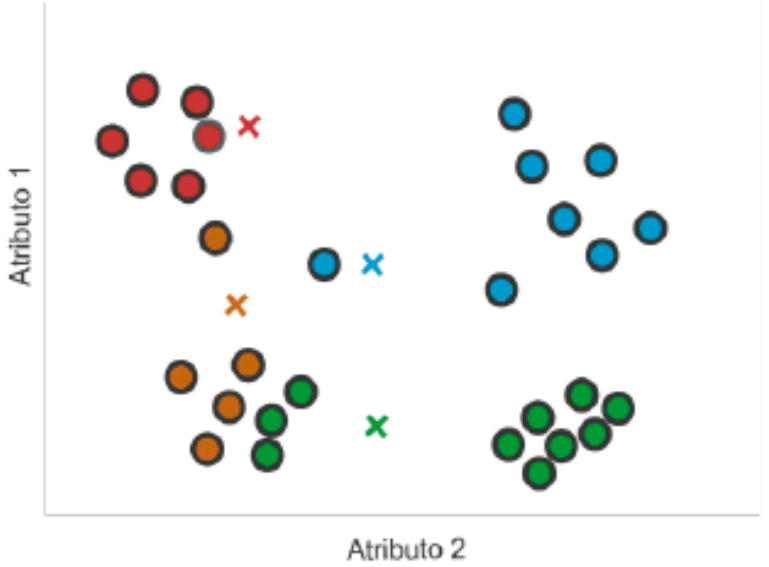
Passos

- Escolha aleatoriamente k centroides
- Determine o Cluster
- Calcule os novos centroides utilizando os grupos que foram criados.

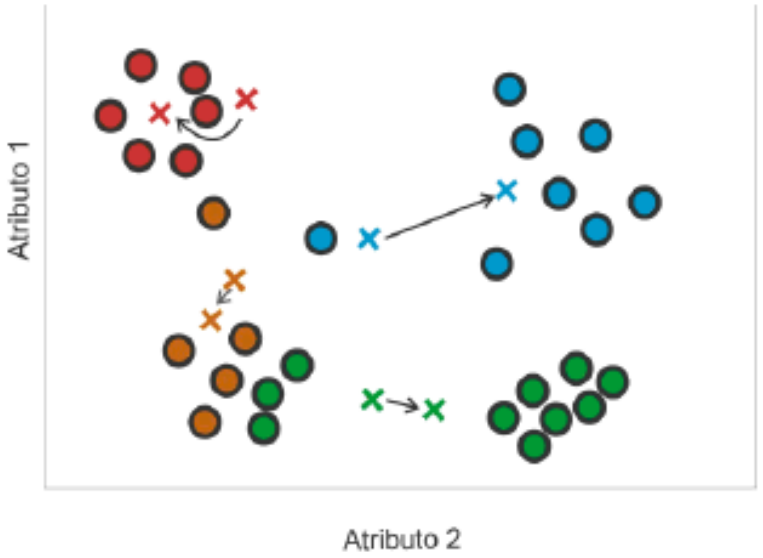
Figura 2. Exemplo do Algoritmo K-means – parte 1.



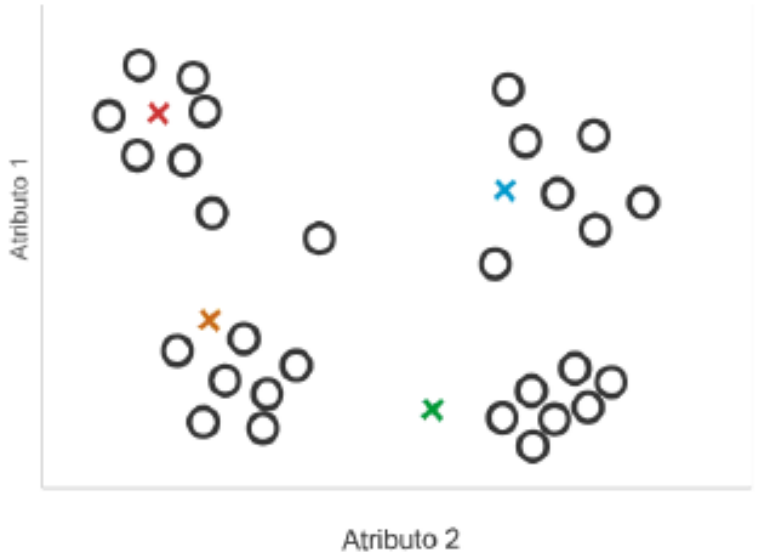
(a) Passo 1 - Sorteio dos valores iniciais dos centróides



(b) Passo 2 - Atribuição dos objetos aos grupos cujo protótipo possua maior similaridade.



(c) Passo 3 - Recalcula os valores dos protótipos como sendo a média dos objetos atuais pertencentes aos grupos.



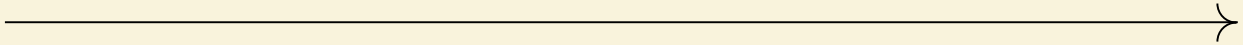
(d) Preparação para repetição dos passos 2 e 3 - Desvincula-se os objetos dos grupos.

Prós

- É um método rápido porque não realiza muitos cálculos
- Não realiza predições ambíguas.

Contras

- Identificar e classificar os grupos pode ser um aspecto desafiador
- Como o centroide do cluster começa em um ponto aleatório, os resultados podem ser inconsistentes



Método do Cotovelo

- Matematicamente falando, nós estamos buscando uma quantidade de agrupamentos em que a soma dos quadrados intra-clusters (ou do inglês within-clusters sum-of-squares, comumente abreviado para `wcss`) seja a menor possível, sendo zero o resultado ótimo.

$$\text{distance}(P_0, P_1, (x, y)) = \frac{|(y_1 - y_0)x - (x_1 - x_0)y + x_1 y_0 - y_1 x_0|}{\sqrt{(y_1 - y_0)^2 + (x_1 - x_0)^2}}$$

Fórmula para o cálculo entre um ponto e uma reta que passa por P0 e P1

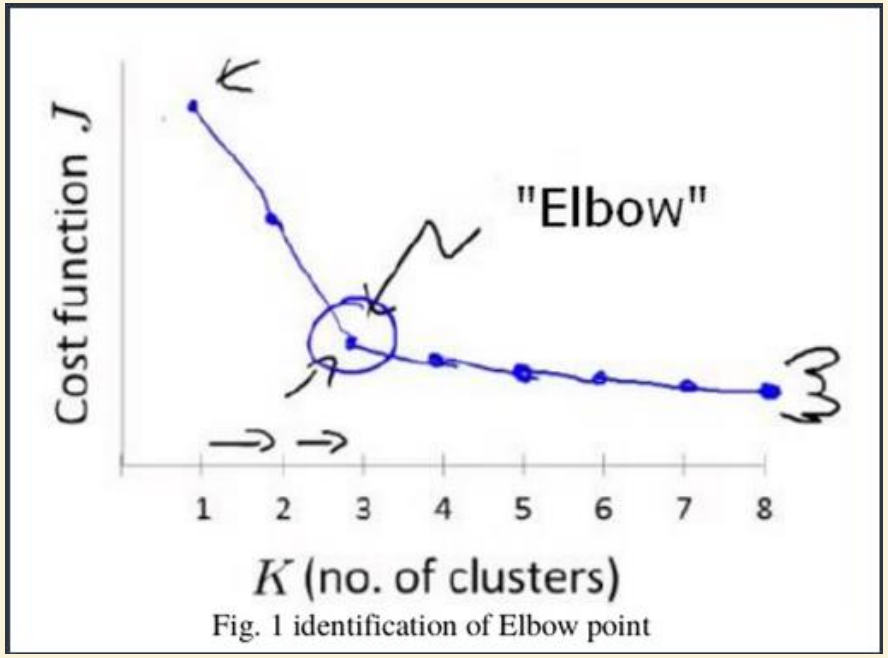


Fig. 1 identification of Elbow point

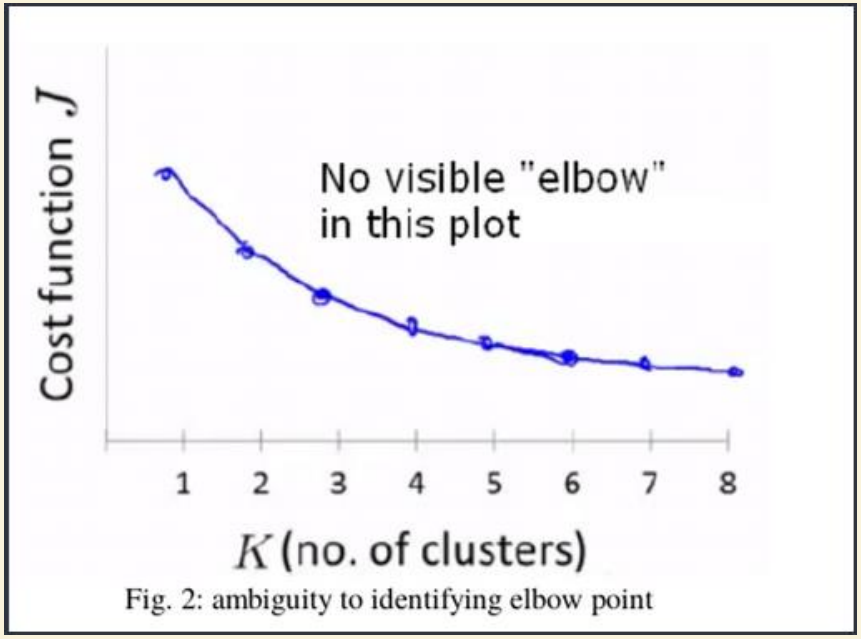
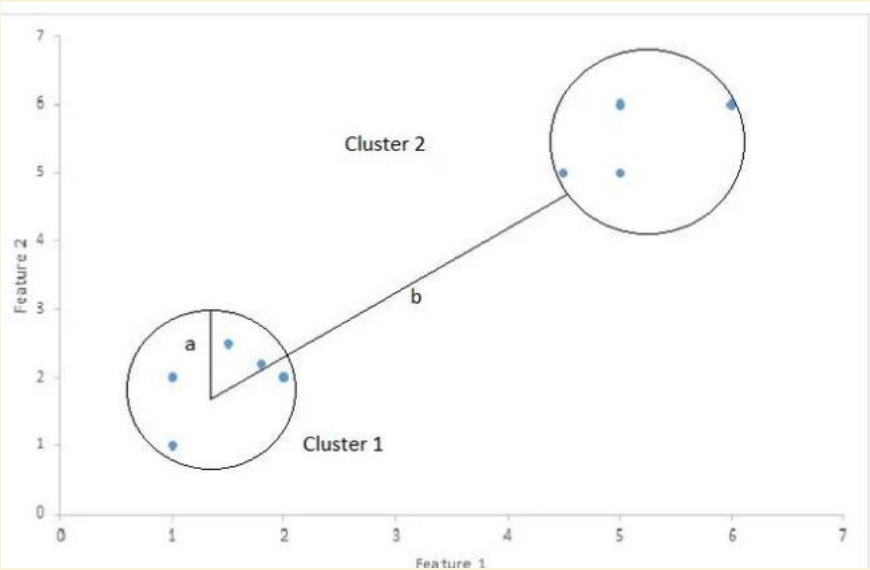


Fig. 2: ambiguity to identifying elbow point

Silhouette Score

É uma métrica usada para calcular o quão boa aquela técnica de cluster é, podendo variar de -1 a 1, sendo que -1 é ruim e 1 seria ótimo.



Silhouette Score= (b-a)/max(a,b)

Onde:

- **a**= distância intra-cluster média, ou seja, a distância média entre cada ponto em um cluster.
- **b**= distância média entre clusters, ou seja, a distância média entre todos os clusters.

Obrigado!

Qr Code para acessar o Linkedin