Introdução

Este relatório apresenta uma análise da complexidade do algoritmo K-Means implementado em Python e orientado a objeto. Esse trabalho é referente ao componente curricular Engenharia de Programas. O K-Means é um algoritmo de clustering amplamente utilizado para particionar um conjunto de dados em K grupos, minimizando a variação dentro dos clusters ao longo de suas iterações. O estudo visa compreender como a complexidade do algoritmo varia com os parâmetros ‘n’ e ‘K’.

Metodologia

Foi realizada a geração de dados usando 8 imagens escolhidas aleatoriamente e fazendo a clusterização para K = 2 até K=7 para a análise de suas variáveis como tempo médio de iterações, desvio padrão e variabilidade. Ainda, foram feita R repetições para obter valores mais estáveis.

Para cada execução foram armazenados os valores de tempo médio, número de iterações, quantidade de clusters, desvio padrão e variabilidade. Em seguida, foram gerados gráficos para identificar correlações entre variáveis e realizar uma regressão linear entre variáveis para obter uma função que seja possível estimar o tempo médio gasto dado um tamanho n do problema.

Ainda, foram armazenados os valores dos slopes por análise de um certo número de clusters para se obter a complexidade da implementação de KMeans em relação às variáveis de n e K.

Resultado:

Por meio dos resultados dos experimentos e dos gráficos dos notebooks para K=2 até K=7 fica claro a relação entre os valores de tempo (tempo médio para cada iteração) e o tamanho do problema (quantidade de pixels de uma imagem), podendo-se observar uma função linear com slope crescente conforme o número de clusters.

Além disso, também foi identificado que não há uma relação entre as variáveis de número de iterações e o tamanho do problema. Pois os pontos no gráfico de dispersão não seguem uma regularidade para todos os K. Ainda, para todas as análises das imagens, a variabilidade permaneceu abaixo de 15% dando confiabilidade para o resultado encontrado. Por fim, foi realizado uma última plotagem de gráfico para verificar o nível de complexidade de O(KN) sendo feito a partir dos valores obtidos dos slope da dependência tempo em função de n de cada um das funções de regressão de cada número de clusters, podendo-se observar uma complexidade linear para esse código de KMeans analisado.