REPORT

PRINCY GAUTAM(B20BB051)

Task 1: K-Means

Method

Import necessary libraries for calculations, preprocessing and plotting the graph.

Download and read the dataset using pandas.

Drop the categorical column from the dataset as K-Means algorithm could not be performed on categorical data.

Convert the leftover data to NumPy array.

Scale the dataset using minmax scaler and obtain the information in a data frame and further convert it into NumPy array.

Import K-Means from the scikit library. Set the number of desired clusters. Initialize the parameters and assign the variable k-means to it.

Predict the cluster for all the samples.

Plot a scatter plot to visualize different clusters. Use different shapes and color to differentiate between any two clusters. Plot the cluster centers and identify them with black circle.

Perform k-means clustering for three different values of number of clusters.

Print the cluster centroids.

Use elbow method to obtain the optimal value of K. perform k-means on a range of number of clusters and append the inertia of k-means in an empty list.

Obtain a scatter plot between any two features (here, protein and ash).

Results & observations

- We see a scatter plot of five different clusters with different shapes and color.
- Prediction array is different for every value of k.
- An array of cluster centers is printed for five clusters.
- The optimal value of K is 2, as seen from the elbow point.

Task 2: Hierarchical clustering and k-Means [w/o inbuilt function]

Method (I)

Import necessary libraries for calculations, preprocessing and plotting the graph.

Download the dataset.

Read the dataset using pandas library and save the dataset in the variable 'df'.

Drop the categorical column from the dataset. Convert the dataset into NumPy array and save it in a variable 'D'.

Import dendrogram and linkage from scipy library. Plot the dendrogram on the dataset.

Perform agglomerative clustering with different values of k and linkage and thereafter plot the scatter plots for them.

Method (II)

Implement k- means clustering from scratch. The steps are given below:

Create a class and initialize with the parameters: dataset and number of clusters.

Randomly initialize the cluster centers of each cluster from the data points. Let's assume k=2 and so we chose randomly 2 data points as assume them as centroids.

For each data point, compute the euclidean distance from all the centroids and assign the cluster based on the minimum distance to all the centroids.

Adjust the centroid of each cluster by taking the average of all the data points belonging to that cluster based on calculations performed in above step. This helps us to predict the clusters.

Repeat the above steps until clusters are well separated or we can say until convergence is achieved.

Obtain a scatter plot of the clusters to visualize.

Results & observations

- Dendrogram is used to study the hierarchical clusters before obtaining the appropriate number of clusters to the dataset. The distance at which two cluster combine is referred to as the dendrogram distance. This distance is a measure of if two or more clusters are disjoint or can be combined to form one cluster together.
- Our dendrogram illustrates the presence of 2 clusters. This is verified with the elbow method where we obtained the optimal value of k = 2.

- With agglomerative clustering, we can visualize the clusters. Changing the number of clusters value, we can see how clusters are combining together.
- At last, we can see the clustered data where we obtained two clusters.

Task 1 colab file link:
https://colab.research.google.com/drive/12ENWHX_9yj9IiNr0vv5ZXubPerVPgw6H#scrollTo=yCPCWI0uFQwS

Task 2 colab file link:
https://colab.research.google.com/drive/1mImmS6ErhVmqGkTX8oil9tXnNHh321zA#scrollTo=6ywjGbH973Oy