National University of Computer and Emerging Sciences

Lab Manual 11
CL461-Data Mining Lab

Course Instructor	Eesha Tur Razia Babar
Lab Instructor (s)	Abdul Rehman Mateen Fatima
Section	BDS-6A
Semester	Spring 2024

Lab Task:

Types of Clustering:

[1] Exclusive (partitioning)

In this clustering method, Data are grouped in such a way that one data can belong to one cluster only.

Example: K-means

[2] Agglomerative

In this clustering technique, every data is a cluster. The iterative unions between the two nearest clusters reduce the number of clusters.

Example: Hierarchical clustering

KMeans Clustering

K means it is an iterative clustering algorithm which helps you to find the highest value for every iteration. Initially, the desired number of clusters are selected. In this clustering method, you need to cluster the data points into k groups. A larger k means smaller groups with more granularity in the same way. A lower k means larger groups with less granularity.

The output of the algorithm is a group of "labels." It assigns data point to one of the k groups. In k-means clustering, each group is defined by creating a centroid for each group. The centroids are like the heart of the cluster, which captures the points closest to them and adds them to the cluster.

Code Example:

from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=4, max_iter=50)

kmeans.fit(dataframe)

For plotting:

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(10, 7))
plt.scatter(df['var1'], df['var2'], c=cluster.labels_)
```

Hierarchical Clustering:

Hierarchical clustering is an algorithm which builds a hierarchy of clusters. It begins with all the data which is assigned to a cluster of their own. Here, two close cluster are going to be in the same cluster. This algorithm ends when there is only one cluster left.

Code Sample:

```
from sklearn.cluster
import AgglomerativeClustering
cluster = AgglomerativeClustering(n_clusters=2, affinity='euclidean', linkage='ward')
cluster.fit_predict(data_scaled)
```

For plotting:

```
import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(10, 7))
plt.scatter(df['var1'], df['var2'], c=cluster.labels_)
```

Lab Work:

Problem 1: (Kmeans)

Online Retail Dataset is attached in the file. You need to load the dataset and preprocess it for missing values and outliers. You need to scale the data for better clusters. You are then required to perform Kmeans clustering on this dataset and try with different number of clusters and visualize it.

Problem 2: (Agglomerative Clustering)

Online Retail Dataset is attached in the file. You need to load the dataset and preprocess it for missing values and outliers. You need to scale the data for better clusters. You are then required to perform Hierarchical clustering on this dataset and try with different number of clusters and visualize it.

Problem 3: (Comparison)

Explain which algorithm is better, the one used in Problem 1 or Problem 2, and why? You claim must be supported by stats and facts showcased in your code.

Instructions:

1. Implement the above-mentioned algorithm
2. Write a detailed analysis