K-MEANS

UM ESTUDO SOBRE O ALGORITMO DE AGRUPAMENTO MAIS UTILIZADO NO MUNDO

Brainer Sueverti de Campos

O QUE É O K-MEANS ?

- É UM ALGORITMO DE APRENDIZADO DE MÁQUINA, NÃO SUPERVISIONADO E DE AGRUPAMENTO.

- É PARTICIONAL E BASEADO EM CENTRÓIDES.

COMO FUNCIONA?

O ALGORITMO FUNCIONA ATRAVÉS DE 4 PASSOS:

(1) SELEÇÃO ALEATÓRIA DE K CENTRÓIDES.

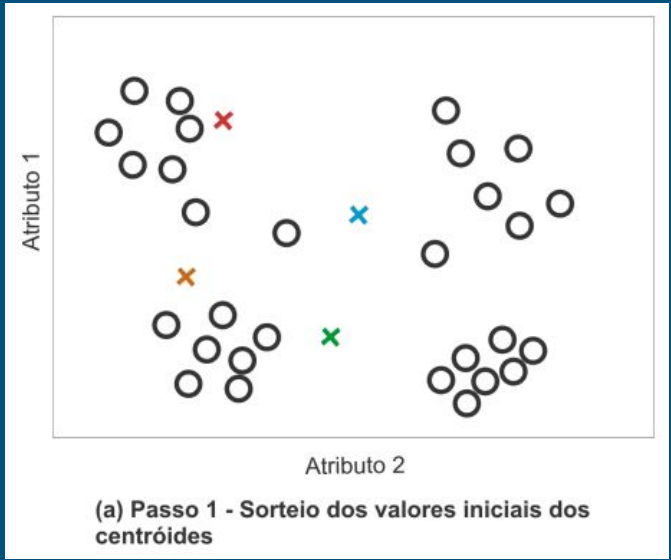
(2) ATRIBUIÇÃO DE CADA ELEMENTO AO CENTRÓIDE MAIS PRÓXIMO.

(3) RECALCULA-SE OS VALORES DOS CENTRÓIDES, DE ACORDO COM A MÉDIA DE CADA AGRUPAMENTO.

(4) REPETIÇÃO DOS PASSOS 2 E 3 ATÉ A CONVERGÊNCIA.

PASSO 1

NO PRIMEIRO PASSO OS CENTRÓIDES SÃO ESCOLHIDOS ALEATORIAMENTE DENTRO DO CONJUNTO.

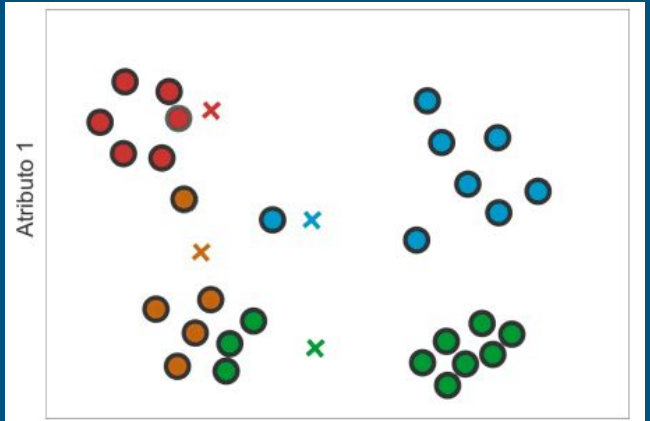


(a) Passo 1 - Sorteio dos valores iniciais dos centróides

FONTE: Referência [2]

PASSO 2

NO PASSO DOIS É CALCULADA A DISTÂNCIA (EUCLIDIANA) DOS ELEMENTOS PARA A CADA CENTRÓIDE E ATRIBUINDO, OS QUE TIVEREM A MENOR DISTÂNCIA , EM UM GRUPO.

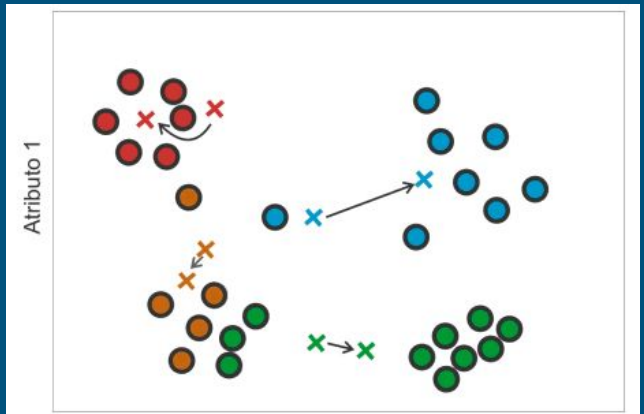


(b) Passo 2 - Atribuição dos objetos aos grupos cujo protótipo possua maior similaridade.

FONTE: REFERÊNCIA [2]

PASSO 3

NO TERCEIRO PASSO AS POSIÇÕES DOS CENTRÓIDES SÃO RECALCULADAS COM BASE NAS MÉDIAS DE CADA AGRUPAMENTO.

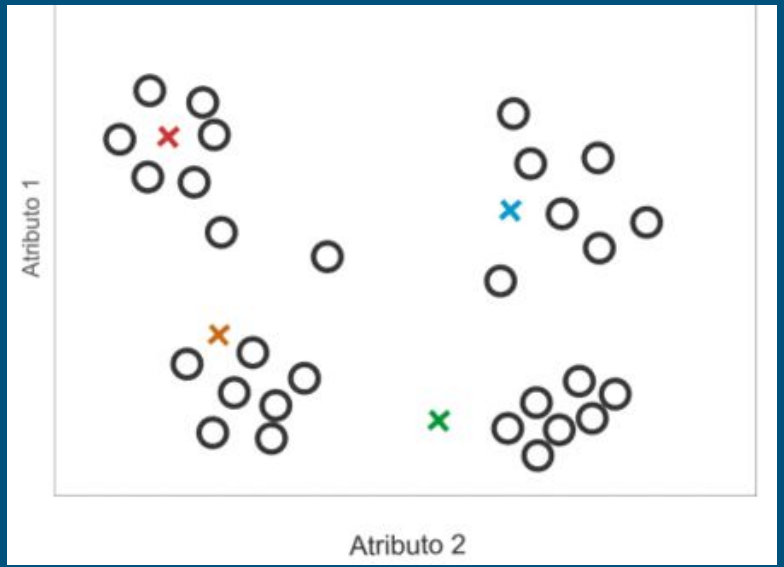


(c) Passo 3 - Recalcula os valores dos protótipos como sendo a média dos objetos atuais pertencentes aos grupos.

FONTE: REFERÊNCIA [2]

PASSO 4

NO QUARTO PASSO, OS PASSOS 2 E 3 SÃO REPETIDOS ATÉ ATINGIR A ESTABILIDADE. A ESTABILIDADE SE DÁ QUANDO NÃO HOUVER MAIS MUDANÇAS NO GRUPO.



FONTE: REFERÊNCIA [2]

COMPLEXIDADE

A COMPLEXIDADE DO ALGORITMO SERÁ O(n * N * T * K), SENDO:

- n, O NÚMERO DE ATRIBUTOS
- N, O NÚMERO DE ELEMENTOS
- T, A QUANTIDADE DE ITERAÇÕES
- K, A QUANTIDADE DE CENTRÓIDES

VANTAGENS

O K-MEANS POSSUI ALGUMAS VANTAGENS EM SUA UTILIZAÇÃO, TAIS COMO:

- TRABALHA BEM COM GRANDES NÚMEROS DE DADOS
- IMPLEMENTAÇÃO SIMPLES
- FÁCIL UTILIZAÇÃO

LIMITAÇÕES

O K-MEANS TAMBÉM POSSUI ALGUMAS LIMITAÇÕES, TAIS COMO:

- É NECESSÁRIO O NÚMERO DE CLASSES (k)
- É AFETADO POR OUTLIERS
- NÃO É ADEQUADO PARA CONJUNTOS NÃO CONVEXOS
- PROBLEMAS PARA DENSIDADES DIFERENTES
- UTILIZA ALEATORIEDADE

LIMITAÇÕES - Número de Classes (K)

PARA QUE O ALGORITMO TEM COMO PARÂMETRO O NÚMERO DE CLASSES (K). DESSA FORMA, FAZ COM QUE TORNE-SE MAIS DIFÍCIL DETERMINAR O NÚMERO DE CLUSTERS ADEQUADO, VISTO QUE TERÁ QUE ESTIMAR OU UTILIZAR OUTROS MÉTODOS.

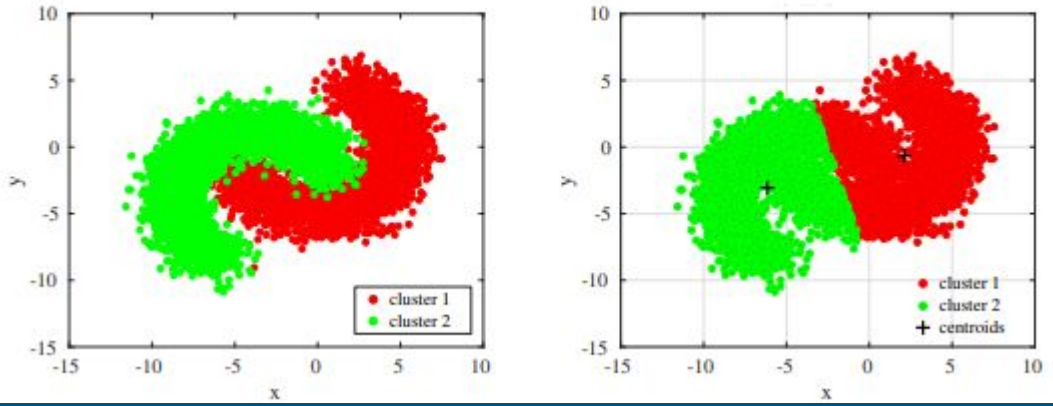
LIMITAÇÕES - Outliers

UM OUTLIER É UM ELEMENTO MUITO DIFERENTE DOS OUTROS ELEMENTOS DENTRO DE UM CONJUNTO. O FATO DO OUTLIER SER MUITO DIFERENTE DOS DEMAIS ELEMENTOS, FAZ COM QUE SUA DISTÂNCIA AFETE NO CÁLCULO DA MÉDIA E, CONSEQUENTEMENTE, NA POSIÇÃO DO CENTRÓIDE QUE, POSTERIORMENTE, AFETARÁ NOS AGRUPAMENTOS.

LIMITAÇÕES - Conjunto Não Convexos

UMA VEZ QUE O ALGORITMO UTILIZA O ESPAÇO EUCLIDIANO COMO BASE, A IDENTIFICAÇÃO DE CONJUNTOS NÃO CONVEXOS FICA MAIS COMPLICADA.

Figura 4.2: Conjunto de Dados: Bananas.
À esquerda, pontos originais; à direita, o agrupamento usando o *K-means* (2 *clusters*)

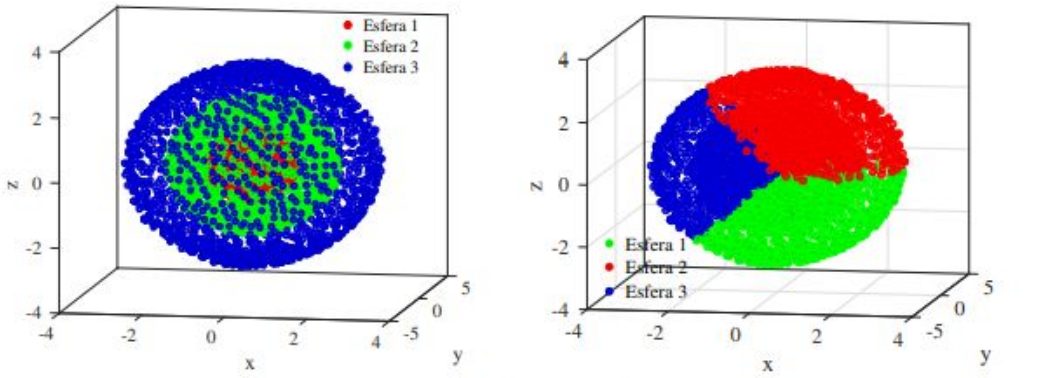


FONTE: REFERÊNCIA [1]

LIMITAÇÕES - Densidades Diferentes

SEMELHANTE A LIMITAÇÃO ANTERIOR, O K-MEANS TEM PROBLEMAS EM IDENTIFICAR GRUPOS COM DENSIDADES (CAMADAS) DIFERENTES.

Figura 4.5: Conjunto de Dados: Esferas.
À esquerda, pontos originais usando projeção; à direita, o agrupamento usando o *K-means* (3 *clusters*)



FONTE: REFERÊNCIA [1]

LIMITAÇÕES - Aleatoriedade

A ESCOLHA DOS CENTRÓIDES INICIAIS AFETA NEGATIVAMENTE NO ALGORITMO, UMA VEZ QUE ESCOLHAS RUINS PODEM LEVAR A AGRUPAMENTOS NÃO IDEAIS. ALÉM DISSO, TAL ALEATORIEDADE O TORNA NÃO DETERMINÍSTICO (PARA UMA MESMA ENTRADA, SAÍDAS DIFERENTES).

EXTRA - IMAGENS HIPERESPECTRAIS

UM EXEMPLO DE DESVANTAGEM AO UTILIZAR O K-MEANS, É NA EXTRAÇÃO DE CARACTERÍSTICAS DE IMAGENS HIPERESPECTRAIS, JÁ QUE O ALGORITMO UTILIZA DISTÂNCIA EUCLIDIANA E ELA NÃO FUNCIONA NO CONJUNTO ABORDADO.

CONCLUSÃO

PORTANTO, O K-MEANS É UM ALGORITMO DE AGRUPAMENTO FÁCIL DE UTILIZAR E BOM EM DETERMINADOS USOS, PORÉM HÁ LIMITAÇÕES QUE DEVEM SER TRATADAS.

REFERÊNCIAS

SOUSA, Maria Cristina Cordeiro. *Uma análise do algoritmo K-means como introdução ao aprendizado de máquinas*. 2019. 74 f. Monografia (Graduação) - Curso de Matemática, Universidade Federal do Tocantins, Araguaína, 2019

Fontana, André, e Murilo Coelho Naldi. *Estudo e comparação de métodos para estimação de números de grupos em problemas de agrupamento de dados*. março de 2009. *repositorio.icmc.usp.br*, http://repositorio.icmc.usp.br//handle/RIICMC/6697.