Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

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

Assignment-2 (Clustering)

Submitted By:
Marshia Muntaka
Id: 1901021
Session: 2019-20

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

Date of Submission: 14 October, 2023**.**

- **Clustering:**

  A machine learning approach called clustering combines related data points based on their intrinsic properties. To make complicated data structures easier to study and comprehend, it is used to find patterns, correlations, or natural divisions within a dataset.

  - ✓ **Example:**

    Imagine you have a dataset of customer information, including age and annual income. By applying clustering algorithms, you can group customers into distinct segments based on similarities. For instance, you might discover two clusters: "young, low-income" and "middle-aged, high-income" customers. This information can help businesses target their marketing strategies more effectively for each group.

- **Types of Clustering:**

  Clustering is a broad category of machine learning techniques, and there are several types of clustering algorithms. Some common types include:

  - ✓ **K-Means Clustering:** Divides data into K distinct, non-overlapping clusters.
  - ✓ **Hierarchical Clustering:** Builds a hierarchy of clusters by either merging or splitting data points.
  - ✓ **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Identifies dense regions of data points as clusters, accommodating varying shapes.
  - ✓ **Agglomerative Clustering:** A bottom-up approach, starting with individual data points and merging them into clusters.
  - ✓ **Mean Shift Clustering:** Identifies modes or peaks in the data distribution and assigns points to these modes.
  - ✓ **Spectral Clustering:** Uses spectral techniques to group data points based on the similarity of their pairwise relationships.
  - ✓ **Gaussian Mixture Models (GMM):** Assumes that data points are generated from a mixture of several Gaussian distributions.
  - ✓ **Fuzzy Clustering:** Assigns data points to multiple clusters with varying degrees of membership.

  The choice of clustering algorithm depends on the nature of the data and the specific problem you want to solve. Different algorithms have different strengths and weaknesses, making them suitable for various scenarios.

Here is a comprehensive description of my methods and the outcomes attained for each clustering algorithm needed in Parts A, B, and C.

PART-A
K-means Clustering

- **K-means Clustering:**

  K-means clustering is an unsupervised machine learning technique that divides a dataset into K distinct clusters, aiming to minimize the within-cluster variance. It assigns data points to the cluster with the nearest centroid, iteratively optimizing cluster boundaries. It's widely used for data segmentation and pattern recognition in various fields, such as customer segmentation and image compression.

- **Method:**

  • To extract the characteristics (Age, Annual Income, and Spending Score) for clustering, we loaded the Mall_Customer dataset into the code.

  • We used K-Means clustering with a range of K values from 1 to 15.

  • To determine how evenly distributed the data points are inside each cluster, we computed the Sum of Squared Errors (SSE) for each K.

  • To estimate the ideal cluster size, we plotted the SSE values against the number of clusters (K) and searched for a "elbow" point.

- **Results:**

  - The Elbow Method determined that K=5 is the ideal number of clusters.
  - Based on similarity in age, annual income, and expenditure score, K-Means divided customers into five groups.
  - These clusters offer insightful information for marketing and company planning

PART-B

Hierarchical Clustering

- **Hierarchical Clustering:**

Hierarchical clustering is a data analysis technique that arranges data points into a hierarchical structure, forming a tree-like diagram called a dendrogram. It iteratively merges or divides clusters based on the similarity between data points, allowing for the exploration of different levels of granularity. Hierarchical clustering is used to reveal relationships, identify natural groupings, and visualize data structures, making it valuable in fields like biology, social sciences, and image analysis. Its flexibility in choosing cluster levels and its visualization capabilities make it a powerful tool for understanding complex data relationships.

- **Method:**

•By specifying the number of clusters (n_clusters) and type of linkage (ward linkage in this case), we carried out hierarchical clustering using the Agglomerative Clustering Algorithm.

• We also produced a dendrogram to show the clusters' hierarchical nature.

- **Results:**

•The dendrogram generated by hierarchical clustering shows the hierarchical relationships between clusters.

•How many clusters there are can be determined by cutting the dendrogram at a specific level.

•This method will assist you in understanding the hierarchical structure of the data.

Part C
Density-Based Clustering

- **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**

  DBSCAN (Density-Based Spatial Clustering of Applications with Noise) is a clustering algorithm that identifies clusters in a dataset based on the density of data points. It groups data points together if they are sufficiently close to each other and have a minimum number of neighboring data points within a specified radius. Data points in sparse regions are labeled as noise, distinguishing outliers. DBSCAN is effective for discovering clusters of arbitrary shapes and is valuable in various applications, such as anomaly detection and spatial data analysis, due to its ability to handle varying cluster densities.

- **Method:**

  •We used StandardScaler to standardize the features and make sure their scales are comparable.

  • We used the DBSCAN algorithm with parameters like eps (the maximum distance between two samples in the same neighborhood) and min_samples (the bare minimum of samples in a neighborhood to be regarded as a core point).

- **Results:**

  •Data points were clustered by DBSCAN based on local density, resulting in clusters of different sizes and shapes.

  • Outliers were classified as noise (-1).

  •By changing the epsilon value, the granularity of the clusters could be managed.

  •This strategy works well with clusters that have a variety of shapes.

- **Link of GitHub:** https://github.com/Muntaka21/Clustering