Bangabandhu Sheikh Mujibur Rahman Digital University

Course Code : Data Science

Course Title : IoT 4313

Assignment : 02

Assessment on : Clustering

Submitted To : Nurjahan Nipa,

Lecturer,

Department of ICT , BDU

Submitted By : Jul Jalal Al-Mamur Sayor

ID **:** 1901029

Session **:** 2019-2020

3rd year 2nd semester

Date **:13th October 2023**

Introduction : Clustering is a technique used in data science to group similar data points together based on certain features or characteristics. It is an unsupervised learning method where the algorithm tries to find patterns in the data without any prior information about the groups. Clustering is widely used in various fields, including machine learning, data analysis, pattern recognition, and image analysis. There are several types of clustering algorithms in data science.

In this assignment we will deploy **K-means , hierarchical and DBSCAN clustering** and will also discuss their outcome and working procedure .

Software Used : 1. Anaconda Navigator

2. Jupyter Notebook

Files are uploaded to my GitHub , please check out the link below :

GitHub Link- https://github.com/MamurSayor/DS_Task2_1901029

** In this documentation, I mainly focus on the theoretical and mathematical explanation of **K-means , hierarchical and DBSCAN clustering** , In the notebook I explained each important block of codes by commenting . **

Pre-processing for **K-means , hierarchical and DBSCAN clustering :**

i. Import Necessary Libraries
ii. Load the csv dataset.
iii. Null Checking
iv. Checking for the better combinations among all features to cluster .
v. Normalizing

PART-(A)

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

Answer :

K-means Clustering works by trying to partition the data into K distinct, non-overlapping clusters, where each data point belongs to the cluster with the nearest mean.

Now let's know from **the theoretical** point of view how K-means Clustering works:

Certainly, let's explain how K-Means works **mathematically** step by step:

1. **Initialization:**
    - Choose the number of clusters, K.
    - Initialize K cluster centroids, denoted as C1,C2….Ck, either randomly or by some other method.
2. **Assignment Step (Data Point to Cluster Assignment):**
    - For each data point *Xi* in our dataset, calculate the Euclidean distance (we can use other distance metrics) between *Xi* and each centroid *Cj*:

$$D(X_i, C_j) = \sqrt{\sum_{k=1}^{n} (X_i[k] - C_j[k])^2}$$

    - Assign the data point *Xi* to the cluster *Cj* that has the minimum distance.

3. **Update Step (Centroid Recalculation):**
    - Calculate the new centroids for each cluster. These are the mean values of all data points in each cluster:

$$C_j[k] = \frac{1}{\text{Number of data points in Cluster}_j} \sum_{X_i \in \text{Cluster}_j} X_i[k]$$

    - This step repositions the cluster centers based on the data points that belong to each cluster.
4. **Repeat Assignment and Update Steps:**
    - Repeat the assignment and update steps iteratively until a stopping condition is met, such as a maximum number of iterations, convergence, or another criterion.
5. **Termination:**
    - Once the algorithm converges or reaches the maximum number of iterations, it terminates. At this point, we have K clusters, and the data points are grouped based on their similarity to the centroids.
6. **Results:**
    - The final clusters are represented as C1,C2….Ck and each cluster contains data points that are closest to its centroid.

Elbow Method

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

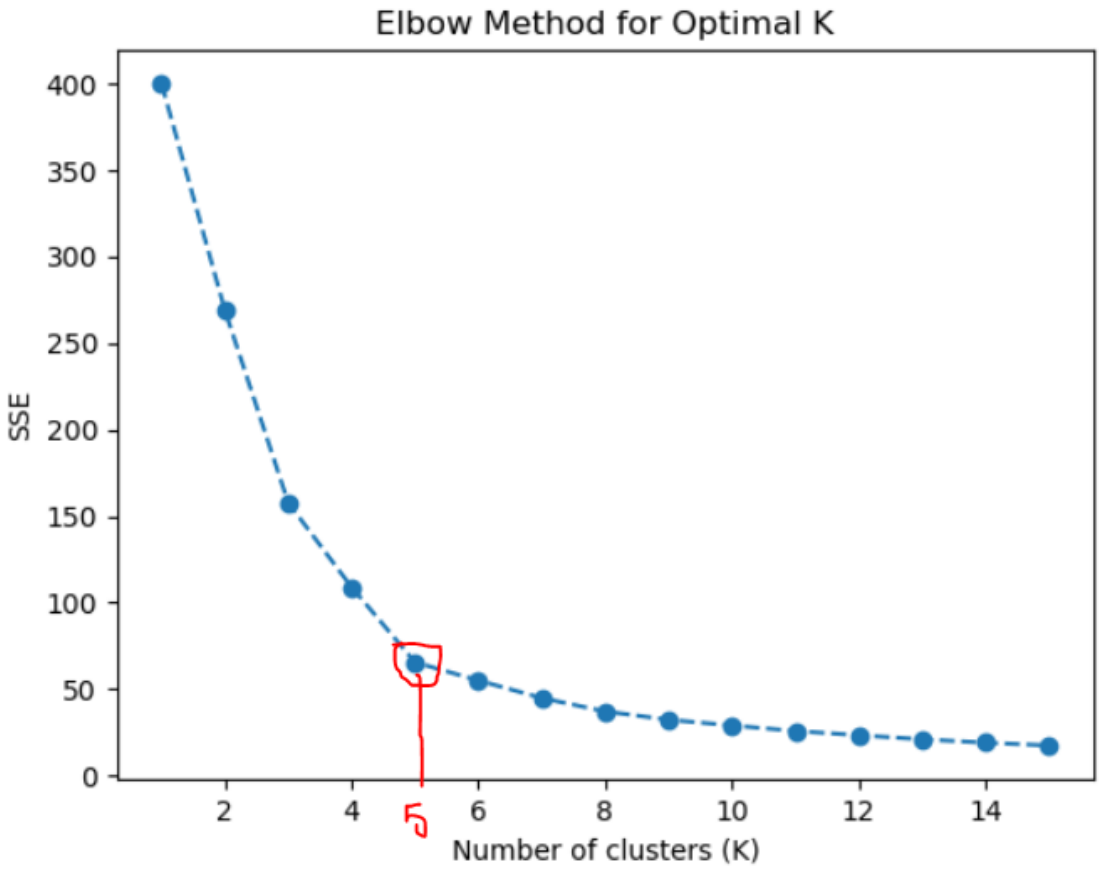
1.Range of K Values: We considered a range of K values from 0 to 15.

2.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. 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.

Result Explanation –

1.We present the Elbow Method plot, depicting K values on the x-axis and SSE values on the y-axis.

2.The plot displays an "elbow point" where the SSE starts to level off, indicating the optimal K value. We can consider optimal value of k as 5 .



PART- (B)

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

Answer :

Hierarchical clustering builds a tree of clusters. It either starts with individual data points as clusters and merges them iteratively, or it starts with all data points as one cluster and splits them iteratively. The result is a dendrogram, which shows the arrangement of the clusters.

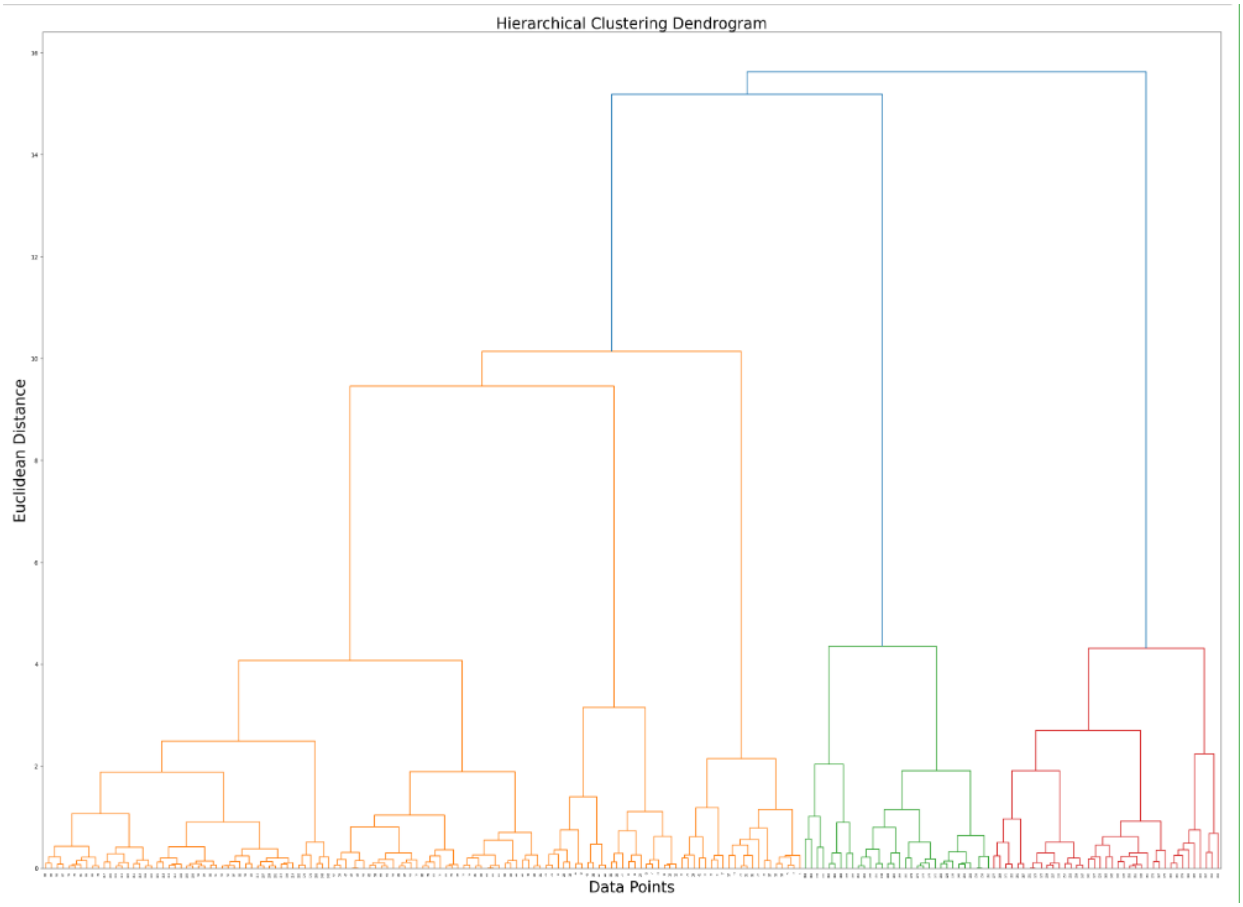
Showing the theoretical and mathematical steps:

Input: A set of data points X and a distance (dissimilarity) matrix D that represents the pairwise distances between the data points.

Output: A hierarchy of clusters, often represented as a dendrogram.

1. **Initialization**:
    - Start with each data point as an individual cluster. At this stage, there are as many clusters as there are data points.
    - Let Ci represent the i-th cluster, and n be the total number of data points.
2. **Calculate Distance Matrix**:
    - Compute the distance between each pair of clusters Ci and Cj using a linkage method.Let $D(Ci,Cj)$ represent the distance between clusters Ci and Cj.
3. **Merge Clusters**:
    - Identify the two clusters Ca and Cb with the smallest distance $D(Ca,Cb)$.
    - Merge these two clusters into a new cluster Cab.
4. **Update Distance Matrix**:
    - Recalculate the distances between the newly formed cluster Cab and all other clusters, using the chosen linkage method.

5. **Repeat Steps 3 and 4**:
    - Continue merging the two closest clusters and updating the distance matrix until only one cluster remains. This process forms a hierarchical structure of clusters.
6. **Dendrogram Visualization** :
    - The result is often visualized as a dendrogram, which is a tree-like structure representing the hierarchy of clusters. The height at which you cut the dendrogram determines the number of clusters you obtain. Clusters are created by cutting the dendrogram at a specific height.



PART-(C)

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

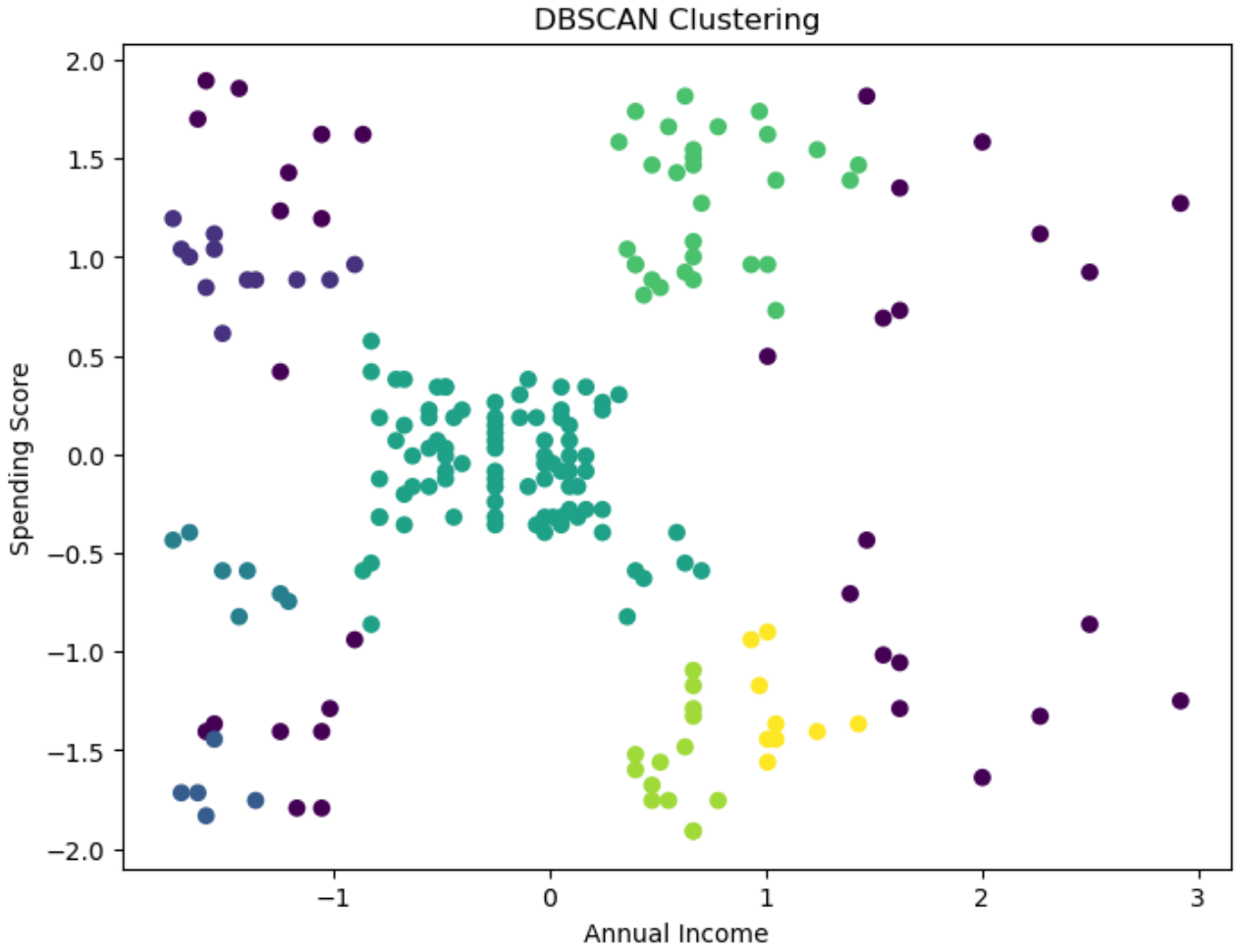
Answer :

Density-Based Spatial Clustering of Applications with Noise: DBSCAN groups together points that are close to each other and marks points as outliers that are in low-density regions.

Visual Result : Our analysis yielded the following results:

1.We applied the DBSCAN algorithm to the dataset to identify clusters and classify data points as core points, border points, or noise points.

2. We present a scatter plot of the DBSCAN clusters, color-coding data points by them.



Reference Links :

1. https://www.analyticsvidhya.com/blog/2019/08/comprehensive-guide-k-means-clustering/
2. https://www.youtube.com/watch?v=oNYtYm0tFso
3. https://www.youtube.com/watch?v=-p354tQsKrs
4. https://www.youtube.com/watch?v=KzJORp8bgqs&t=475s

--END--