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

Student's Name: Madhur Jajoo **Mobile No: 7597389137**

Roll Number: B20211 **Branch:DSE**

1

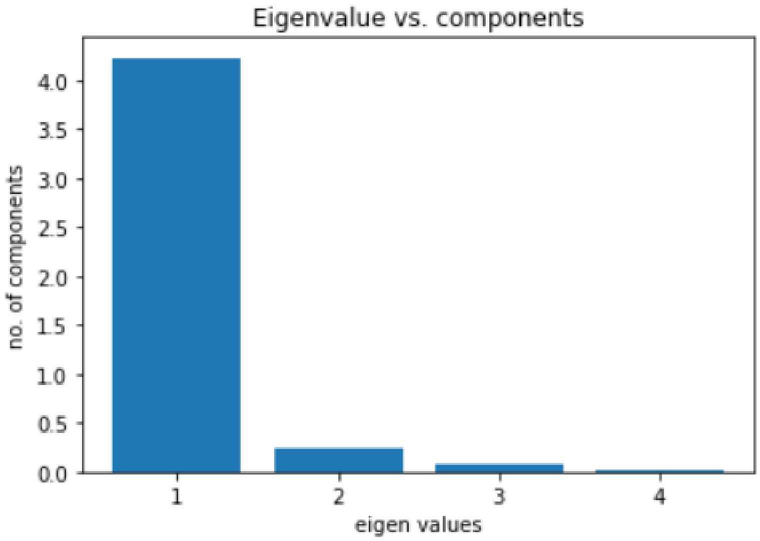

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigen value decreases with increase in component value.
2. Eigen value represent variance of component, some of the components have covered more covariance as compared to others.

2 a.

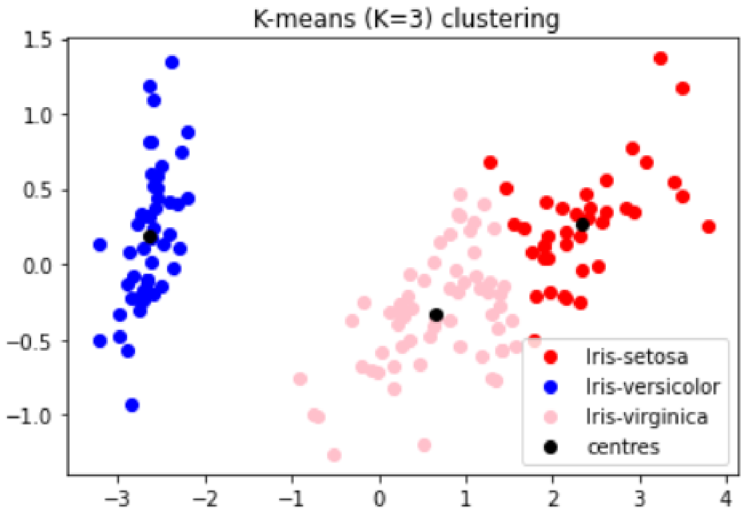

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering prowess of the algorithm is good.
2. The boundaries of the clusters seems to be in straight line, they are not circular.

b. The value for distortion measure is 63.874.

c. The purity score after examples are assigned to the clusters is 0.887.

3

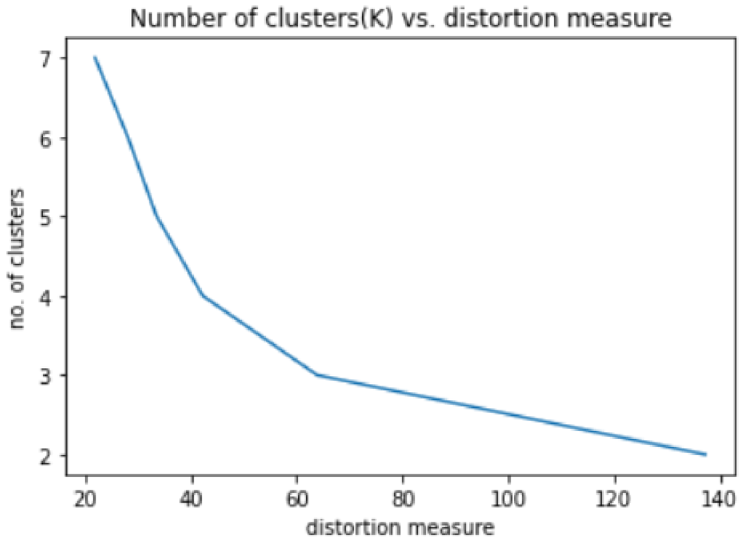

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with increase in value of K.
2. Here, the distortion measure is the sum of squared distance of the points from center, as we increase the no. of clusters the data gets more close to respective centers and the distance decreases which implies the squared distance decreases resulting in the decreased distortion measure.
3. Intuitively there should be 3 clusters only in the dataset. No, because Kmeans is unsupervised clustering i.e. it didn't use labels of the data points and only 2 clusters can be seen therefore it gives 2 as optimum no. of clusters.

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	0.673
6	0.527
7	0.513

Inferences:

1. The highest purity score is obtained with K =3.
2. From 2 to 3 the purity score increases but after that it decreases.

3. As we have only 3 labels in our dataset, more no. of clusters will lead some points in a cluster which doesn't even exist in the original data which leads to decrease in the purity score.
4. After the maximum value of the purity score it decreases with increase in no. of clusters.

4 a.

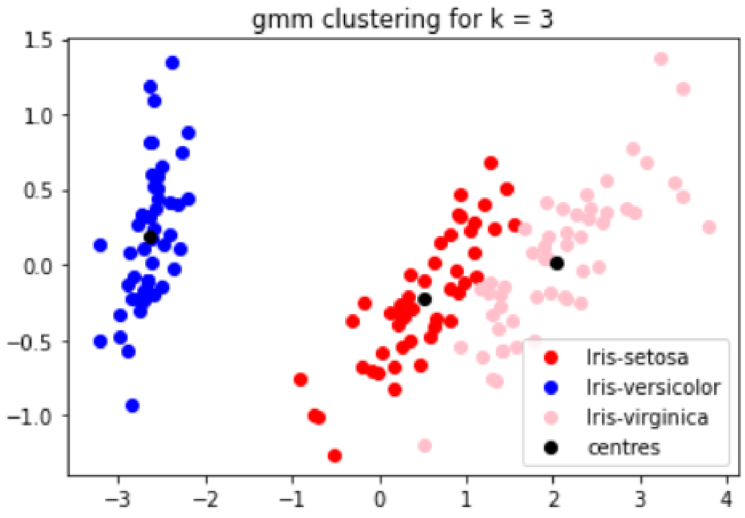

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering prowess of the algorithm is good.
2. From the output, the boundary doesn't seem to be circular.
3. There is no visual difference between the cluster plots of the k means and the gmm algorithms.

b. The value for distortion measure is -280.960

c. The purity score after examples are assigned to the clusters is 0.98.

5

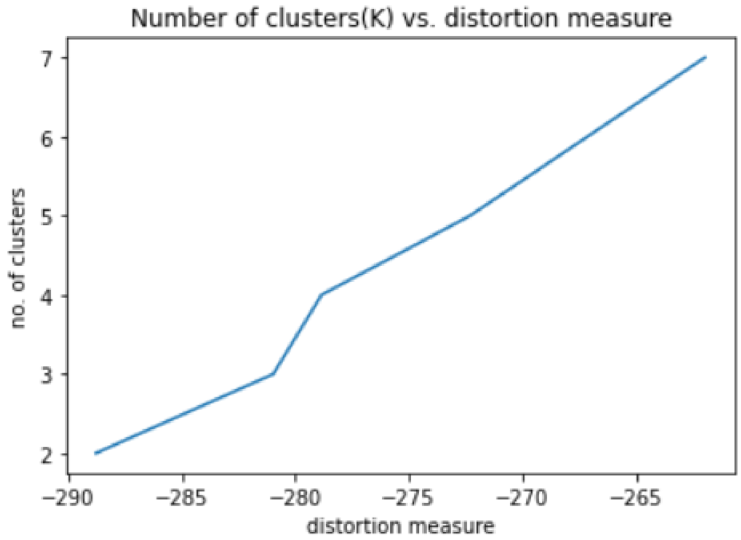

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure increases with increase in no. of clusters.
2. Intuitively there should be 3 clusters only in the dataset. No, because gmm is unsupervised clustering i.e. it didn't use labels of the data points.

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

K value	Purity score
2	0.667
3	0.98
4	0.82
5	0.773
6	0.747
7	0.68

Inferences:

1. The highest purity score is obtained with K =3.
2. The purity score first increases from K=2 to 3 but after that it decreases with increase in value of K.
3. As we have only 3 labels in our dataset, more no. of clusters will lead some points in a cluster which doesn't even exist in the original data which leads to decrease in the purity score.
4. Once the purity score attains its maximum value it decreases with increase in k.

6

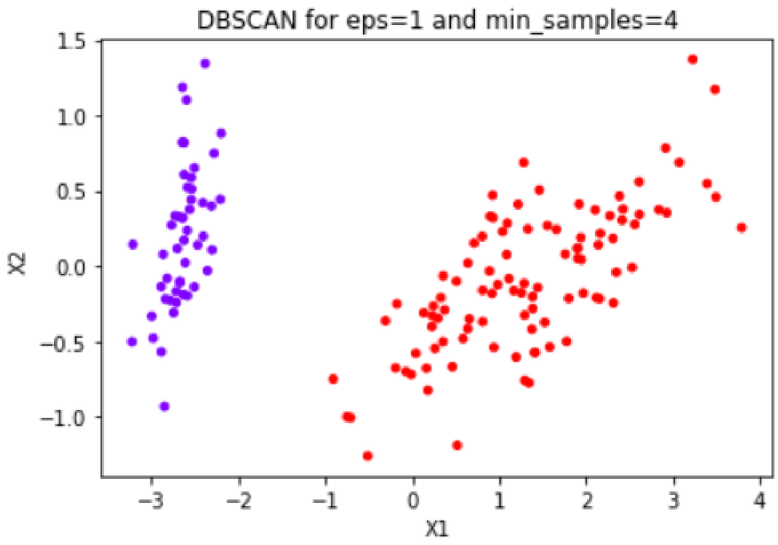

Figure 6 DBSCAN clustering on Iris flower dataset

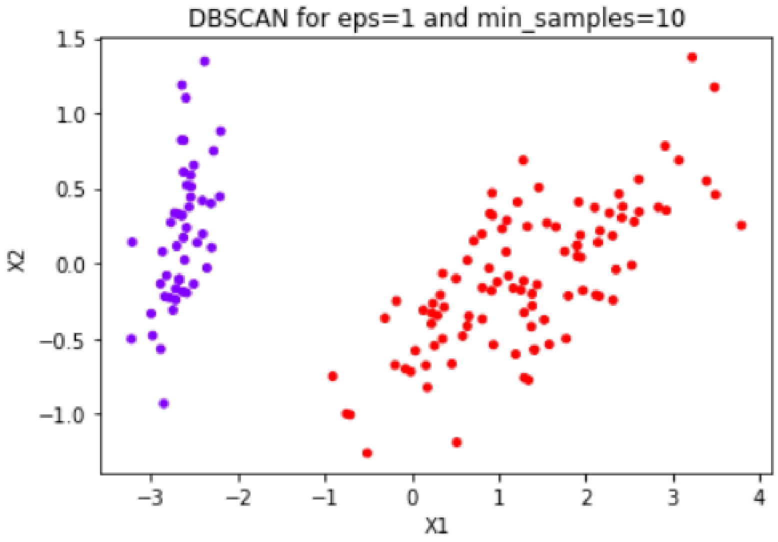

Figure 7 DBSCAN clustering on Iris flower dataset

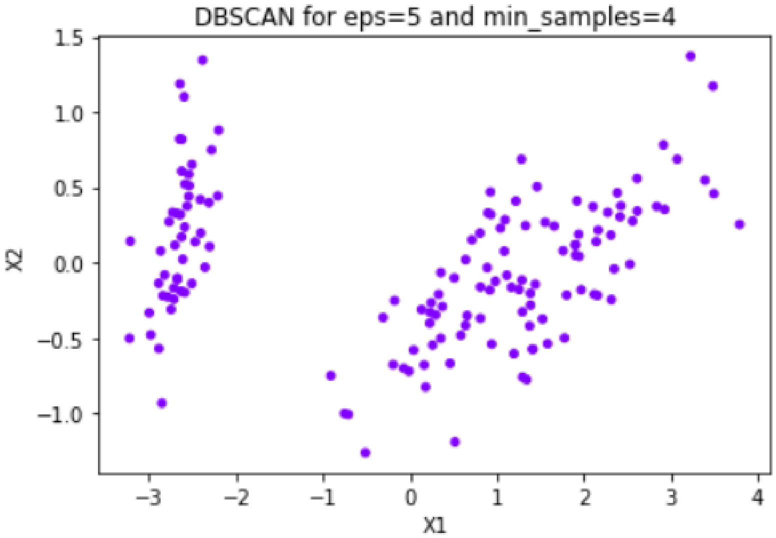

Figure 8 DBSCAN clustering on Iris flower dataset

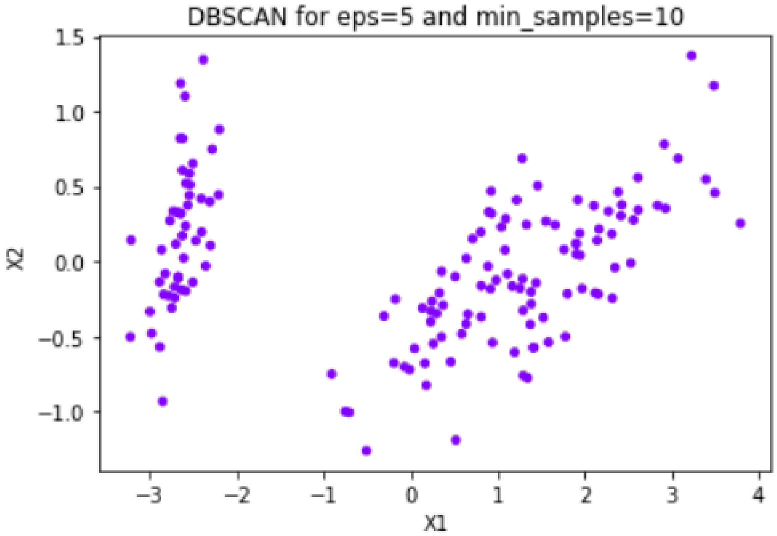

Figure 9 DBSCAN clustering on Iris flower dataset

Inferences:

1. For the optimum values of esp and min_samples the xlustering prowess is good.
2. For questions 2.a and 4.a we have manually given the no. of clusters while this algorithm decides the no. of clusters by itself.

b.

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

Inferences:

1. For the same eps value, increasing min_samples purity score is same.
2. For the same min_samples, increasing eps value the purity score is same.