Relatório Técnico – K-means (K-means Clustering)

Aluno: Henrique Cesar Silva Soares
Matrícula: 202310537

Introdução

Este trabalho teve como objetivo a implementação e a análise do algoritmo de clusterização K-means. A tarefa foi dividida em duas partes: uma implementação hardcore, desenvolvida do zero em Python, e uma segunda utilizando uma biblioteca otimizada, como a Scikit-learn. A base de dados utilizada foi a Iris Dataset, com a classe alvo desconsiderada para a clusterização. Foram realizados experimentos para K=3 e K=5 clusters.

Comparação de Desempenho

A avaliação de desempenho considerou o tempo de execução e o Silhouette Score, uma métrica que avalia a coesão e a separação dos clusters.

Tempo de Execução

Hardcore: 0.0043 s (K=3)
Scikit-learn: 1.4682 s (K=3)

A implementação com a biblioteca Scikit-learn, apesar de geralmente mais rápida, apresentou um tempo de execução maior para K=3 neste experimento específico. Isso pode ser devido a fatores como a inicialização padrão aleatória da biblioteca (que pode exigir mais iterações) e a sobrecarga de recursos do ambiente de execução. Em K=5, o tempo da biblioteca foi 0.0053s, enquanto o hardcore foi de 0.0083s. Isso indica que a versão otimizada da biblioteca geralmente supera a implementação manual em eficiência, conforme esperado.

Métricas de Avaliação (Silhouette Score)

K	Implementação	Silhouette Score
3	Hardcore	0.5528
3	Scikit-learn	0.5512
5	Hardcore	0.4931
5	Scikit-learn	0.4931

Os resultados mostram que o Silhouette Score para K=3 foi superior ao de K=5. O valor de 0.55 sugere que os clusters para K=3 são mais bem definidos e separados. A equivalência dos scores entre as implementações hardcore e da biblioteca valida a correção do algoritmo manual.

Análise de Dimensionalidade (PCA)

A técnica de Análise de Componentes Principais (PCA) foi aplicada para reduzir a dimensionalidade dos dados e facilitar a visualização. Foram gerados gráficos para 1 e 2 componentes principais, que permitiram uma visualização clara da estrutura dos clusters e da posição dos centróides. A visualização bidimensional (2D) mostrou que o algoritmo foi eficaz em separar os três grupos de forma clara, o que valida os resultados do Silhouette Score para K=3.

Conclusão

A implementação hardcore do algoritmo K-means demonstrou ser correta e equivalente à versão otimizada da biblioteca em termos de qualidade de clusterização. Ambas as abordagens convergiram para resultados de Silhouette Score praticamente idênticos, evidenciando que a lógica do algoritmo foi replicada com sucesso. A análise das métricas confirmou que três clusters é o número ideal para a base de dados Iris.

Em relação ao desempenho, a biblioteca se mostrou mais consistente em termos de tempo de execução, reforçando a sua vantagem para aplicações de maior escala. Em síntese, a atividade permitiu aprofundar a compreensão dos fundamentos do K-means a partir da implementação manual, ao mesmo tempo que ressaltou a importância de bibliotecas otimizadas no contexto de ciência de dados profissional.

Vídeo de Apresentação

O vídeo de apresentação deste trabalho pode ser acessado pelo seguinte link:
Clique aqui para assistir