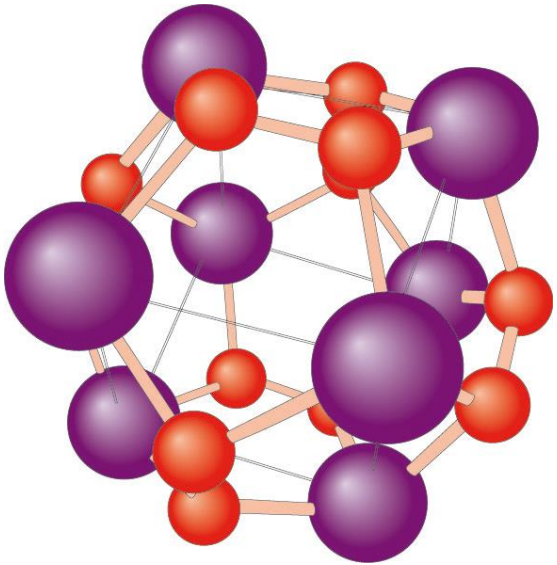
Kmeans

Técnicas de clusterização



Introdução ao Kmeans

- Clustering é uma técnica de aprendizado **não supervisionado** usada para agrupar dados em grupos ou clusters, com base na semelhança entre eles.

- O algoritmo K-means é uma das técnicas de clustering **mais populares.** Ele é usado para dividir um conjunto de dados em um número específico de grupos, onde cada ponto de dados pertence a um grupo específico.

Agenda Kmeans

01. Algoritmos particionais

02. O que é o kmeans?

03. Como escolher o número de clusters k e inicializar os centroides?

04. Vantagens e Desvantagens

05. Hands-on

01. Algoritmos particionais

- Em algoritmos de clusterização existem **3** grandes tipos de modelos, **particionais**, **hierárquicos** e baseados em **densidade**
- Hoje iremos forcar nos particionais que o método mais popular onde o objetivo é dividir o conjunto de dados em partições, existem **2 tipos de clusters particionais**

1 Partição Rígida

- Conhecido como hard clustering a ideia é que cada observação pertença a um único grupo de forma integral.
- Comum em algoritmos como k-means, k-medians, k-medoides, k-modes, k-prototype

2 Partição com sobreposição

- Ao final do algoritmo a observação não necessariamente precisa pertencer a um único cluster
- Resolver problemas onde categorias se sobrepõem, quando há fronteiras bem definida entre clusters
- 3 tipos soft, fuzzy e probabilisticos

02. O que é kmeans

- Clusterização **particional**, onde cada ponto de dados é atribuído a um único cluster **(partição rígida).**
- Objetivo minimizar a soma das distâncias euclidianas ao quadrado de cada ponto em relação ao centroide do cluster ao qual ele foi atribuído

Minimizar a variância intra cluster

$$J = \sum_{c=1}^{k} \sum_{\mathbf{x}_j \in \mathbf{C}_c} d\left(\mathbf{x}_j, \bar{\mathbf{x}}_c\right)^2$$

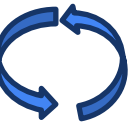
Onde:

- **d =** Distância Euclidiana***** observação Xj ao Centroide do grupo Xc

* É possível o uso de outras medidas de dissimilaridade

K-means tende a criar clusters com variações mínimas dentro do cluster **(coesão ou homogeneidade)** e máxima separação entre os clusters **(separabilidade ou heterogeneidade).** Ou seja, clusters parecidos entre si e diferentes entre os outros

Passo a Passo

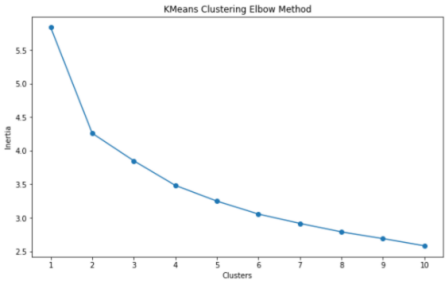


1. **Escolha K centros iniciais para o clusters (centroides)**
2. **Atribua cada observação/objeto para o cluster de centro mais próximo**
3. **Atualiza o centro de cada cluster com base no ponto médio das observação (centroide)**
4. **Repete passo 2 e 3 até que tenha-se um critério de parada**
   - Número máximo de iterações
   - Baixa mudanças nos recálculos dos centroides dos clusters após 2 iterações consecutivas

03. Como escolher o número de cluster k?

Método Elbow (critério de validação interno)

- Técnica gráfica que pode ser usada para determinar o número ideal de clusters no K-means.
- Plota a soma das distâncias quadradas (SSE) em relação ao número de clusters.



Silhouette Score (critério de validação relativo)

- Mede a semelhança entre um ponto e seus próprios clusters em comparação com outros clusters.
- O Silhouette Score varia de -1 a 1, onde um valor mais próximo de 1 indica que o ponto está bem ajustado ao seu próprio cluster e mal ajustado a outros clusters.
- O número ideal de clusters é aquele que maximiza o Silhouette Score médio para todos os pontos.

Como melhorar a inicialização dos K centroides?

- O K-means é **sensível à inicialização dos centróides**, o que pode levar a resultados diferentes dependendo da escolha inicial dos centróides.
- Uma inicialização aleatória pode resultar em uma solução subótima.

Melhorando a inicialização com o K++

1. O primeiro centróide é escolhido aleatoriamente a partir dos dados.
2. Para cada ponto, calcula-se a distância mínima ao centróide mais próximo que já foi selecionado.
3. O próximo centróide é escolhido aleatoriamente a partir dos pontos, com uma probabilidade proporcional à distância ao centróide mais próximo já selecionado.
4. Repetir os passos 2 e 3 até que todos os centróides tenham sido escolhidos.

Beneficios do K++

- O método K++ tende a escolher centróides mais distantes uns dos outros, o que pode levar a uma solução melhor do que a inicialização aleatória.
- Ele também ajuda a evitar a inicialização de centróides próximos a pontos sem variação, o que pode levar a um cluster vazio.

05. Vantagens e desvantagens

Vantagens

- Simples, intuitivo e fácil implementação
- Complexidade linear (número de clusters, número de observações, números de variáveis e iterações)

Desvantagens

- Sensibilidade a outliers
- Sensibilidade a escala dos dados
- Escolha do K
- Sensibilidade a inicialização (problema de mínimo local)
- Clusters globulares
- Limita-se a atributos númericos