Assignment 3

A1(a) Density based clustering:
https://www.kaggle.com/datasets/vjchoudhary7/customer-segmentation-tutorial-in-python
This dataset contains data of a supermarket mall and through membership cards , there is some basic data about its customers like Customer ID, age, gender, annual income and spending score.
We want to understand the customers like who can easily converge [Target Customers] so that the sense can be given to marketing team and plan the strategy accordingly.
Theoretical reasoning:
The dataset contains attributes like annual income and spending score of the customers which are relevant to us for targeting customers. Since different customers will have different values of spending scores and incomes, we need to consider different possible combinations of these values. If we plot these points on a graph in the form of dots, we can cay that the dots that closer together having high density will have a common pattern for spending and earning money, and thus these points can be put into the same cluster.
And there can some customers who earn too much and may have very different patterns. These points will behave as outliers.
The clusters, in this case, can be irregularly shaped due to different customer behaviours.
Reasoning:
Plotted Kernel Density Estimation plots for all individual features. Since there is a concentrated peak in the KDE plots and a normal distribution shaped curve, the data is likely to have areas of high density corresponding to those points.

The clusters in the tsne plot are also irregular shaped so density based clustering would be suitable.

A1(b)
Agglomerative hierarchical clustering:
https://www.kaggle.com/datasets/himanshunakrani/iris-dataset
It includes three iris species with 50 samples each as well as some properties of each flower. One flower species is linearly separable from the other two, but the other two are not linearly separable from each other.

Theoretical Reasoning:
This dataset contains properties of three different species of flowers. And it has been given that one flower is different from the other two, but the other two are not very different. This means that there is a hierarchy in difference between the flowers, and this is exactly what is suitable for agglomerative hierarchical clustering.
Reasoning:
From the scatter plot in the code file, we can see that the wine dataset fits the best on prototype based clustering algorithm. And this can also be seen in the purity values.

A1(c)

Prototype-based clustering
https://www.kaggle.com/datasets/harrywang/wine-dataset-for-clustering
This dataset contains the results of a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.

Theoretical reasoning:
Since this dataset contains the quantities of each of the components in the three types of wines. Since the wines are produced in different cultivars, they are going to have some differences in these components. The wines produced in the same cultivars will have similar quantities of these components. Hence the cultivars can be a big factor for the difference in the wines and can be the reason for the wines to be in different clusters.
Reasoning:
From the scatter plot in the code file, we can see that the wine dataset fits the best on prototype based clustering algorithm. And this can also be seen in the purity values.

2. (b) Advantages:

1. It is very efficient when we want to classify high-density clusters from low-density clusters in a dataset.
2. It can handle the outlier points efficiently.

Disadvantages:

1. This algorithm does not work nicely if the dataset points have a uniform or near-uniform density in the dataset.
2. Since it draws a shape of n-dimensions and includes every point that lies in that shape, the complexity of the algorithm increases with the increasing value of n. Thus, for high dimensional data, this algorithm gives bad performance.

(c) HDBSCAN is one such algorithm that makes clusters based on differences in density.

The main idea of HDBSCAN is to build a hierarchy of cluster that are formed on the basis of density of connected components. It identifies cluster based on density variations.

Advantages of HDBSCAN **over DBSCAN:**

1. HDBSCAN can identify clusters of varying densities in the same dataset as well. This overcomes a limitation of DBSCAN as DBSCAN assumes that clusters roughly have a uniform density.
2. With HDBSCAN we can find clusters of non-convex shapes as well which is not the case with DBSCAN.
3. In the HDBSCAN we do not need to specify the eps parameter which can be quite challenging. We just need to specify min number of points in cluster. Clusters are identified on the basis of density variations.

2. (d)

Best results on datasets:
Mall: Silhouette Coefficient is 0.3835838178228424
Iris: Silhouette Coefficient is 0.43467143265757235
Wine: Silhouette Coefficient is 0.03957050764295184

2. (e) Parameters changed: Min number of points in cluster, and cluster_selection_epsilon

2. (f) The hypothesis is confirmed that the 1st dataset is best for density-based clustering. Since we do not have the true labels in this case, we visually analyzed the result on the mall data set across the algorithms and found out that the best clustering was done by the density-based algorithm since it was able to identify different clusters across.

2. (g)

In order to verify the results of the clustering, bootstrapping was done on each dataset. After bootstrapping, clustering was performed on each dataset and the results were compared to the original dataset.

For iris: 0.47336839452966634
For wine: Silhouette Coefficient is 0.04121106385558935
For Mall: Silhouette Coefficient is 0.3478214497496882

Since no significant deviation was found, we can conclude that there is an underlying distribution and the results obtained are not random.

3. (a)

The two types of hierarchical clustering algorithms are:

- Agglomerative algorithm
- Divisive algorithm

The agglomerative algorithm is used more than the divisive algorithm because:

- It is easier to understand as it shows a dendrogram that displays the merging and hierarchy of classes.
- It enables us to choose the number of clusters, too, according to the level of granularity we want our classification to go to.
- This also provides it the ability to handle outliers as it can include those points in bigger clusters

3. (c)

Here are the values of the Silhouette Coefficient for all the datasets for all the algorithms.

For wine dataset:

Single : 0.18584684069789328
Complete: 0.20028718167300175
Average: 0.15802505775764714
Ward: 0.27631805856826086

For iris dataset:

Single: 0.4784399371880524
Complete: 0.5120749374629816
Average: 0.44330017257110443
Ward: 0.45183200660097467

For Mall customers dataset:

Single: 0.365677402459876
Complete: 0.4218340266628975
Average: 0.42761358158775435
Ward: 0.4618340266628975

3. (d)

Here, we can see that the Silhouette Coefficient has the highest value in the iris dataset. In the iris dataset, all the methods give a high value for the Silhouette Coefficient.
In contrast, the wine dataset gives a very low value of the Silhouette Coefficient, and the mall customer dataset also gives comparatively lower values.
Here, this dataset is best suited to be run on hierarchical clustering algorithms.

(e)
In order to verify the results of the clustering, bootstrapping was done on each dataset. After bootstrapping, clustering was performed on each dataset and the results were compared to the original dataset.

For iris: Silhouette Coefficient is 0.47457345781895754
For wine: Silhouette Coefficient is 0.2694070979503702
For Mall: Silhouette Coefficient is 0.4494857706000199

Since no significant deviation was found, we can conclude that there is an underlying distribution and the results obtained are not random.

4.(a)
(i) K-means clustering is sensitive to outlier points. This is because K-means clustering works by defining points called centroids, and then it takes data points in their vicinity and puts them in the same cluster. Now, these centroids and points are selected using the “mean” metric. And since the mean is a metric that can easily get affected by outliers, the centroids, and other chosen points get disturbed from what they should have been and thus give bad results.

(ii) When the number of dimensions is high, it becomes hard for the K-means algorithm to work efficiently. This is because when there are too many dimensions, the distance between the data points becomes small, and the points come closer together. This is not good for an algorithm that uses distances as a metric to calculate means and then classify them according to those distances.

(b) Fuzzy c-means is one such algorithm. It can handle outliers better than Kmeans. This is because it uses soft clustering whereas Kmeans uses hard clustering. Soft clustering means that a point can belong to more than 1 cluster. It

assigns weight with which each point can belong to a cluster. The sum of weights for a point across all clusters is 1. So in the case of outliers, the outlier would not clearly belong to one cluster so its weight would be spread out across all the clusters. Since the weight of an outlier in each cluster would be low so the cluster centroid would not be influenced that greatly as compared to if it the point belonged to only 1 cluster.

Mixture Model Clustering is one algorithm that can handle scaling with a number of dimensions better as compared to Kmeans. In this, we use a probabilistic framework and model the clusters as Gaussian distributions. This allows for both spherical and elliptical shapes whereas Kmeans only allows for spherical shapes. With increasing number of dimensions the shape of clusters may not remain spherical only so mixture model clustering would handle it better. Also Mixture model clustering can capture the data distribution better using Gaussian distribution and EM algorithm. Also this algorithm can handle sparsity which often occurs with high dimensional data better than Kmeans since it uses a gaussian distribution.

4(c)
Here are the values of the Silhouette Coefficient for the three algorithms on the wine dataset.
K-means: 0.2835806364948403
GaussianMixture/ Mixture Model: 0.2829669172404775
Fuzzy cmeans: 0.2835806364948403

Here are the values of the Silhouette Coefficient for the three algorithms on the iris dataset.
K-means: 0.452949780355554
GaussianMixture/ Mixture Model: 0.44249183226269917
Fuzzy cmeans: 0.45231919470216203

Here are the values of the Silhouette Coefficient for the three algorithms on the mall customer dataset.
K-means: 0.46761358158775435
GaussianMixture/ Mixture Model: 0.4310103691306562
Fuzzy cmeans: 0.46761358158775435

Both algorithms give the same performance as kmeans on our dataset. This may be because our dataset performs too well on means because of only a few outlier points.

4. (d) Hyperparameters changed: Num clusters, covariance_matrix_type.

4(e)
Here are the values of the Silhouette Coefficient for all the datasets for all the algorithms.
For wine dataset:
GaussianMixture/ Mixture Model: 0.5029669172404775
Fuzzy cmeans: 0.5035806364948403

For iris dataset:
GaussianMixture/ Mixture Model: 0.44249183226269917
Fuzzy cmeans: 0.45231919470216203

For Mall customers dataset:
GaussianMixture/ Mixture Model: 0.4310103691306562
Fuzzy cmeans: 0.46761358158775435

Here, we can see that the Silhouette Coefficient has the highest value in the wine dataset. In the wine dataset, all the methods give a high value for the Silhouette Coefficient.
In contrast, the mall customer and iris dataset gives a lower value of the Silhouette Coefficient.
Here, this dataset is best suited to be run on prototype clustering algorithms.

4. (f)
In order to verify the results of the clustering, bootstrapping was done on each dataset. After bootstrapping, clustering was performed on each dataset and the results were compared to the original dataset.
Results on the bootstrapped dataset:

Wine: Silhouette Coefficient is 0.2664546615310252
Iris: Silhouette Coefficient is 0.49162911834836187
Mall: Silhouette Coefficient is 0.4191137821992875

Since no significant deviation was found, there is an underlying distribution and the results obtained are not random.