Clustering:

1. O que é e para que serve?

↳ Forma de aprendizagem não supervisionada, utilizado quando existem instâncias de treino disponíveis. O objetivo é determinar o agrupamento intrínseco num conjunto de dados não rotulados. Marketing, Biologia, Bibliotecas, seguros ...

2. Tipos de Algoritmos de Clustering

2.1. Clustering Hierárquico:

(1) Aglomerativo → Começa por considerar cada elemento como 1 cluster e vai recursivamente aglomerando outros clusters.

1. Associar cada item a um cluster (n items => n clusters).
2. Encontrar o par de clusters mais próximos e fundi-los.
3. Recalcular distância entre o novo cluster e os outros.
4. Repetir até estarem todos num só cluster.

(2) Divisive → começa por assumir todos num só cluster e vai-se dividindo.

2.1.1. Métricas de distância em clustering Hierárquico: (Linkage):

- Single Linkage: Distância entre os 2 pontos mais próximos de 2 clusters.
- Complete Linkage: Distância entre os 2 pontos mais afastados de 2 clusters.
- Average Linkage: Média das distâncias entre todos os pares de pontos.*
- Ward's Linkage: Encontra o par de clusters que, após fusão, leva ao aumento mínimo da variância total dentro do cluster.

* Cada par é formado por 1 ponto de cada cluster.

2.2. Clustering Particional:

↳ Estes métodos dividem os pontos de dados em clusters de uma só vez. Todos os clusters têm o mesmo nível de hierarquia.

(1) baseado em centroides → Clusters são representados por um vetor central (centroide).

↓

k-means → 1. Colocar k pontos no espaço representados pelos objetos que estão a ser agrupados. Estes pontos representam centroides de grupos iniciais.
2. Atribuir cada objeto ao grupo que tem centroide mais próximo
3. Quando todos os objetos tiverem sido atribuídos, recalcular as posições dos k centroides.
4. Repetir os passos 2 e 3 até que os centroides não se movam

Nota: O número de clusters (k) é feito por visualização, Elbow Method ou abordagens evolutivas.
Escolher os centroides iniciais é feito aleatóriamente, usando médias de n distâncias, maximizar a distância de Clusters ou k-means ++.

(2) baseado em densidade → Clusters definidos com base em areas de grande densidade de pontos.

DBSCAN → Classifica pontos em:
- Core points → É central se minPts pontos estiverem a uma distância ε dele.
- density-reacheable → Pontos alcançaveis a partir de um ponto central.
- outliers → os que não são reachable.

-