HyperionDev

Unsupervised Learning – Hierarchical Clustering

Model Answer Approach

Visit our website

Auto-graded task

This approach involves using hierarchical clustering techniques on the Iris dataset, focusing on two features: Sepal Length and Sepal Width. The data is first scaled using `StandardScaler` to ensure that all features have equal weight in the clustering process. Dendrograms are generated using both single and complete linkage methods, combined with two distance metrics: Euclidean and city block (Manhattan). These dendrograms visually represent how the hierarchical clustering process merges data points, with the distance between merges reflecting their similarity.

The `AgglomerativeClustering` algorithm is used to create clusters, where the optimal number of clusters is determined based on the dendrogram. For evaluation, the silhouette score is calculated to assess the quality of the clustering solution. This score ranges between -1 and 1, where a higher value indicates well-separated clusters. In this case, the silhouette score for Sepal Length and Sepal Width was moderate, suggesting some overlap between clusters.

One possible pitfall in this task is the reliance on only two features for clustering, which may not provide a clear separation of clusters. As highlighted, a low-to-moderate silhouette score indicates that the chosen features might not be ideal. Using additional features from the dataset or trying different combinations could yield better-defined clusters and higher silhouette scores. Therefore, the clustering outcome can be sensitive to feature selection, and testing with all available features may lead to a more accurate model.