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

Student's Name: Amit Maindola **Mobile No: +91 7470985613**

Roll Number: B20079 **Branch: CSE**

1

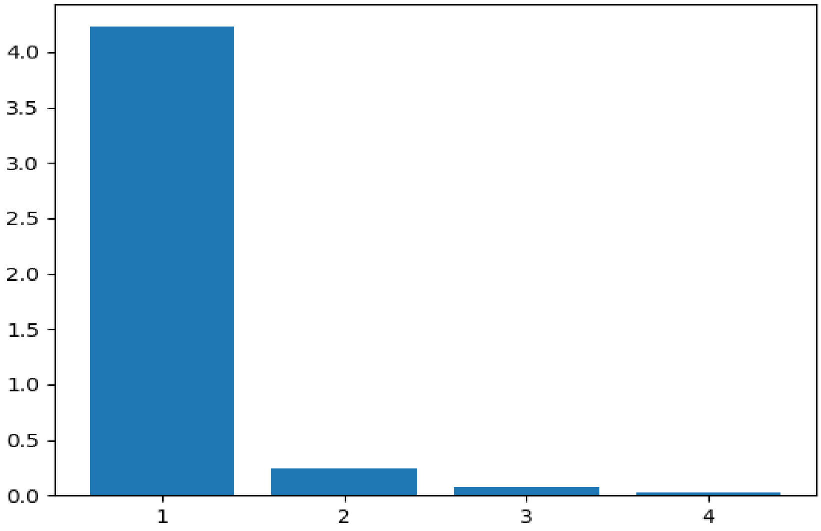

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigenvalue is decreasing with every successive increase in component value.
2. Eigenvalue represent variance of components so, it is quite natural that some of the components will cover more variance compared to other. Higher eigenvalues represents whole dataset better compared to others.

2 a.

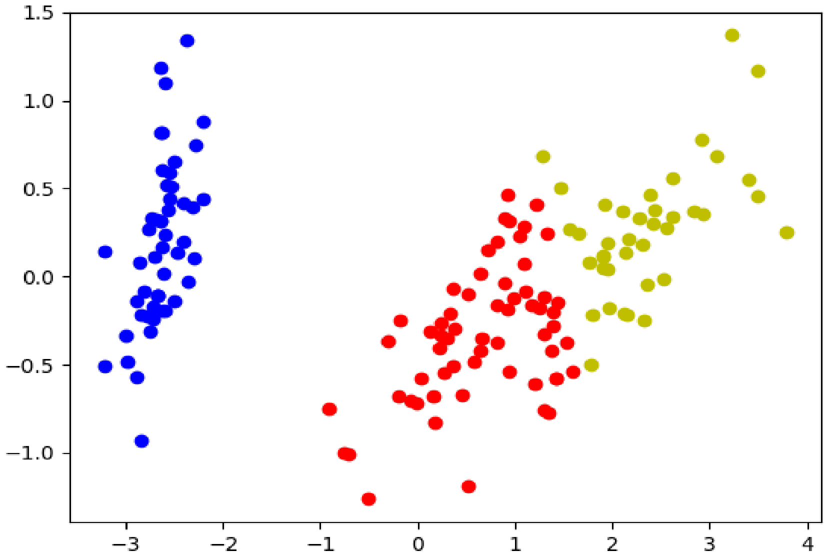

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, we can say that clustering prowess of algorithm is very good.
2. No, boundaries of clustering are more in straight line.

b. The value for distortion measure is **63.874**

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

3

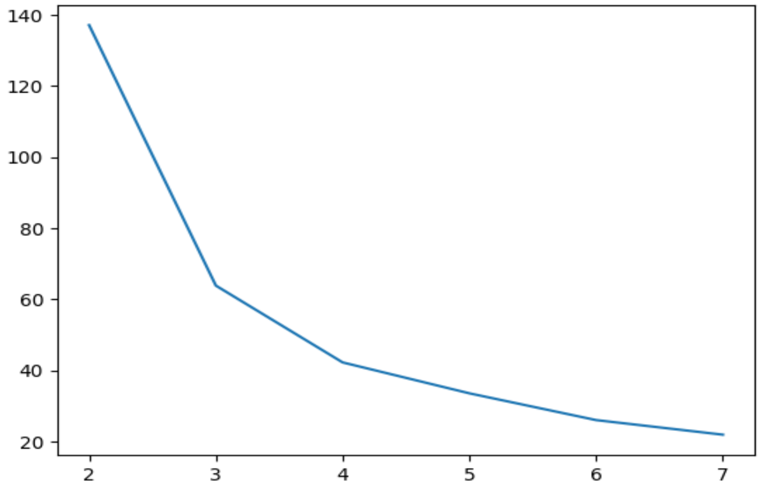

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases on increasing the value of K.
2. The distortion measure is the sum of squared distance of the points from center, as we increase the number of clusters the data get closer to respective centers and the distance decreases therefore distortion decreases.
3. There should be 3 clusters only in the dataset. No, Kmeans is unsupervised clustering hence it did not use labels of the data points and only 2 clusters can be seen therefore it gives 2 as optimum number 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. Initially increases (here till 3) then decreases.
3. Since, in original dataset there are only 3 labels. So, if we increase the value of k more than 3 than some points will get assigned to which they don't which leads in decrease in purity score..
4. Yes, after maximum value of purity score, its value decreases with the increase in K.

4 a.

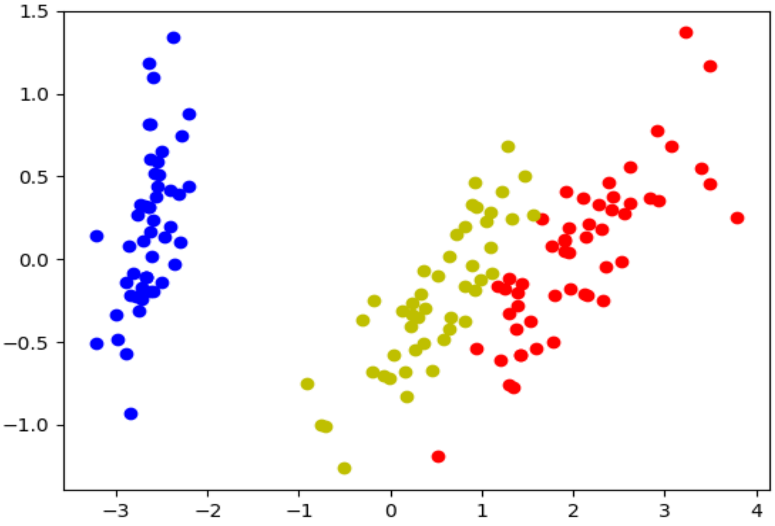

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, we can say that clustering prowess of algorithm is very good.
2. No, the boundary do not seem to be circular?
3. No.

b. The value for distortion measure is **-280.960**

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

5

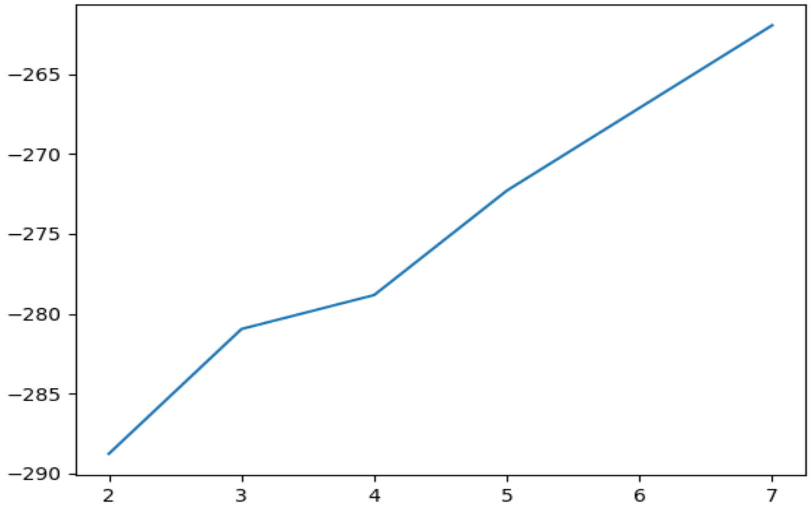

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Magnitude of Distortion measure increases with increase in no. of clusters.
2. By intuition, there should be 3 clusters only in the dataset. No, since GMM is unsupervised clustering therefore 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	0.980
4	0.820
5	0.773
6	0.747
7	0.680

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. Since 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 value of k.
5. Compare K-means and GMM based on inferences in Q3 and Q5.

6

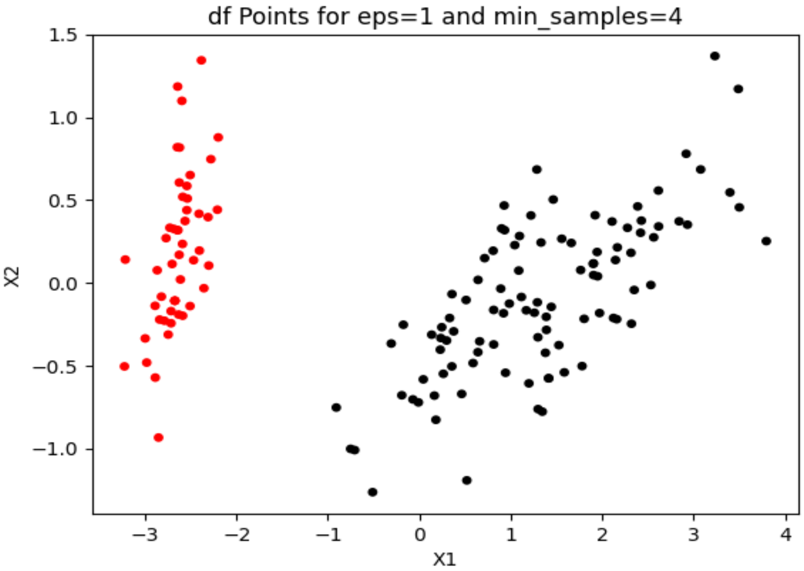

Figure 6 DBSCAN clustering on Iris flower dataset

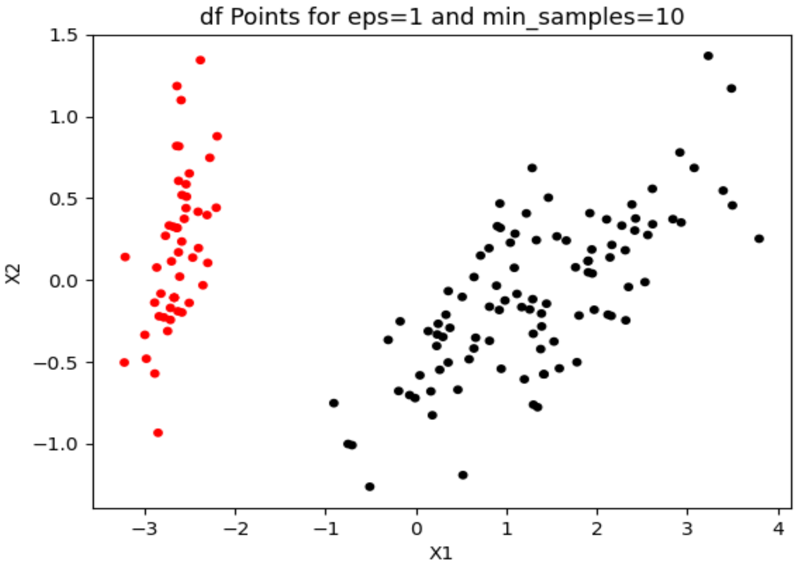

Figure 7 DBSCAN clustering on Iris flower dataset

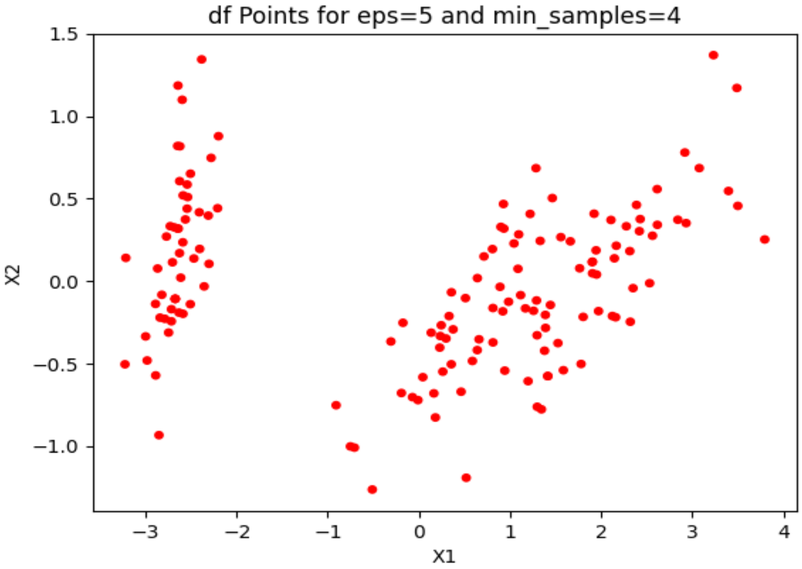

Figure 8 DBSCAN clustering on Iris flower dataset

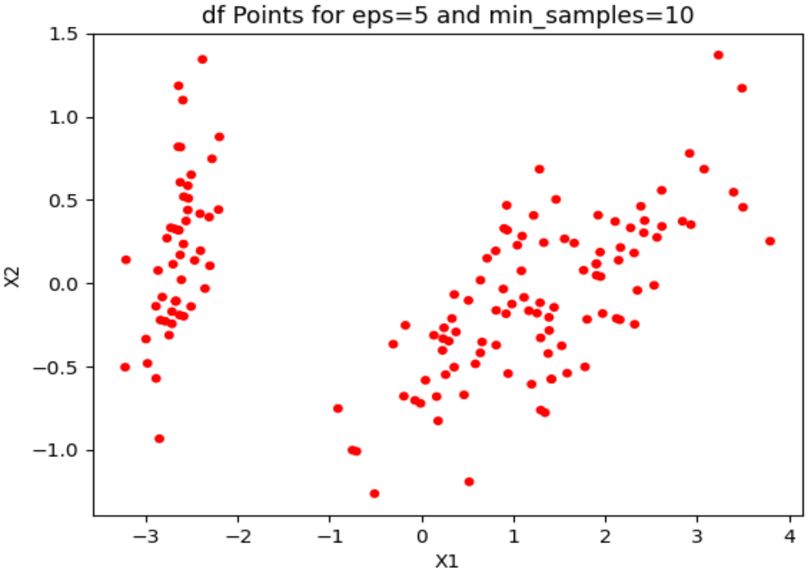

Figure 9 DBSCAN clustering on Iris flower dataset

Inferences:

1. For optimum values of esp and min_samples the clustering prowess of the algorithm is good.
2. For questions 2(a) and 4(a) we have manually given the number of clusters while this algorithm decides the number 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.