K-Means

Guilherme Pereira Paiva
Matrícula 181057600032

30 de junho de 2019

Informações

Para a execução do programa basta executar o arquivo main.py via linha de comando.
Devido à posição inicial das centróides ser aleatórias, a classificação pode sofrer variações
Versão do python utilizada: Python 3.7.3

O trabalho foi dividido nos arquivos:

- read.py: Faz a leitura do arquivo "data.txt" e prepara as demais variáveis;
- clustering.py: Faz o processo de clustering do algoritmo, calcula as distâncias, e reposiciona as centróides para os devidos lugares;
- write.py: Produz o arquivo final "classes.txt" especificando a média de consumo e de emissão de carbono de cada classe;
- main.py: Faz a chamada das funções e mostra o gráfico.

Bibliotecas utilizadas

- Numpy - para cálculos de distância euclidiana e Arrays;
- CSV - "Comma separated values" para trabalhar com o arquivo csv separado por vírgulas;
- Matplotlib - Para gerar o gráfico dos dados após a Clusterização.

Número de Classes

Para determinar o número de classes (K) foi utilizado o método cotovelo (Do inglês *Elbow Method*), que consiste em executar o K-Means variando a quantidade de classes e calcular a soma dos quadrados intra-clusters. Esta soma mede a distância dos pontos observados e os clusters posicionados. A partir destas distâncias observadas, podemos gerar um gráfico para diferentes números de K.

Analisando o gráfico, podemos perceber que a partir de uma certa quantidade de clusters, a distância não diminui tão significativamente, o que significa que neste ponto há o número ideal de clusters.

Para este trabalho foi utilizado K = 4, a partir da análise do gráfico 1.

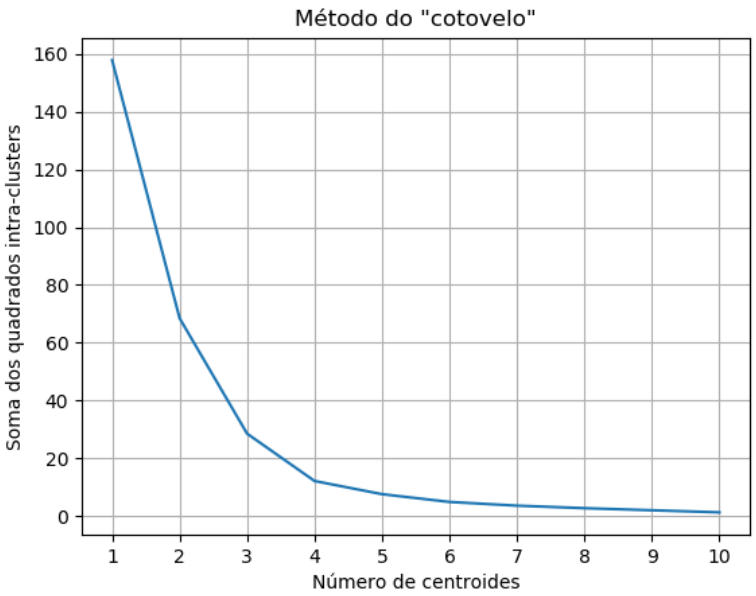

Figura 1: Gráfico do Método Cotovelo.