Aprendizado **NÃO** supervisionado

Definição

Aprendizado não supervisionado, usa algoritmos de aprendizado de máquina (ML) para analisar e agrupar conjuntos de dados não rotulados. Esses algoritmos descobrem padrões ocultos ou agrupamentos de dados sem a necessidade de intervenção humana.

Fonte: https://www.ibm.com/think/topics/unsupervised-learning

Motivação

Você tem um conjunto de dados e necessita dividi-los em grupos, como faria para encontrar um critério de seleção ?

- Seleção aleatória.
- Utilizando padrões.
- Utilizando métricas.

Tipos de Algoritmos

1. Algoritmos de Clusterização (K-Means, DBSCAN, Hierarchical Clustering, Gaussian Mixture Models)

Utilizados para segmentar dados em grupos com características semelhantes.

2. Algoritmos de Redução de Dimensionalidade (Principal Component Analysis, t-SNE, UMAP)

Usados para reduzir o número de variáveis mantendo a maior parte da informação.

3. Algoritmos de Detecção de Anomalias (Isolation Forest, Local Outlier Factor, One-Class SVM)

Identificam padrões incomuns nos dados.

4. Algoritmos de Associação (Apriori, Eclat)

Descobrem regras de associação entre variáveis.

K-means

O K-means consiste em um agrupamento ou clusterização.

No aprendizado não supervisionado, não temos um conjunto de treinamento e teste, mas todos os pontos de dados que temos são pontos de dados de 'treinamento' e construímos os grupos (que definirão o hiperplano) a partir deles. Os vetores de linha de entrada não têm um rótulo; eles consistem apenas em recursos.

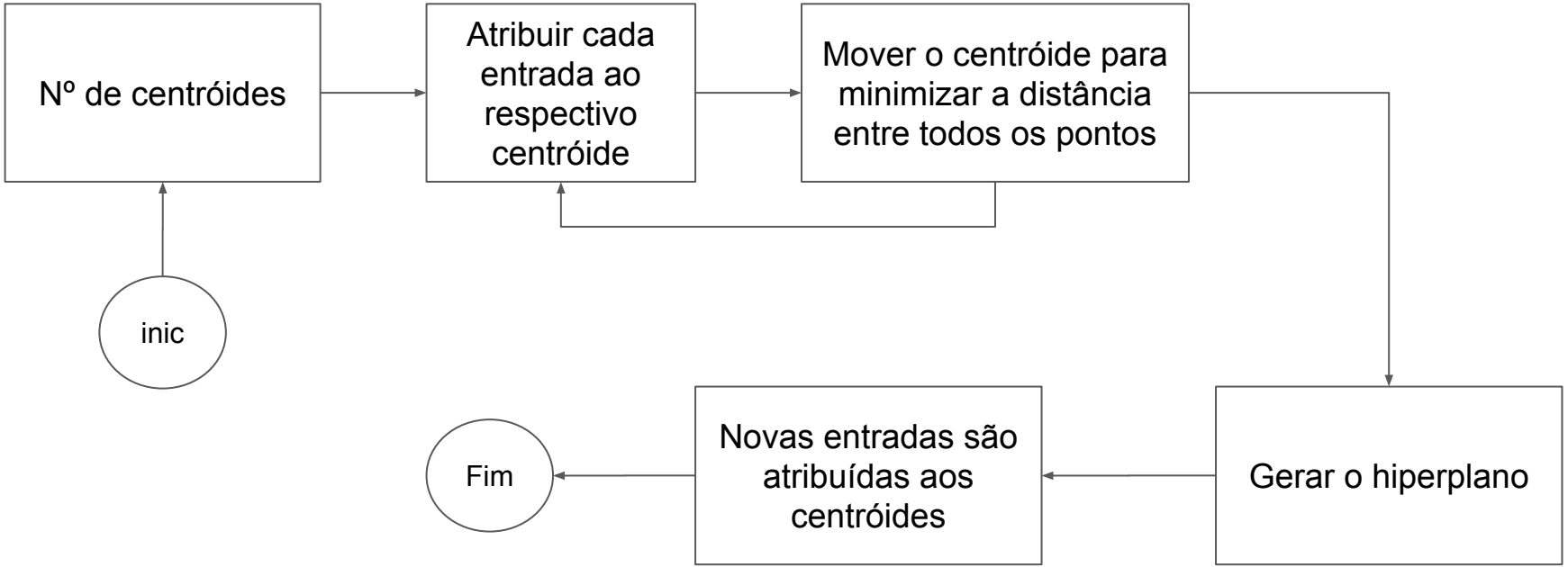
K-means

Entrada: O algoritmo K-means recebe como entrada o número de centróides (grupos) a serem usados. No início do algoritmo, os centróides são colocados em um local aleatório no espaço vetorial do ponto de dados**.**

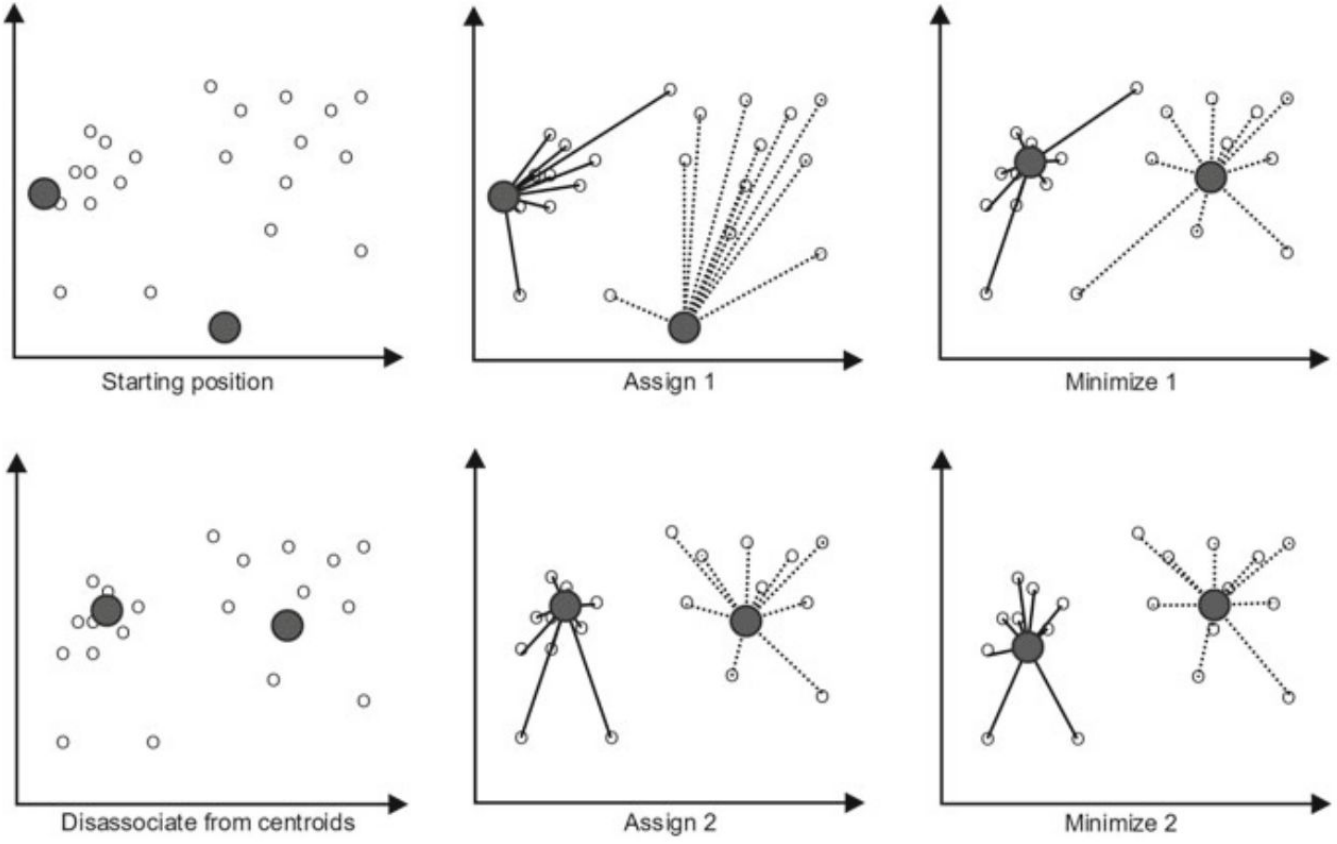
Etapas: Atribuir e Minimizar.

Durante a fase de atribuição, cada ponto de dados é atribuído ao centróide mais próximo em termos de distância euclidiana.

Durante a fase de ‘minimizar’, os centróides são movidos em uma direção que minimiza a soma da distância de todos os pontos de dados atribuídos a ele.


Nº de centróides
Atribuir cada entrada ao respectivo centróide
Mover o centróide para minimizar a distância entre todos os pontos
inic
Fim
Novas entradas são atribuídas aos centróides
Gerar o hiperplano

K-means

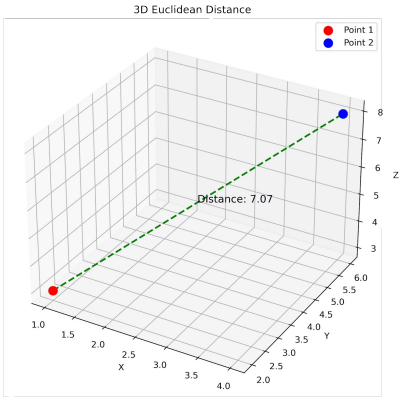

Starting position
Assign 1
Minimize 1
Disassociate from centroids
Assign 2
Minimize 2

Distância Euclidiana

A distância euclidiana representa o caminho mais curto entre dois pontos no espaço euclidiano. A distância euclidiana permite o cálculo da distância em n-dimensões.

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$



$$d = \sqrt{\sum_{i=1}^{n} (b_i - a_i)^2}$$

Exemplo do k-means

x	y
2.50	1.86
2.65	3.52
1.77	1.77
3.58	2.77
1.53	2.54
1.54	1.53
2.24	0.09
0.28	1.44
0.99	2.31
1.09	0.59

Centróides Iniciais:

- C1=(1.53,2.54)
- C2=(1.09,0.59)
- C3=(7.83,7.63)

Métrica

- **Quantos ciclos devem ser executados ?**

  O número de ciclos pode estar relacionado com a mudança após o cálculo da média. Caso a quantidade de elementos em cada um dos grupos não se altere, este pode ser considerado o ponto de parada.

K-means - *Dunn Coefficient*

- Como saber se o cluster está correto ?

A densidade de um cluster pode ser medida utilizando o índice de Dunn (*Dunn Coefficient*). O Dunn Coefficient é uma métrica usada para avaliar a qualidade de clusters em um algoritmo de agrupamento, como o K-Means. Ele mede a compactação e a separação dos clusters, ajudando a determinar se os grupos formados são bem definidos.

K-means - *Dunn Coefficient*

$$D_C = \frac{min\{d(i, j) | i, j \in Centroids\}}{d^{in}(C)}$$

$d(i, j)$ É a distância euclidiana entre os centróides (i,j)

$$d^{in}(C) = max\{d(x, y) | x, y \in C\},$$

É a maior distância intra-cluster

K-means - *Dunn Coefficient*

$$min\{d(i, j)|i, j \in Centroids\}$$

Mede a menor distância entre os elementos de clusters diferentes. Clusters bem separados possuem valores altos.

$$d^{in}(C) = max\{d(x, y)|x, y \in C\},$$

Mede o diâmetro do cluster, ou seja, a maior distância entre dois pontos dentro do mesmo cluster.

K-means - Dunn Coefficient

Interpretação do Índice de Dunn

Valores altos: Indicam clusters bem separados e compactos (boa qualidade de clusterização).

Valores baixos: Indicam clusters muito espalhados ou sobrepostos (má qualidade).

K-means aplicado na segmentação de imagens