Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

Faculty of Cyber Physical System
Department of Internet of Things and Robotics Engineering
B.Sc. in Internet of Things and Robotics Engineering
Course Title: Data Science
Course Code: IoT 4313
Assignment 02: Clustering

Submitted By:
Mahbuba Tabassum
Id: 1901026
Session: 2019-20

Submitted to:
Nurjahan Nipa
Lecturer,
Department of IRE, BDU.

Date of Submission: 14 October 2023

Introduction: A data analysis technique called clustering brings together comparable data points. It is employed to find structures and trends in datasets devoid of labeled information. There are numerous clustering approaches, uses, and difficulties. For activities like customer segmentation, anomaly detection, and more, the objective is to find natural groupings.

Here is a detailed explanation of my approaches and the results obtained for all of the clustering algorithms required in Parts A, B, and C.

Part A: K-Means Clustering

K-Means clustering is a centroid-based algorithm that partitions data into K clusters, where K is predefined.

Approach:

- In the code, we loaded the **Mall_Customer** dataset and extracted the features (Age, Annual Income, and Spending Score) to be used for clustering.
- We applied K-Means clustering with a range of K values from 1 to 15.
- For each K, we calculated the Sum of Squared Errors (SSE), which is a measure of how spread out the data points are within each cluster.
- We plotted the SSE values against the number of clusters (K) and looked for an "elbow" point to determine the optimal number of clusters.

Results:

- The Elbow Method identified K=5 as the optimal number of clusters.
- K-Means grouped customers into five clusters based on similarities in age, annual income, and spending score.
- These clusters provide valuable insights for marketing and business strategy.

Part B: Hierarchical Clustering

Hierarchical clustering is a method that builds a hierarchy of clusters by successively merging or splitting existing clusters.

Approach:

- We performed hierarchical clustering using the AgglomerativeClustering algorithm, specifying the number of clusters (n_clusters) and linkage type (ward linkage in this example).
- We also created a dendrogram to visualize the hierarchical structure of clusters.

Results:

- Hierarchical Clustering produced a dendrogram showing hierarchical relationships between clusters.
- The number of clusters can be determined by cutting the dendrogram at a chosen level.
- This method is useful for understanding the hierarchical structure of the data.

Part C: Density-Based Clustering

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that identifies clusters based on the density of data points.

Approach:

- We standardized the features using StandardScaler to ensure that they have similar scales.
- We applied DBSCAN with parameters such as eps (maximum distance between two samples in the same neighborhood) and min_samples (minimum number of samples in a neighborhood to be considered a core point).

Results:

- DBSCAN clustered data points based on local density, producing clusters of varying shapes and sizes.
- Outliers were labeled as noise (-1).
- The granularity of clusters was controlled by adjusting the epsilon value.
- This approach is suitable for irregularly shaped clusters.