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

Student's Name: Shubham Shukla **Mobile No: 8317012277**

Roll Number: B20168 **Branch: CSE**

1

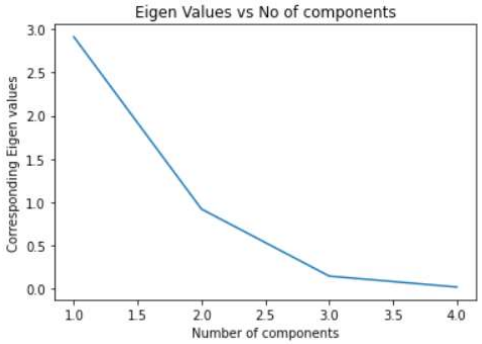

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigenvalue decrease corresponding to each component increase.
2. As the number of eigenvalues increases then the less infered data is also we have so our eigenvalues decreases.

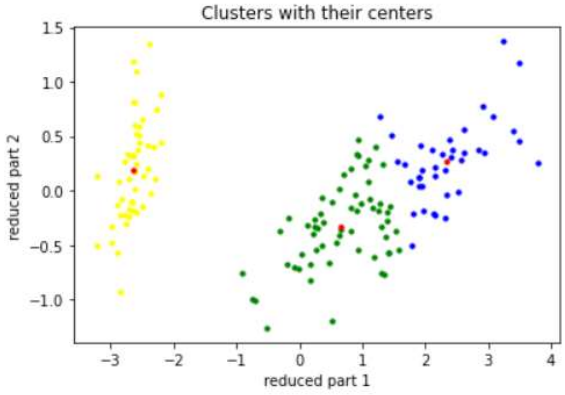

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. Clustering prowess of the algorithm is very fine.
2. No, the boundary seem more to be 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`.

2

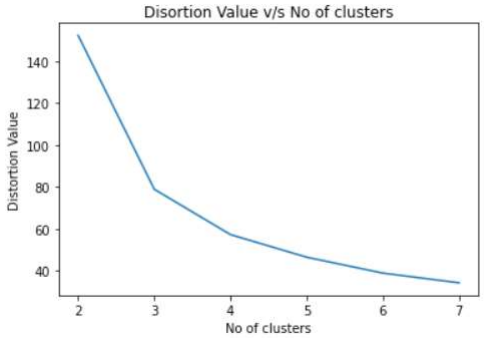

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure decreases with an increase in K.
2. As we have more number of clusters so we are more near to real data hence we will get less distortion value.

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

K value	Purity score
2	0.667
3	0.893
4	0.88
5	0.907
6	0.907
7	0.967

Inferences:

1. The highest purity score is obtained with K =7.
2. Increasing the value of K increases the purity score.
3. As we have more number of clusters so we are approaching more to real data hence we will get less distortion value so purity score increases.
4. Purity score is inversely proportional to distortion measure.

3 a.

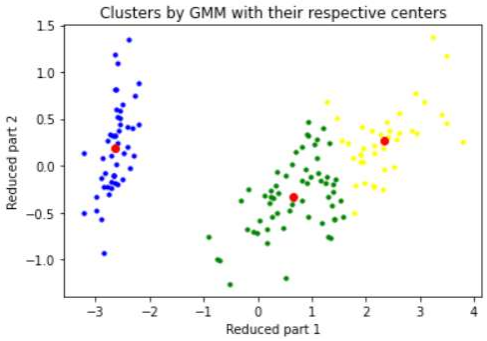

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. Clustering prowess of the algorithm is very good.
2. No, the boundary seem more to be straight line.

b. The value for distortion measure is `-16316.773`.

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

4

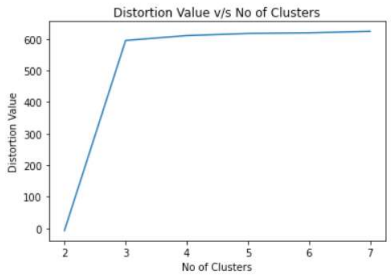

Figure 5 Number of clusters(K) vs. distortion measure

Inferences

1. Distortion measure increase with an increase in K.
2. We can see that boundary doesn't matching on increasing so distortion increases.

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

K value	Purity score
2	0.667
3	0.887
4	0.887
5	0.887
6	0.887
7	0.96

Inferences:

1. The highest purity score is obtained with K = 7.
2. Increasing the value of K increases the purity score.
3. Purity score and distortion is direct.
4. K-means is better than GMM.

5

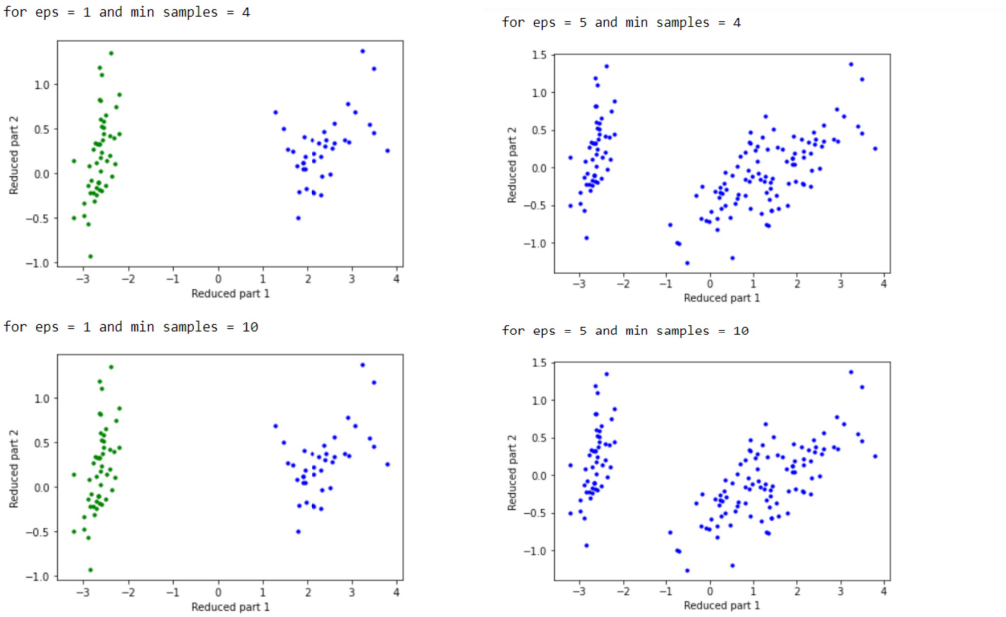

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Here the accuracy is not very good.
2. The number of clusters are less than those in K-means and also the boundaries are not defined.

b.

Eps	Min_samples	Purity Score
1	5	0.667
	10	0.887
4	5	0.667
	10	0.887

Inferences:

1. For the same eps value, Increasing min_samples doesn't purity score.
2. For the same min_samples, increasing eps value increase purity score.