
FIAP

Aprendizado não supervisionado - KMeans

Objetivos da aula:

- Apresentar e praticar conceitos de aprendizado não supervisionado
- Compreender o conceito de cluster
- Apresentar os algoritimos KMeans

Métodos de agrupamento

Uma das técnicas de aprendizagem não supervisionada é o agrupamento automático de dados ou clusterização.

Lembre-se: Em aprendizagem não supervissionada, o nosso conjunto de dados para treino não possui label.

Essa técnica classifica os dados em conjuntos que apresentam alguma similaridade (distância das observações). Os grupos gerados nesse processo são chamados de clusters.

EXEMPLO 1

Acesse o link: https://miro.com/app/board/o9J_l2e1B1U=/

K-Means

- O K-Means é SIMPLE por isso é um dos metodos de clusterização mais "clásicos" e recebe esse nome pois encontra uma quantidade K de clusters, sendo K um parâmetro para o modelo e para cada cluster é atribuído um centro chamado de centroide. Tambem conhecido método de partição sem sobreposição.

EXEMPLO 2

- **Acesse o link:** http://alekseynp.com/viz/k-means.html

Curiosidades

- Esse algoritimo está entre os top 10 algoritimos de mineração de dados (data mining)

- É um algoritimo utilizado a mais de 70 anos, existe papers das decadas de 50 e 60 que falam sobre ele.

Algoritmo K-Means

O algoritimo K-Means trata-se de um método iterativo e pode ser definido em 4 etapas, são elas:

1) Escolher aleatoriamente k protótipos (centros) para os clusters

2) Atribuir cada objeto para o cluster de centro mais próximo (segundo alguma distância, Euclidiana)

3) Mover cada centro para a média (centróide) dos objetos do cluster correspondente

4) Repetir os passos 2 e 3 até que algum critério de convergência seja obtido:

número máximo de iterações
limiar mínimo de mudanças nos centróides

Vantagens

- Simples e intuitivo

- Complexidade computacional linear em todas as variáveis críticas

- Eficaz em muitos cenários de aplicação e produz resultados de interpretação simples

Desvantagens

- k = ?
- Sensível à inicialização dos protótipos (mínimos locais)
- Limita-se a encontrar clusters volumétricos / globulares
- Cada item deve pertencer a um único cluster (partição rígida, ou seja, sem sobreposição)
- Limitado a atributos numéricos
- Sensível a outliers

Desafio

Faça agora uma exploração de dados em outra base, conheça a base, utilize KMeans, exiba os resultados e defina uma conclusão.

Copyright © 2023 Prof. Arnaldo Jr/Yan Coelho