IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VII

Clustering

Student's Name: Vaibhav **Branch: CSE**

Roll Number: B20140

Mobile No: 9812276214

1.

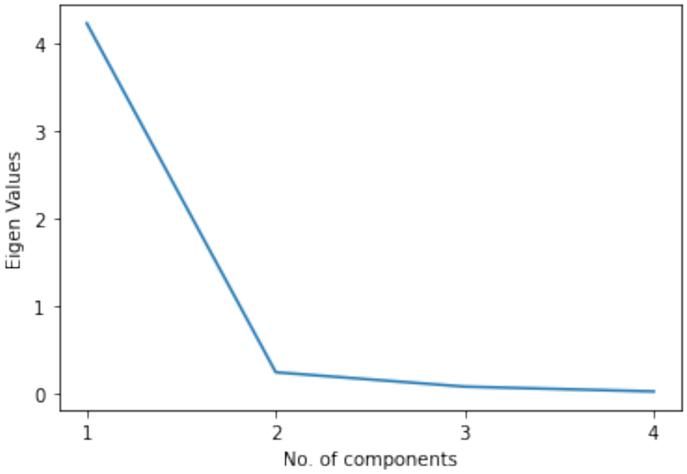

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigen value decreases as the number of components increases. Eigen value represents the amount of variance captured by the Eigen vector corresponding to that Eigen value. The amount of variance captured by each principal component decreases as the number of component increases. Hence, the Eigen value decreases as number of PC increases.

2. a.

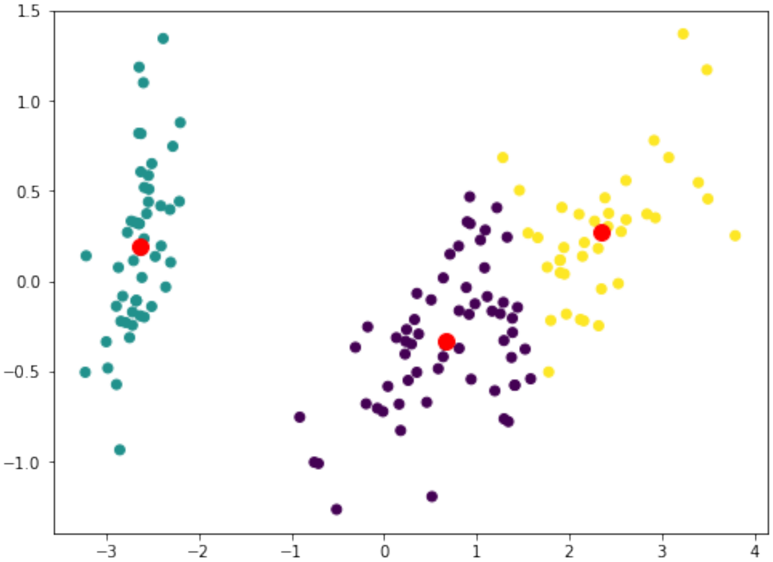

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering prowess of the algorithm is good.
2. If two clusters are intersecting each other they will have a linear boundary between them because cluster is assigned to a sample based on the Euclidean distance between the sample and the cluster centre.

b. The value for distortion measure is 63.87.

c. The purity score after examples are assigned to the clusters is 0.8867.

3.

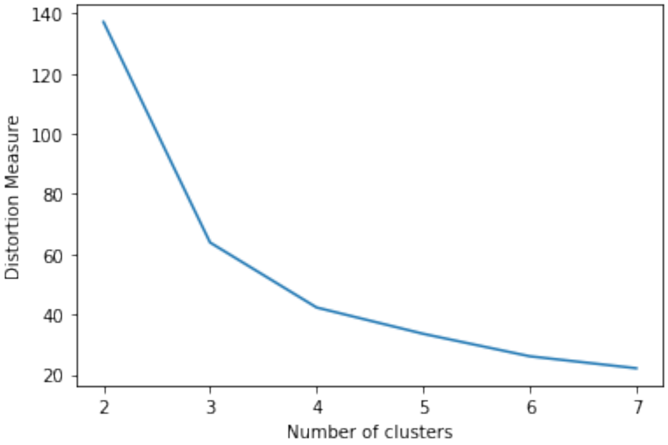

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with an increase in K.
2. As the number of clusters increase the cluster centres will be more closer to the sample and hence the distortion measure will be less.
3. From the number of species in the given dataset, intuitively the number of optimum clusters is 3. The elbow and distortion measure plot follow the intuition.

LAB ASSIGNMENT – VII

Clustering

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

K value	Purity score
2	0.667
3	0.887
4	0.693
5	0.68
6	0.527
7	0.52

Inferences:

1. The highest purity score is obtained with K = 3.

2. Purity score is max when we have optimal value of K. Optimal value of K depends on the dataset. If K is less model is under fitted while if K is large we have an overfitted model and for both of these purity score is less.

3. Purity score depends on choosing the optimal value of K. But the distortion measure if bound to decrease as K increases. Hence, there is no observable relationship between purity score and distortion measure.

4. a.

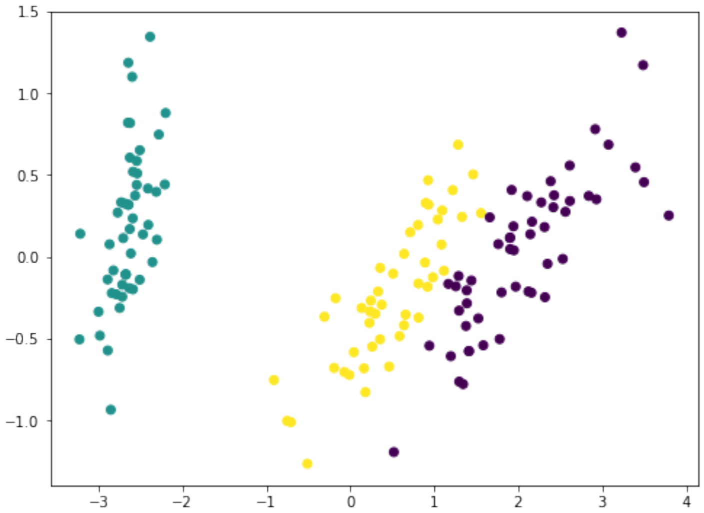

Figure 2 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering prowess of the algorithm is good.
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D. From the output, the boundary shape is not clear.
3. There is observable difference between clusters formed using K-means in 2.a and GMM in 4.a. The shapes of some of the clusters are different.

b. The value for distortion measure is -280.869.

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

5.

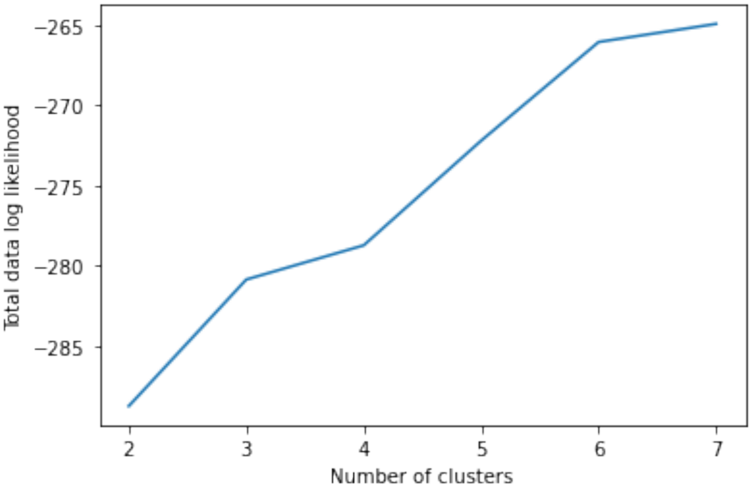

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Total data log likelihood increases as the number of clusters increases.
2. More the number of clusters better the model will fit on the training data. This is because the model moving towards becoming more and more overfitted.
3. From the number of species in the given dataset, intuitively the number of optimum clusters is 3. The elbow and distortion measure plot follow the intuition.

LAB ASSIGNMENT – VII

Clustering

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	0.773
6	0.7
7	0.627

Inferences:

1. The highest purity score is obtained with K = 3.

2. Purity score is max when we have optimal value of K. Optimal value of K depends on the dataset. If K is less model is under fitted while if K is large we have an overfitted model and for both of these purity score is less.

3. Purity score depends on choosing the optimal value of K. But the distortion measure if bound to decrease as K increases. Hence, there is no observable relationship between purity score and distortion measure.

4. GMM preformed better since the number of clusters were not very less (> 2).

6.

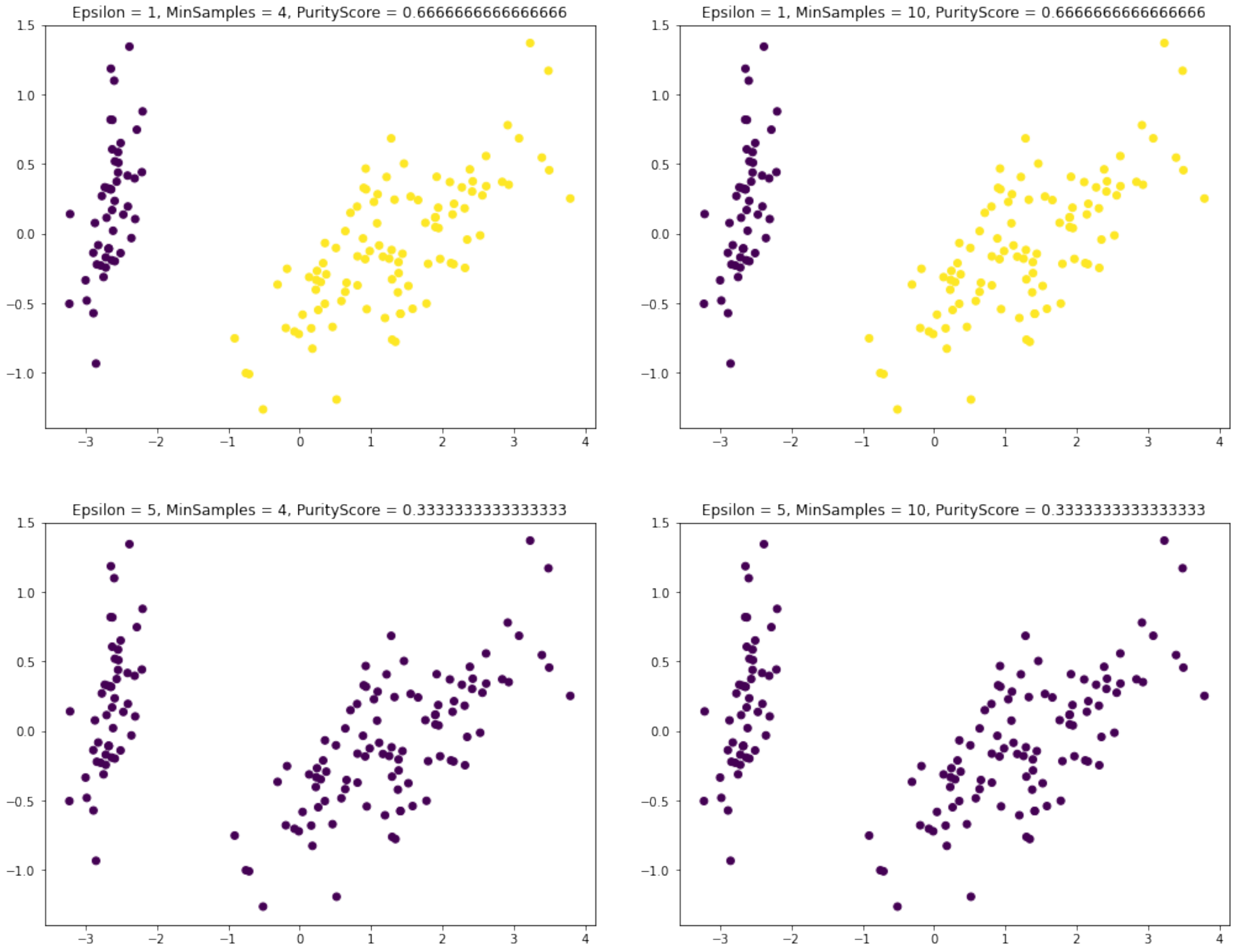

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. The clustering prowess of the algorithm is bad in this case.
2. There any observable difference between clusters formed using K-means in 2.a, GMM in 4.a and DBSCAN in 6.a.

b.

Eps	Min_samples	Purity Score
1	4	0.67
	10	0.67
5	4	0.33
	10	0.33

Inferences:

1. For the same eps value, does increasing min_samples has no effect on purity score in this case.
2. For the same min_samples, does increasing eps value decreases the purity score in this case.