Relatório Final do Trabalho Prático relacionado ao algoritmo K-Means

GCC128 -Inteligência Artificial:

Sistemas de informação – 14A

Gustavo de Jesus Teodoro - 202311146
Thiago Lima Pereira - 202310057

Lavras - MG

Relatório — K-means

Neste trabalho foi aplicada a técnica de **clusterização K-means** sobre a base de dados *Iris*, com o objetivo de comparar o desempenho entre uma **implementação manual (hardcore)** e a implementação já disponível na biblioteca `scikit-learn`. O algoritmo foi executado para diferentes valores de `k` (3 e 5), permitindo avaliar o impacto da escolha do número de clusters no desempenho.

A versão hardcore foi desenvolvida do zero em Python, utilizando apenas bibliotecas básicas para manipulação de dados e cálculo de distâncias. Já a versão do sklearn foi instanciada com múltiplas inicializações (`n_init=20`), garantindo maior robustez e estabilidade dos resultados. Ambas as abordagens foram avaliadas segundo **inércia**, **silhouette score**, número de iterações, tempo de execução e tamanhos de clusters.

Resultados obtidos:

- Para **k=3**, a implementação hardcore apresentou inércia ≈ 140.0 e silhouette ≈ 0.463, enquanto o sklearn alcançou inércia ≈ 139.8 e silhouette ≈ 0.460.

- Para **k=5**, os valores de silhouette foram menores (≈ 0.39 no hardcore e ≈ 0.35 no sklearn), indicando clusters menos bem definidos.

- Em ambos os casos, o tempo de execução foi da ordem de centésimos de segundo, com ligeira vantagem para o hardcore em k=3, mas convergência em menos iterações pelo sklearn.

Análise comparativa:
Os resultados confirmam que o valor mais adequado de k é **3**, coerente com as três espécies de flores da base Iris. As duas implementações apresentaram métricas próximas, validando a corretude da versão manual. O sklearn, no entanto, se destacou por convergir em menos iterações e oferecer maior estabilidade devido às suas otimizações internas e múltiplas inicializações, enquanto o hardcore, apesar de mais simples, é útil para fins didáticos e de compreensão do algoritmo.

Redução de dimensionalidade (PCA):
Com o melhor k=3, aplicou-se PCA para projetar os dados em espaços de 1 e 2 componentes. A versão com **1 componente** explicou cerca de 73% da variância, mas apresentou sobreposição entre os clusters. Já a projeção com **2 componentes** explicou aproximadamente 96% da variância total, evidenciando uma separação mais clara entre os grupos e facilitando a visualização dos centróides.

Conclusão:
O trabalho demonstrou com sucesso a aplicação do K-means em duas abordagens distintas. Conclui-se que o valor ótimo de clusters é 3, e que a versão sklearn apresenta desempenho mais eficiente e consistente, enquanto a versão hardcore reforça o entendimento dos conceitos fundamentais. A análise com PCA confirmou visualmente a separação dos clusters e a adequação da escolha de k.