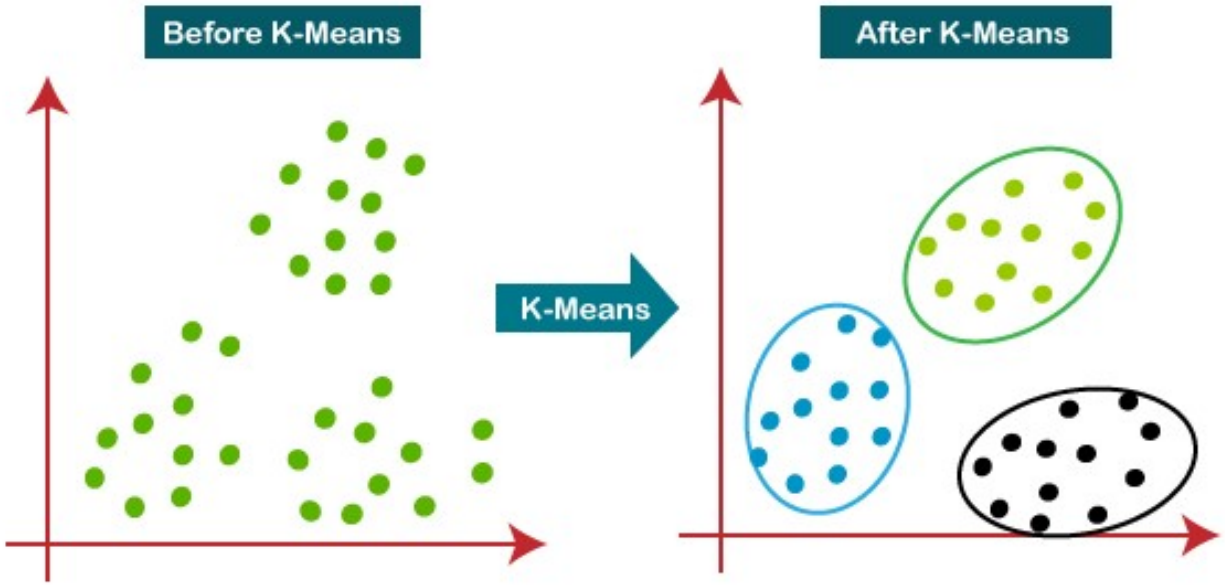
Classificação Não Supervisionada

Prof. Gustavo Willam Pereira

INSTITUTO FEDERAL
Sudeste de Minas Gerais

K Means Clustering in Python

- K-Means é um dos algoritmos de aprendizado de máquina não supervisionados mais populares usados para resolver problemas de classificação.
- K-Means separa os dados não rotulados em vários grupos, chamados clusters, com base em características semelhantes, padrões comuns.



Fonte: https://www.analyticsvidhya.com/blog/2021/04/k-means-clustering-simplified-in-python/

O que é Clustering ?

- Suponha que temos um número N de conjuntos de dados multivariados não rotulados de vários animais como cães, gatos, pássaros etc.
- A técnica para segregar conjuntos de dados em vários grupos, com base em características e características semelhantes, está sendo chamada de agrupamento.
- Os grupos que estão sendo formados estão sendo conhecidos como Clusters.
- A técnica de clustering está sendo usada em vários campos, como reconhecimento de imagem, filtragem de spam.
- O clustering está sendo usado no Algoritmo de Aprendizado Não Supervisionado em Aprendizado de Máquina, pois pode segregar dados multivariados em vários grupos, sem nenhum supervisor, com base em um padrão comum oculto nos conjuntos de dados.

O que o Algoritmo K-Means ?

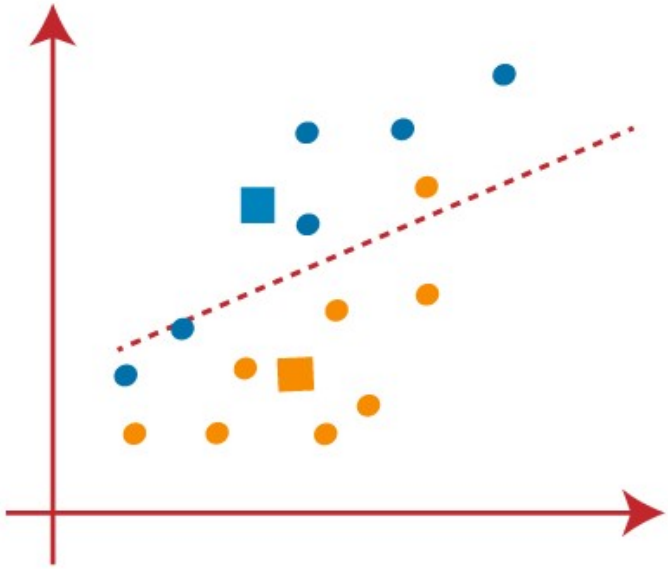
- O algoritmo K-means é um algoritmo iterativo que divide um grupo de n conjuntos de dados em k subgrupos/clusters com base na similaridade e sua distância média do centroide daquele subgrupo/formado em particular.
- K, aqui é o número pré-definido de clusters a serem formados pelo algoritmo.
- Se K=3, significa que o número de clusters a serem formados a partir do conjunto de dados é 3.

Passos do Algoritmo K-Means ?

- O funcionamento do algoritmo K-Means é explicado nas etapas abaixo:
- Passo-1: Selecione o valor de K, para decidir o número de clusters a serem formados.
- Passo-2: Selecione K pontos aleatórios que atuarão como centróides.
- Passo-3: Atribua cada ponto de dados, com base em sua distância dos pontos selecionados aleatoriamente (Centroid), até o centroide mais próximo que formará os clusters predefinidos.
- Passo-4: coloque um novo centroide de cada cluster.
- Etapa 5: Repita a etapa 3, que reatribui cada ponto de dados ao novo centroide mais próximo de cada cluster.
- Etapa 6: Se ocorrer alguma reatribuição, vá para a etapa 4, caso contrário, finalize o algoritmo.

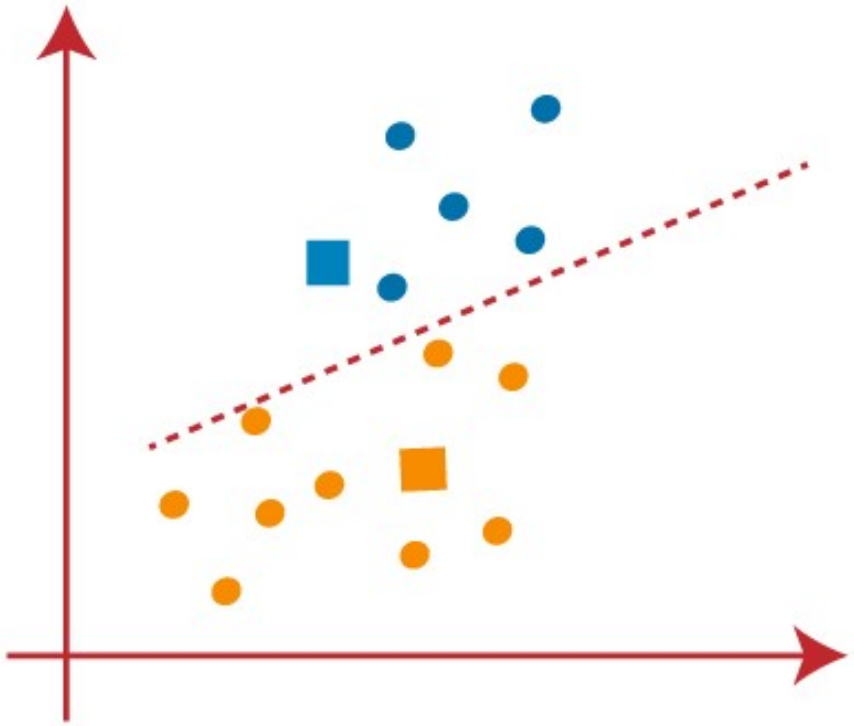
Implementação esquemática do K-Means Clustering ?

- PASSO 1: Vamos escolher o número k de clusters, ou seja, K=2, para segregar o conjunto de dados e colocá-los em diferentes clusters respectivos. Escolheremos alguns 2 pontos aleatórios que atuarão como centróides para formar o cluster.
- PASSO 2: Agora vamos atribuir cada ponto de dados a um gráfico de dispersão com base em sua distância do ponto K ou centroide mais próximo. Isso será feito traçando uma mediana entre ambos os centróides. Considere a imagem abaixo:



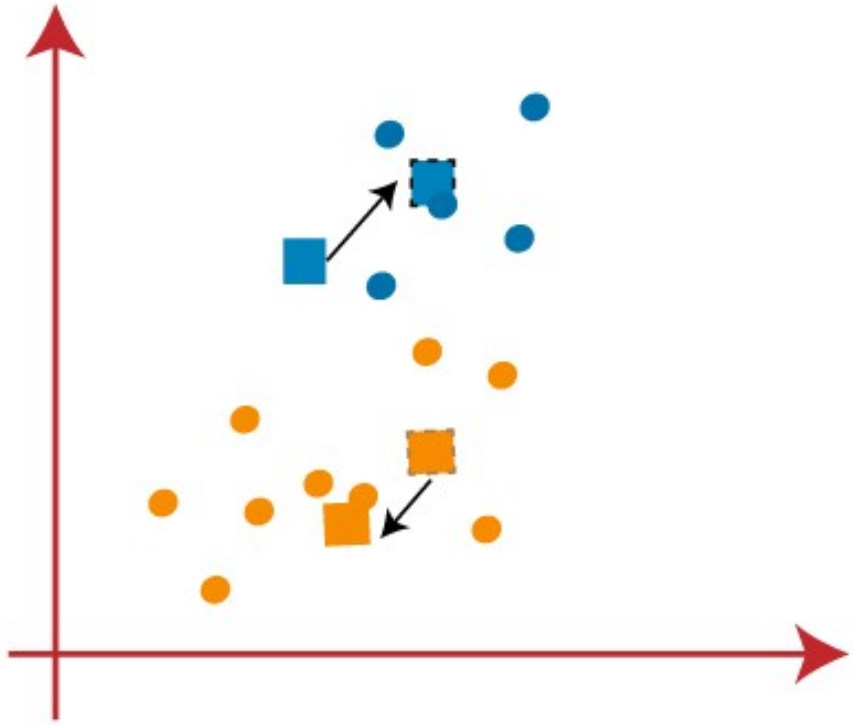
Implementação esquemática do K-Means Clustering ?

- PASSO 3: os pontos do lado esquerdo da linha estão próximos do centróide azul e os pontos à direita da linha estão próximos do centróide amarelo. O da esquerda forma o cluster com o centroide azul e o da direita com o centroide amarelo.



Implementação esquemática do K-Means Clustering ?

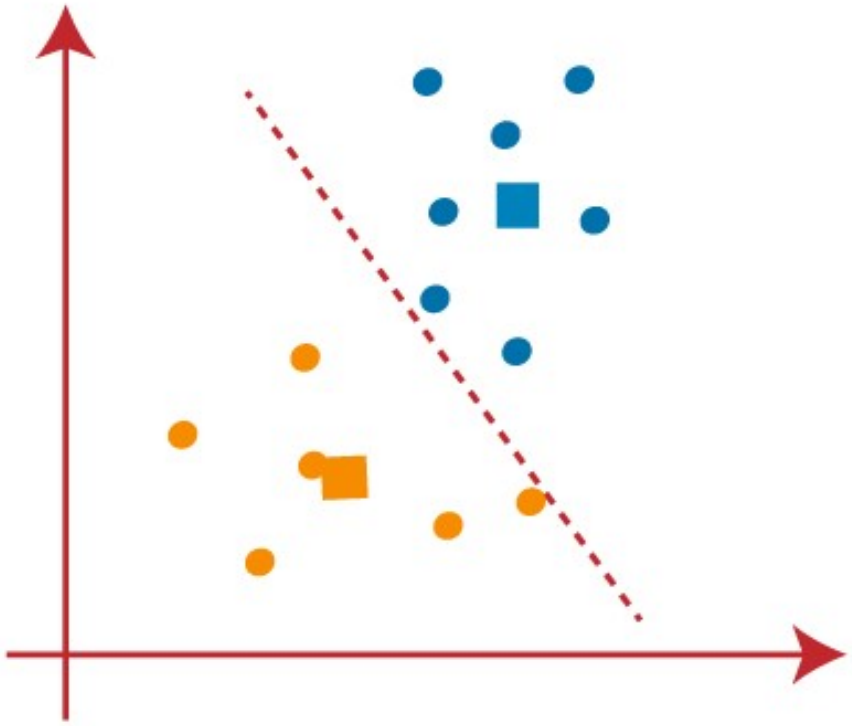
- PASSO 4: repita o processo escolhendo um novo centroide. Para escolher os novos centróides, encontraremos o novo centro de gravidade desses centróides, que está representado abaixo:



INSTITUTO FEDERAL
Sudeste de Minas Gerais

Implementação esquemática do K-Means Clustering ?

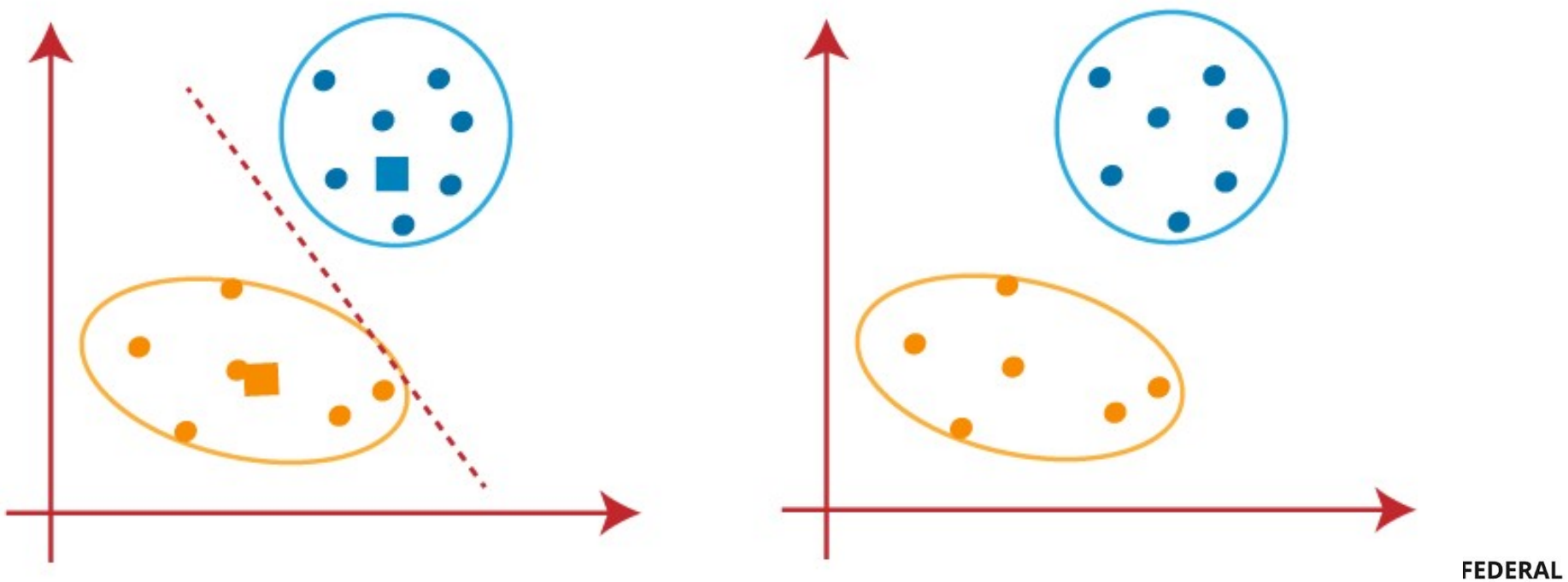
- PASSO 5: Em seguida, vamos reatribuir cada ponto de dados ao novo centroide. Vamos repetir o mesmo processo acima (usando uma linha mediana). O ponto de dados amarelo no lado azul da linha mediana será incluído no cluster azul:



INSTITUTO FEDERAL
Sudeste de Minas Gerais

Implementação esquemática do K-Means Clustering ?

- PASSO 6: Finalmente vamos segregar os pontos com base na linha mediana, de modo que dois grupos estejam sendo formados e nenhum ponto diferente seja incluído em um único grupo. O cluster final que está sendo formado é o seguinte



Escolhendo o número certo de clusters

- O número de clusters que escolhemos para o algoritmo não deve ser aleatório. Cada cluster é formado calculando e comparando as distâncias médias de cada ponto de dados dentro de um cluster a partir de seu centroide.
- Podemos escolher o número certo de clusters com a ajuda do método Within-Cluster-Sum-of-Squares (WCSS).
- WCSS Representa a soma dos quadrados das distâncias dos pontos de dados em cada cluster de seu centroide.
- A ideia principal é minimizar a distância entre os pontos de dados e o centroide dos clusters.
- O processo é iterativo até atingirmos um valor mínimo para a soma das distâncias.

Método do Cotovelo (Elbow Method)

- Para encontrar o valor ideal de clusters, o método do cotovelo segue as etapas abaixo:
- 1) Execute o agrupamento K-means em um determinado conjunto de dados para diferentes valores de K (variando de 1 a 10).
- 2) Para cada valor de K, calcula o valor WCSS.
- 3) Traça um gráfico/curva entre os valores WCSS e o respectivo número de clusters K.
- 4) O ponto agudo da curva ou um ponto (parecendo uma articulação do cotovelo) do gráfico como um braço, será considerado como o melhor/ótimo valor de K.

INSTITUTO FEDERAL
Sudeste de Minas Gerais