109550198 卜銳凱

Introduction to Data Science
HW8
Report

This report describes the analysis of the dataset from file ‘'Mall_Customers.csv'.

K-Means Clustering

Importing Libraries: The code begins by importing necessary libraries: numpy for numerical operations, matplotlib.pyplot for plotting, and pandas for data handling.

Importing the Dataset: The dataset 'Mall_Customers.csv' is loaded using pandas. The code then extracts the third and fourth columns, likely representing key features like 'Annual Income' and 'Spending Score', for clustering.

Elbow Method for Optimal Clusters:
The elbow method is employed to find the optimal number of clusters. It involves running the K-Means clustering for a range of cluster numbers (1 to 10 in this case) and calculating the Within-Cluster-Sum-of-Squares (WCSS) for each.
The results are plotted, showing how WCSS changes with the number of clusters. The 'elbow point' in this plot indicates the optimal cluster number where adding another cluster doesn't significantly improve the variance explained by the model.

Training K-Means Model:
A K-Means model is trained with 5 clusters (as suggested by the elbow method), using the 'k-means++' initialization and a fixed random state for reproducibility.
The fit_predict method assigns each data point to one of the 5 clusters.

Visualizing K-Means Clusters:
The clusters are visualized in a scatter plot, with different colors representing different clusters. The centroids of the clusters (the mean of the points in each cluster) are also plotted in yellow.

Hierarchical Clustering

Dendrogram for Optimal Clusters:
The code transitions to Hierarchical Clustering, beginning with the creation of a dendrogram using the Ward method. This is a visual tool to help determine the number of clusters by observing where the merges of clusters occur at a significant increase in distance.
Training Hierarchical Clustering Model:

The AgglomerativeClustering from scikit-learn is used with 5 clusters, 'euclidean' affinity, and 'ward' linkage, as suggested by the dendrogram.
The fit_predict method is again used to classify each data point into one of the clusters.

Visualizing Hierarchical Clusters:
Similar to the K-Means section, the hierarchical clusters are visualized using scatter plots with distinct colors for each cluster.

Summary
This notebook is a thorough application of two clustering techniques to a customer dataset, likely aimed at understanding customer segments based on their income and spending behavior. The use of both the elbow method and dendrograms for determining the optimal number of clusters demonstrates a comprehensive approach to cluster analysis. The visualizations provide an intuitive understanding of the data's structure and the characteristics of each identified customer segment.

Discuss possible problems you plan to investigate for future studies

- Neural Network and DeepLearning