IC 272: DATA SCIENCE - III
LAB ASSIGNMENT – VII
Clustering

Student's Name: Yash Sharma **Mobile No: 8802131138**

Roll Number: B20241 **Branch: CSE**

1

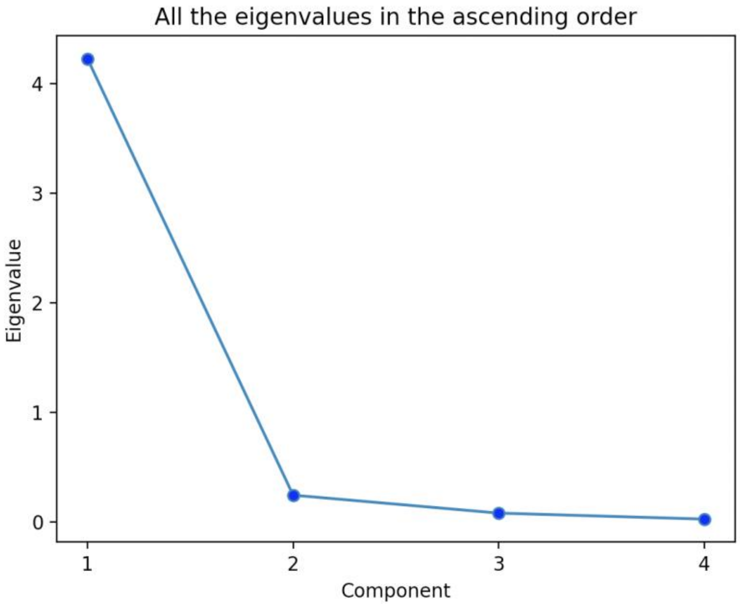

Figure 1 Eigenvalue vs. components

Inferences:

1. The eigenvalue continuously decreases corresponding to each component increase successively.
2. A very sharp decreases is visible from first eigenvalues to second eigenvalue then the decrease becomes quite negligible. This proves that the first director (eigenvector) captures most of the data variance.
3. This trend is consistent with the concepts of Principal Component Analysis which states that the data projected along the eigen-vector corresponding to biggest eigen-value has the most information about the original data.
4. The given data has dimension of 4 (excluding the column 'class') and thus a total of 4 eigenvector would be possible. And the eigenvalues corresponding to these directions would have different values and when sorted in descending order would mostly give a decreasing trend. This happens because in some direction represents or captures the variance/ data variations in the best possible

way. And further on as eigenvalue decreases which states that particular eigenvector direction captures less variance of data as compared to its predecessor.

2 a.

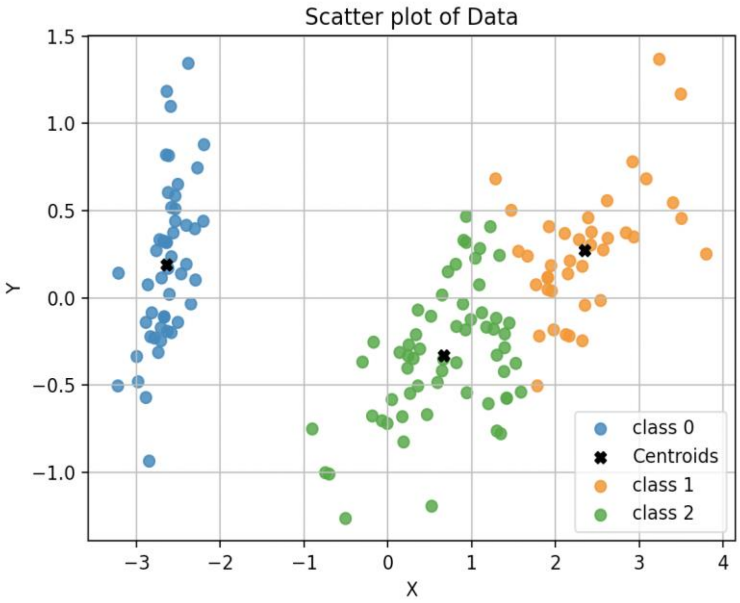

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The concept behind the K-means clustering algorithm is very simple to implement, yet very effective. The meat of the K-means clustering algorithm is just two steps, the cluster assignment step and the move centroid step. It's an unsupervised learning algorithm that is easy to implement and can handle large data sets, K-means clustering is a good starting point. From above plots its quite visible that the clusters have been very effectively distinguished and a judicial clustering has been observed using this algorithm. This is a hard clustering technique.
2. We see from scatter plot that the data is not quite circular in shape and is more elongate or elliptical in shape. And as the k-means model places a circle (or, in higher dimensions, a hyper-sphere) at the center of each cluster, with a radius defined by the most distant point in the cluster. This radius acts as a hard cutoff for cluster assignment within the training set: any point outside this circle is not considered a member of the cluster. That's why k means fails to make an accurate / flexible boundary shape when data is randomly shaped.
3. No, the boundary does not seem to circular at all. The boundary is quite linear. The boundary is linear because we are using Euclidean distance and the Euclidean distance is calculated for the data points

using 1st order statistics that is mean and variance thus giving us linear boundary. There are no second order terms as such.

4. These two disadvantages of k-means are —its lack of flexibility in cluster shape and lack of probabilistic cluster assignment—mean that for many datasets (especially low-dimensional datasets) it may not perform as well as we might hope.

b. The value for distortion measure is 63.874

c. The purity score after examples are assigned to the clusters is 88.667 %

3

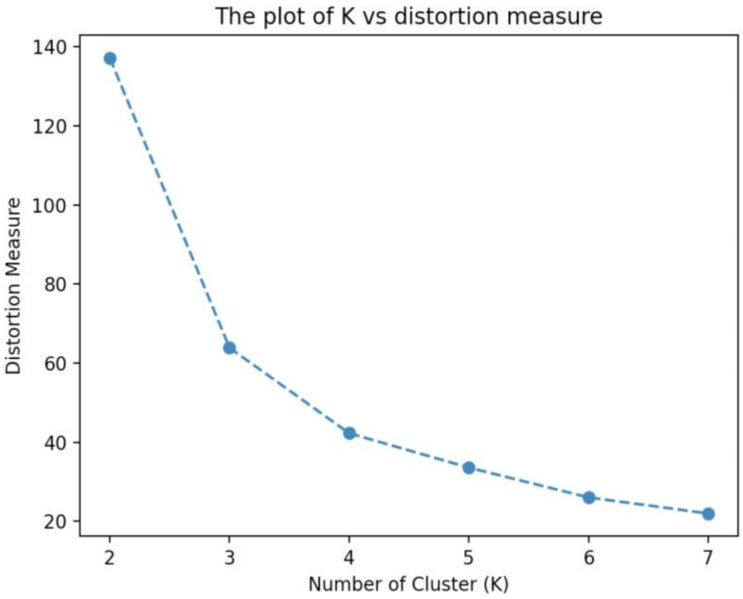

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The Distortion measure decreases with increase in K (Number of clusters).
2. The above trend is observed because as the number of clusters increases then so does the cluster centers and thus now points are allotted to cluster center more near to them than the previous one (in most of the cases) thereby decreasing the sum of squared distances of points from cluster centers and thus distortion measure decreases gradually. The distortion measures become zero when number of clusters equals the number of points as now every point is itself center thus distance becomes zero.

3. With each iteration, the data is more correctly clustered based on its distance from the nearest cluster center. As we repeat the process over and over, the data will become completely clustered at one point.
4. From our intuition the optimum number of clusters shall be 3 as there are 3 different species in our Iris flower dataset. From elbow method, the optimum number of clusters are also coming out to be 3. Thus, the elbow and distortion measure plot follow our intuition quite well.

Table 1 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	0.667
3	0.887
4	0.687
5	0.680
6	0.527
7	0.507

Inferences:

1. The highest purity score is obtained with K =3.
2. On Increasing the value of k shows two types of trends here. Initially a sharp rise in Purity score is seen when k moves from 2 to 3. There onwards a gradual decrease in Purity score magnitude when number of clusters are increased further beyond 3.
3. Initially as from our intuition there are 3 species thus 3 clusters must be there. Therefore, a sharp rise was seen when k changed from 2 to 3. There onwards, as there are only 3 species or 3 cluster for Iris flower in our dataset thus when clusters are increased beyond 3 then some addition random classes are introduced in model which classifies the data for the class which is not at all present in our dataset. Thus, making totally different predictions and so the classes falling into the 3 given original cluster/species reduces thereby decreasing the net purity score gradually when K is greater than 3.
4. It may be observed that for K greater than 3 both Purity score and distortion measure are decreasing and somewhere saturated.

4 a.

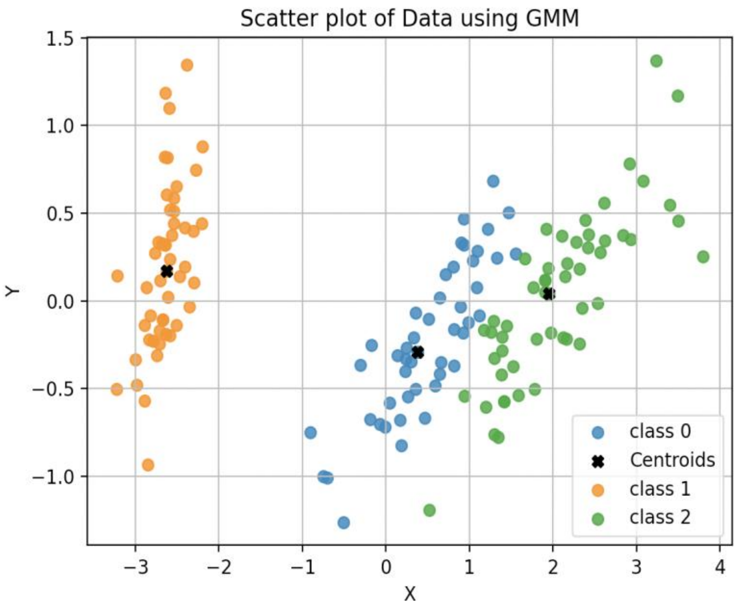

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. As GMM is a soft clustering algorithm and assigns points to cluster using its likelihood probability thus giving more flexible boundaries rather than linear as seen in k-means. A very clustering a visible from the scatter plot also.
2. The boundary seems to be quite elliptical instead of being circular.
3. The GMM approach easily fits any stretched dataset, if allowed for a full covariance, the model will fit even very oblong, stretched-out clusters. That’s the power of GMM model and flexible cluster boundaries. That’s why it’s being referred to as soft clustering algorithm. That’s the main difference between GMM and the K mean s algorithm. The k-means algorithm fails to cluster out elliptical or stretched out dataset and generally the boundaries are quite ‘linear’ while on the other hand GMM algorithm provides us much flexibility towards boundary shape. It can easily cluster very complex cluster shapes while k-means will be unable to do so.

b. The value for distortion measure is -280.869

c. The purity score after examples are assigned to the clusters is 98.0 %

5

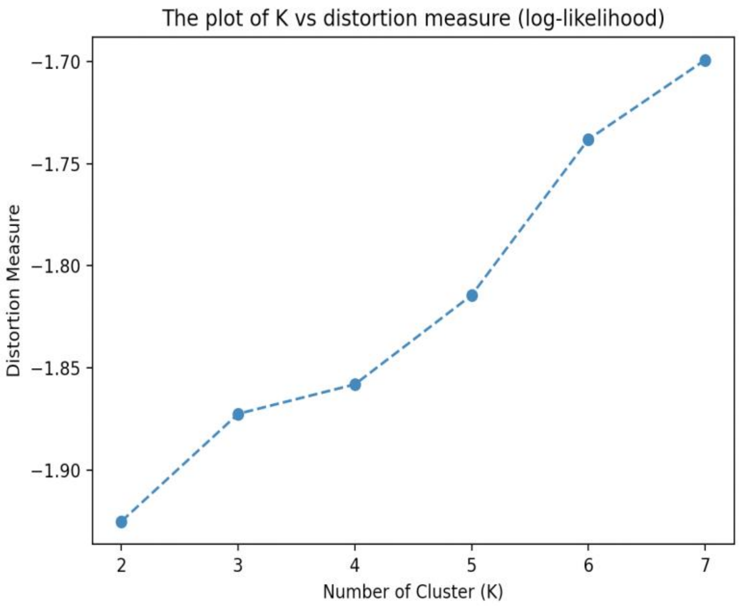

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increases with increase in K.
2. As the number of clusters are increases using GMM algorithm and since it's a soft clustering algorithm thus the probability of each point belonging to its nearest cluster increase. And as K increases the cluster comes near to any point under observation. This means that probability that that point belong to that cluster increases. And the probability becomes 1 when K equals number of points in the dataset. With increase in probability the total maximum log-likelihood increases and thus this type of trend is observed.
3. From our intuition the optimum number of clusters shall be 3 as there are 3 different species in our Iris flower dataset. From elbow method, the optimum number of clusters are also coming out to be 3. Thus, the elbow and distortion measure plot follow our intuition quite well.

Table 2 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	0.667
3	0.98
4	0.833
5	0.78
6	0.700

7	0.620

Inferences:

1. The highest purity score is obtained with K =3.
2. On Increasing the value of k shows two types of trends here. Initially a sharp rise in Purity score is seen when k moves from 2 to 3. There onwards a gradual decrease in Purity score magnitude when number of clusters are increased further beyond 3.
3. Initially as from our intuition there are 3 species thus 3 clusters must be there. Therefore, a sharp rise was seen when k changed from 2 to 3. There onwards, as there are only 3 species or 3 cluster for Iris flower in our dataset thus when clusters are increased beyond 3 then some addition random classes are introduced in model which classifies the data for the class which is not at all present in our dataset. Thus, making totally different predictions and so the classes falling into the 3 given original cluster/species reduces thereby decreasing the net purity score gradually when k is greater than 3.
4. It may be observed that for K greater than 3 both Purity score are decreasing, and distortion measure is increasing (though its valid for K<3 also). One thing is worth noting that the value of K from elbow method also shows highest Purity Score among all other value and matches our intuition too. This shows how powerful GMM can get.
5. GMM seems to be producing somewhat s better results at K=3. Apart from K=3 as well, GMM clusters the data with a bit of more purity score. Moreover, at K=3, it has very high Purity score as compared to k-means. It has also flexible boundaries of its clusters.

6

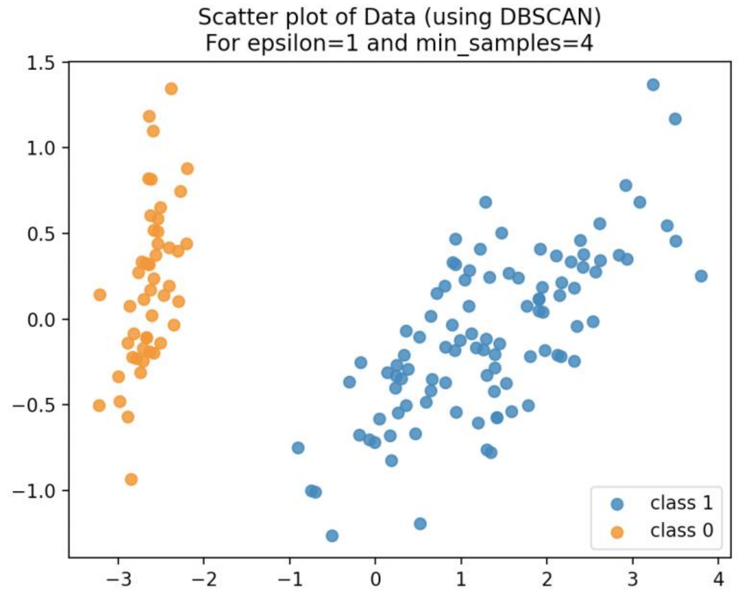

Figure 6 DBSCAN clustering for epsilon=1 and min_samples=4 on Iris flower dataset

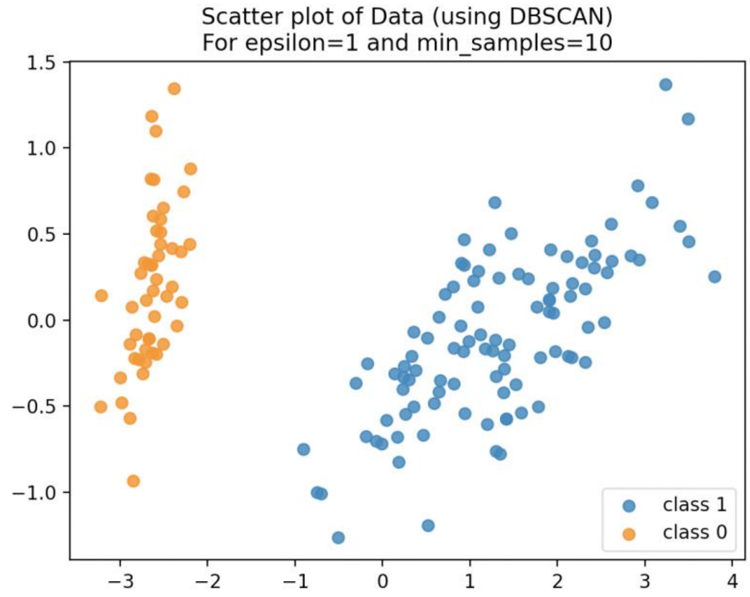

Figure 7 DBSCAN clustering for epsilon=1 and min_samples=10 on Iris flower dataset

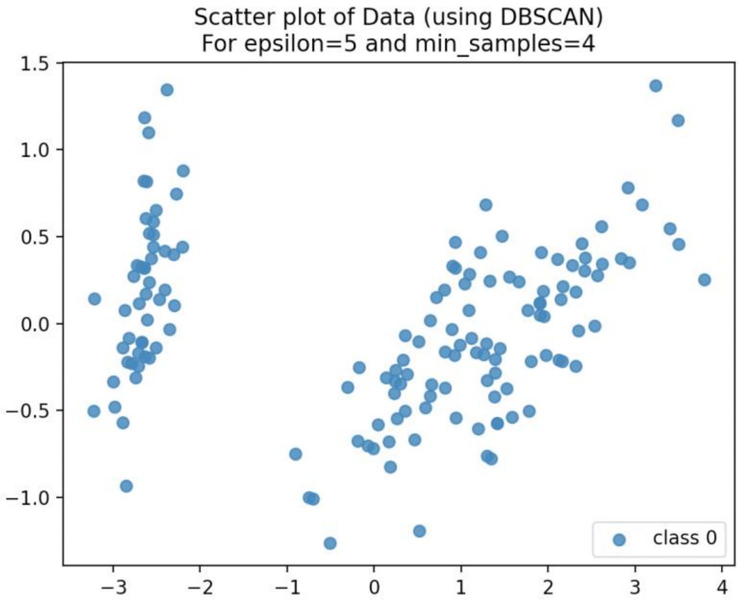

Figure 8 DBSCAN clustering for epsilon=5 and min_samples= 4 on Iris flower dataset

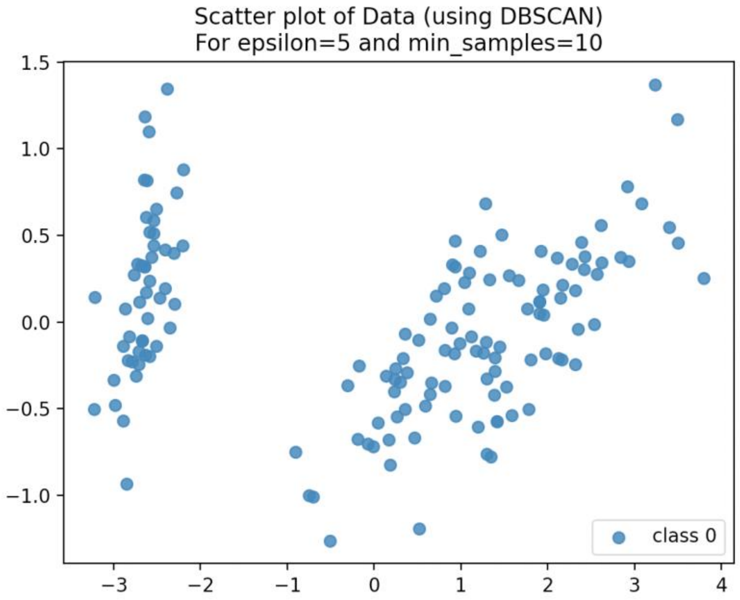

Figure 9 DBSCAN clustering for epsilon=5 and min_samples=10 on Iris flower dataset

Inferences:

1. We may observe that the cluster formed are not proper or more feasible. The number of clusters are even less than our intuition which is K=3 as 3 species are present in our dataset.
2. In case of GMM and K-means algorithms they were able to cluster the available data into 2 or more cluster. While in this scenario only two clusters are made in initial stages when epsilon is 1 and only 1 cluster is made when epsilon is 5. This way the algorithm is less efficient and feasible at the same time as compared to the previous algorithms.
3. The possible reason for such observation can be the incorrect values chosen for epsilon and *min_points*. May be lower values of epsilon than 1 may produce satisfying results.

b.

Eps	Min_samples	Purity Score
1	4	0.667
	10	0.667
5	4	0.333
	10	0.333

Inferences:

1. No, for same *epsilon* value, on increasing *min_samples*, the purity score remains exactly same. No change is observed in its value.
2. For same *min_samples*, increasing *epsilon* value is decreasing the Purity score.
3. We observe that no specific. Trend can be observed from available results and thus non conclusive arguments can be made in this respect.
4. It may be possible that better results which makes our more feasible can be obtained if the values of *epsilon* and *min_samples* is chosen more carefully.