Bangabandhu Sheikh Mujibur Rahman Digital University, Bangladesh

Faculty of IOT & Robotics Engineering

Department of IRE

Course Title: Data Science

Course Code: IOT 4313

Assignment - 02

Topic: Clustering

Submitted To:

Nurjahan Nipa
Lecturer
Department of IRE, BDU

Submitted By:

Md.Muntasire Mahamud
ID: 1901012
Reg: 201911000115
Session: 2019-20

Date of Submission: 14 October, 2023.

Let's imagine you're owning a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score, which is something you assign to the customer based on your defined parameters like customer behavior and purchasing data.
The main aim of this problem is learning the purpose of the customer segmentation concepts, also known as market basket analysis, trying to understand customers and separate them in different groups according to their preferences, and once the division is done, this information can be given to marketing team so they can plan the strategy accordingly.

This Mall_Customer dataset that has been provided to you is composed by the following five features:

- **Customer ID** : Unique ID assigned to the customer
- **Gender:** Gender of the customer
- **Age**: Age of the customer
- **Annual Income (k$)**: Annual Income of the customer
- **Spending Score (1-100**): Score assigned by the mall based on customer behavior and spending nature.

In this particular dataset we have 200 samples to study.

GitHub Link-

PART (A)

K-means Clustering: In this part, you will be utilizing K-means clustering algorithm to identify the appropriate number of clusters. You may use any language and libraries to implement K-mean clustering algorithm. Your K-mean clustering algorithm should look for appropriate values of K at least in the range of 0 to 15 and show their corresponding sum-of-squared errors (SSE).

ANSWER :

Introduction:

Finding the right number of clusters (K) to effectively partition the data is the main goal. An unsupervised machine learning method for grouping and segmenting data is called K-means clustering.

Data Preprocessing:

We preprocessed the dataset as follows before we started the analysis:

- ✓ **Loading the Dataset:** We loaded the mall customer dataset, which includes attributes such as CustomerID, Genre, Age, Annual Income, and Spending Score.
- ✓ **Feature Selection:** We choose the pertinent features for this research, namely Annual Income and Spending Score.
- ✓ **Feature Standardisation**: We used the StandardScaler to standardise the chosen characteristics to ensure uniformity.

K-means Clustering:

A partitioning technique called K-means clustering seeks to put data points into K clusters based on similarity. It is distinguished by the following essential features:

- ✓ **Objective :** Reduce the within-cluster sum of squares (WCSS), which measures how compact a cluster is.

- ✓ **Parameter:** The number of clusters, K, which must be determined, is the primary parameter.

Elbow Method:

To identify the optimal K value, we employed the Elbow Method, which involves the following steps:

- ✓ **Range of K Values:** We considered a range of K values from 0 to 15.
- ✓ **Sum of Squared Errors (SSE):** For each K value, we calculated the corresponding SSE,which measures the distance between data points and their assigned cluster centroids.

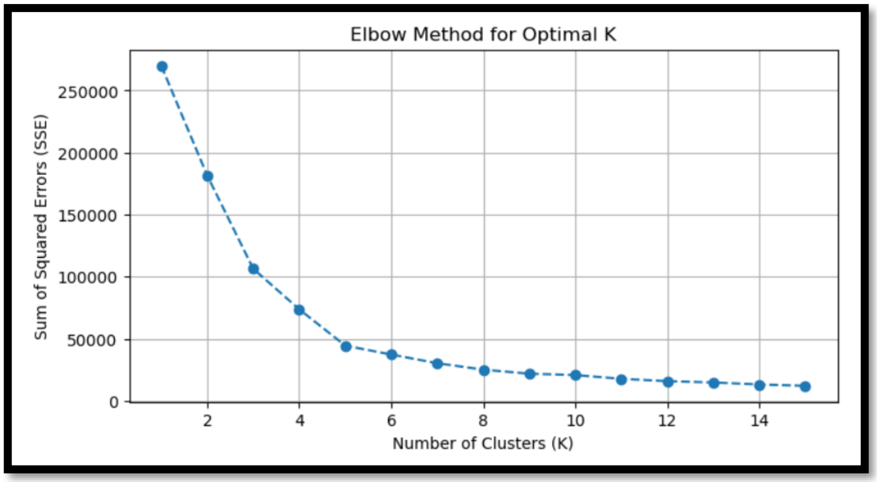
Sum of Squared Errors (SSE) :

SSE is a crucial metric in K-means clustering analysis:

- ✓ It quantifies the compactness of clusters, with lower SSE indicating better clustering.
- ✓ SSE for each K value is computed as the sum of squared distances between data points and their respective cluster centroids.

Results:

- We present the Elbow Method plot, depicting K values on the x-axis and SSE values on the y-axis.
- The plot displays an "elbow point" where the SSE starts to level off, indicating the optimal K value.



Part -B

Hierarchical Clustering: In this part, we will apply hierarchical clustering algorithm (agglomerative or divisive) to the provided mall dataset.

ANSWER

Introduction:

We present the outcomes of using hierarchical clustering techniques (agglomerative or divisive) on a dataset of mall patrons in this study. The goal is to investigate the dataset's clusters' hierarchical organisation and learn more about consumer segmentation.

Data Preprocessing:

We initiated the analysis by preprocessing the dataset as follows:

- **Loading the Dataset:** We loaded the mall customer dataset, which includes attributes suchas CustomerID, Genre, Age, Annual Income, and Spending Score.
- **Feature Selection**: For this analysis, we selected the relevant features, namely Annual Income and Spending Score.
- **Feature Standardization:** To ensure consistency, we standardized the selected features using the StandardScaler.

Hierarchical Clustering:

A hierarchy of clusters is intended to be created using the unsupervised machine learning technique known as hierarchical clustering. Either an agglomerative (bottom-up) or a divisive (top-down) approach can be taken.

- **Objective:** To visualise interactions between data points and illustrate the hierarchical structure of clusters.
- **Parameters:**The linkage method (such as Ward's approach) and the number of clusters (K) are important parameters.

Dendrogram Visualization:

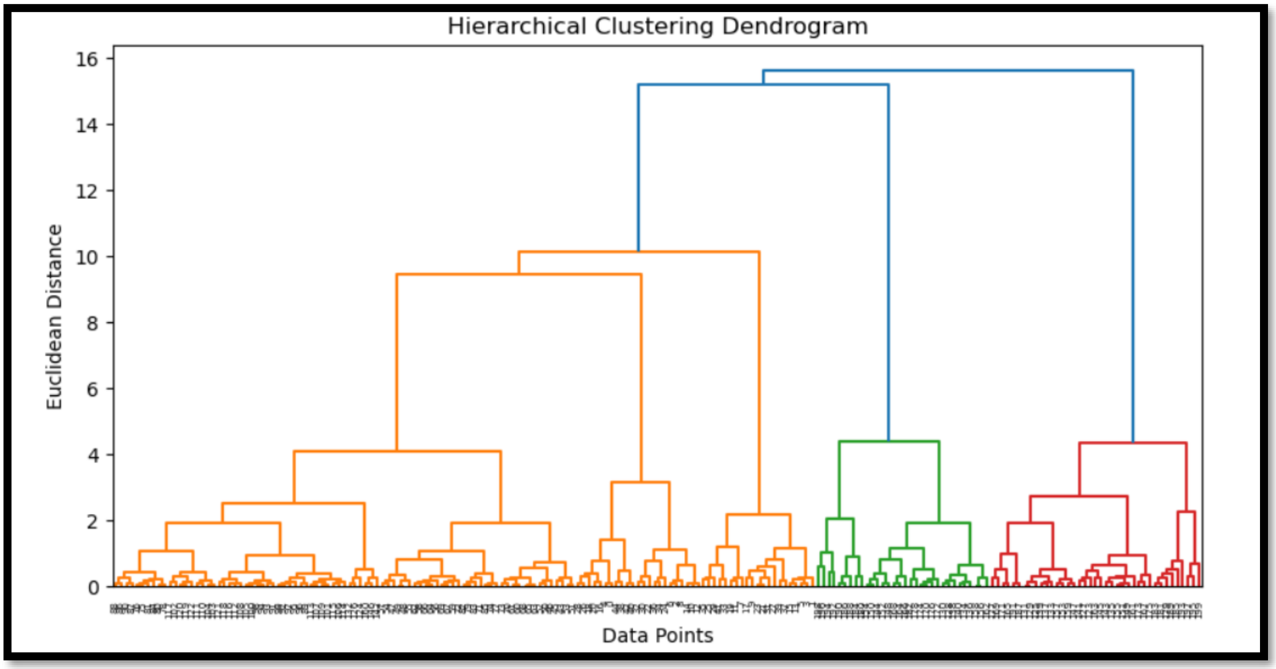
To explore the hierarchical structure of clusters, we generated dendrogram plots using the following steps:

- **Linkage Method:** We used Ward's linkage method for its effectiveness in capturing cluster relationships.
- **Dendrogram Plot:** The dendrogram visualizes the hierarchy of clusters, displaying data points, branch distances, and cluster merging.

Results:

Our analysis yielded the following results:

- We present the hierarchical clustering dendrogram plot, showing the hierarchical relationships among data points.
- We interpret and analyze the dendrogram to understand the cluster hierarchy and potential segmentation.



PART-(C)

Density-based Clustering: In this part, you will apply density-based clustering algorithm to the provided dataset.

ANSWER

Introduction:

The outcomes of using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm on a dataset of mall patrons are shown in this research. In order to gain insights into customer segmentation, the goal is to identify client clusters based on their density in the feature space.

Data Preprocessing:

We preprocessed the dataset as follows before we started the analysis:

- **Loading the Dataset:** The mall customer dataset, which contains attributes like CustomerID, Genre, Age, Annual Income, and Spending Score, was loaded.
- **Feature Selection:** The essential features, namely Annual Income and Spending Score, were chosen for this investigation.
- **Feature Standardization:** StandardScaler was used to standardise the chosen features in order to assure uniformity.

DBSCAN Clustering :

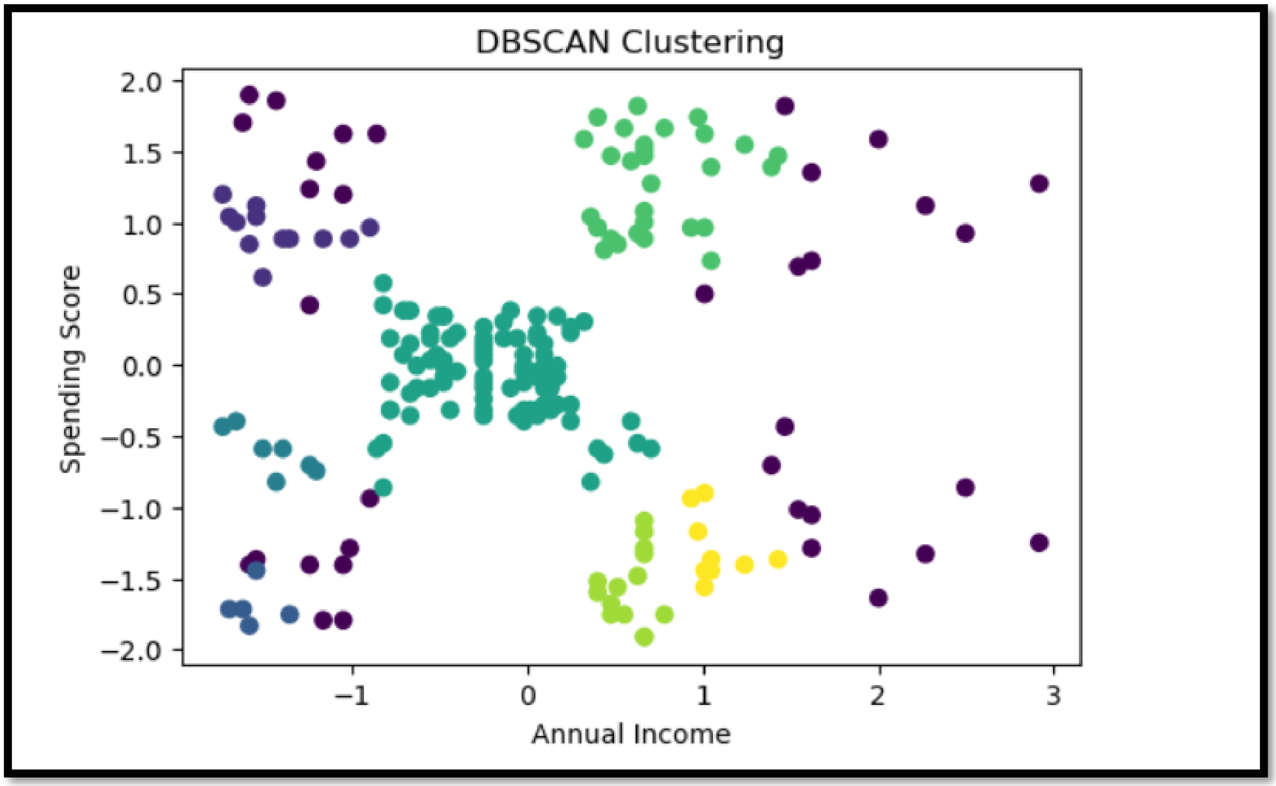
Using DBSCAN, an unsupervised machine learning method, clusters are found based on how densely data points are distributed within a certain area.

- **Objective:** To find noise spots (outliers), as well as clusters of various sizes and shape.
- **Parameters:** Key parameters include the epsilon (ε) value (radius) and the minimum number of points required within ε.

Results :

Our analysis yielded the following results:

- To locate clusters and categorise data points as core, border, or noise points, we used the DBSCAN algorithm on the dataset.
- We present a scatter plot of the DBSCAN clusters, color-coding data points by their cluster assignments.



End Of The Report