LISTA 8 IA
Bruno Braga
767681

Link do código: https://github.com/Bruno0926/IA/blob/main/Quest%C3%A3o%20lista%208.py
“Questão lista 8.py”

Questão 1:
1) Identificação de outlier e normalização (presentes no código).

2) Encontrar e Avaliar Agrupamentos, Métodos K-means (presentes no código).

3) Explicação das métricas:

Silhouette Score:

O Silhouette Score é uma métrica que avalia a qualidade dos agrupamentos formados. Ele mede quão semelhantes os pontos são ao seu próprio cluster em comparação com outros clusters. A equação do coeficiente de Silhouette para um ponto iii é dada por:

$s(i)=b(i)−a(i) / max(a(i),b(i))$

- a(i) é a distância média entre o ponto i e todos os outros pontos no mesmo cluster.

- b(i)b(i)b(i) é a menor distância média entre o ponto i e todos os pontos em qualquer outro cluster.

- O valor de s(i)s(i)s(i) varia de -1 a 1: Valores próximos de 1 indicam que o ponto está bem ajustado ao seu próprio cluster e mal ajustado aos outros clusters. Valores próximos de 0 indicam que o ponto está em cima da fronteira ou sobreposto entre dois clusters. Valores negativos indicam que o ponto foi mal agrupado.

Método Elbow:

O método Elbow é utilizado para encontrar o número ideal de clusters (k) ao analisar a soma das distâncias quadradas dentro do cluster (WCSS - Within-Cluster Sum of Squares). Para isso, plotamos o WCSS para diferentes valores de k e procuramos um ponto no gráfico onde a diminuição da WCSS começa a se estabilizar, formando um "cotovelo". Este ponto é o número ideal de clusters. A soma das distâncias quadradas dentro do cluster é calculada como:

$WCSS=∑i=1k∑x∈Ci||x−μ||²$

- Ci é o i-ésimo cluster.

- μi é o centróide do i-ésimo cluster.

- x são os pontos de dados pertencentes ao cluster CiC_iCi.

4) Métrica Adicional: Índice de Davies-Bouldin

O índice de Davies-Bouldin (DB) é outra métrica usada para avaliar a qualidade dos agrupamentos. Ele mede a média das razões entre a soma das variâncias intra-cluster e a distância inter-cluster. A equação para o índice de Davies-Bouldin é:

$DB=1/k\sum i=1k\ maxj\neq i((si+sj)\ /\ dij)$

- si é a média das distâncias intra-cluster para o cluster i.

- dij é a distância entre os centroids dos clusters i e j.

- O índice de Davies-Bouldin busca minimizar a distância intra-cluster e maximizar a distância inter-cluster. Valores menores do índice indicam melhores partições dos dados.

5) Visualização também presente no código.

Para visualizar as instâncias incorretamente agrupadas, comparamos os clusters obtidos com as classes reais. Foram usados gráficos de dispersão para observar a correspondência entre as classes verdadeiras (setosa, virginica, versicolor) e os clusters formados pelo K-means.

Ao comparar os clusters com as classes reais, podemos identificar instâncias que foram agrupadas incorretamente. Isso pode ser feito visualmente, onde cada ponto é colorido de acordo com sua classe real e estilizado de acordo com seu cluster atribuído. A partir disso, podemos discutir se o K-means conseguiu capturar bem as estruturas naturais dos dados ou se houve confusão significativa entre as classes.

Por exemplo, podemos observar que o K-means pode confundir instâncias das classes virginica e versicolor devido à sua proximidade no espaço das características. No entanto, a classe setosa pode ser facilmente separável, resultando em uma correspondência mais precisa para essa classe específica.

Ao avaliar os agrupamentos utilizando as métricas Silhouette, Elbow e Davies-Bouldin, conseguimos ter uma visão abrangente da qualidade dos agrupamentos e das possíveis limitações do método K-means no contexto dos dados analisados.

6) Relatório presente no código.

Questão 2:

Está tudo no código: https://github.com/Bruno0926/IA/blob/main/Lista%208%20questao%202.py

“Lista 8 questao 2.py”

Interpretação dos resultados, fora do console.

Accuracy: A precisão do Naive Bayes é de 85%, enquanto a do SVM é de 90%. Isso indica que o SVM tem um desempenho melhor na classificação dos textos neste caso específico.

Precision, Recall, e F1-score: Estes valores são fornecidos para cada classe (0 e 1). Valores mais altos indicam um melhor desempenho do modelo para aquela métrica.

Support: Mostra o número de instâncias de cada classe no conjunto de teste.