IC 272: DATA SCIENCE - III

LAB ASSIGNMENT – VII

Clustering

Student's Name: Arpit Singh **Mobile No: 6265104315**

Roll Number: B20084 **Branch: CSE**

1

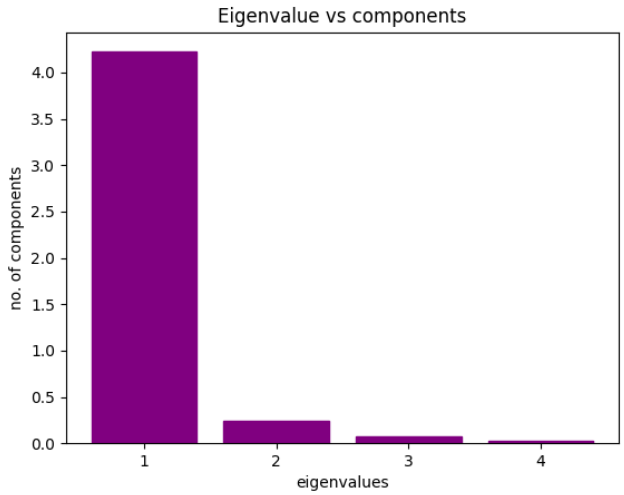

Figure 1 Eigenvalue vs. components

Inferences:

1. The eigenvalue decreases with increase of component value.
2. The eigenvalue decreases because the represented variance is more in 1st eigenvalues as it has higher covariance than others. All the eigenvalues follow this as some have more and some have les covariance.

2 a.

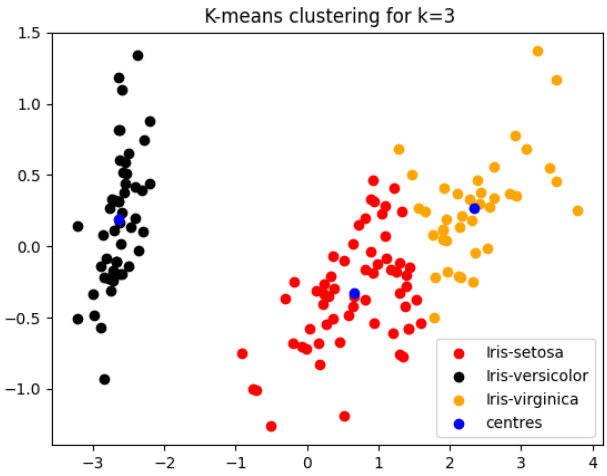

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, the clustering prowess of the algorithm is good.
2. The K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, we can see that boundaries are straight lines and not circular.

b. The value for the 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 an increase in K.
2. So, the distortion measure is the sum of squared distance of the points from the center. therefore, as we increase the no. of clusters the data gets closer to the centers and the squared distance decreases resulting in the decreased distortion measure.
3. From the number of species in the given dataset, intuitively the number of optimum clusters should be 3.No, the elbow and distortion measure plot follow the intuition.

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	0.68
6	0.527
7	0.5

Inferences:

1. The highest purity score is obtained with K = 3.
2. Initially it was increasing but after 3 increasing the value of K decreases the purity score.

3. As we have only 3 labels in our dataset, increasing the no. of clusters will lead to a point in a cluster which doesn't even exist in the original data that is the reason for decrease in the purity score.
4. When we get the maximum value of purity score, increasing no. of clusters will decrease the purity score.

4 a.

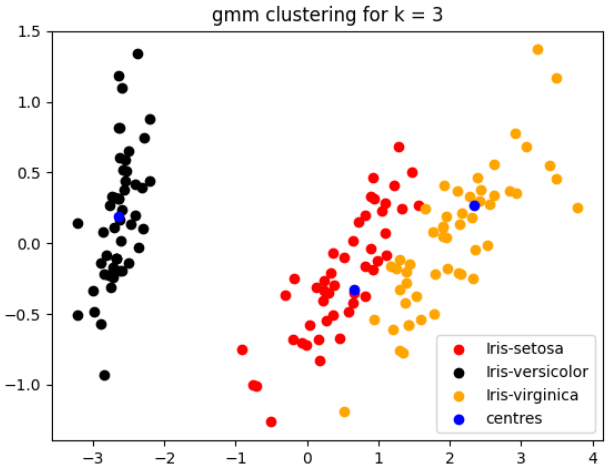

Figure 2 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, the clustering prowess of the algorithm is good.
2. The K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, we can see that boundaries are straight lines and not circular.
3. No, there is not any observable difference between clusters formed using K-means in 2.a and GMM in 4.a.

b. The value for the distortion measure is -280.96.

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

5

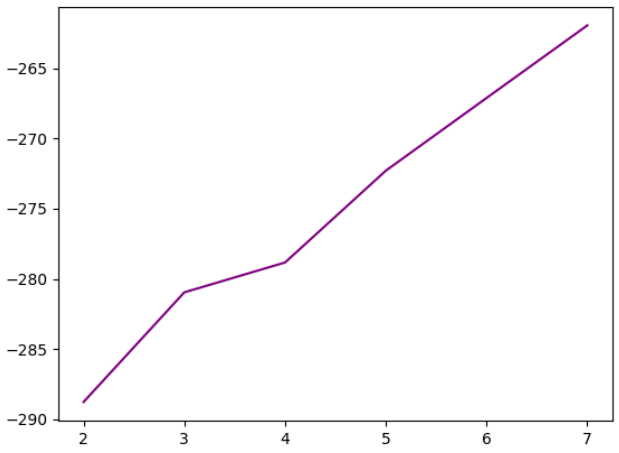

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increases with an increase in K.
2. So, the distortion measure is the sum of squared distance of the points from the center. therefore, as we increase the no. of clusters the data gets closer to the centers and the squared distance decreases resulting in the decreased distortion measure.
3. From the number of species in the given dataset, intuitively the number of optimum clusters should be 3.No, the elbow and distortion measure plot follow the intuition.

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	0.774
6	0.747
7	0.68

Inferences:

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

2. Initially it was increasing but after 3 increasing the value of K decreases the purity score.
3. As we have only 3 labels in our dataset, increasing the no. of clusters will lead to a point in a cluster which doesn't even exist in the original data that is the reason for decrease in the purity score.
4. The purity score of GMM is more than that of K-means.
5. When we get the maximum value of purity score, increasing no. of clusters will decrease the purity score.

6

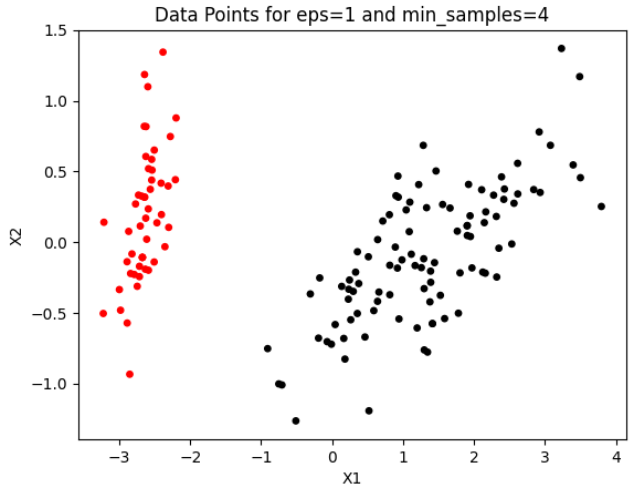

Figure 6 DBSCAN clustering on Iris flower dataset

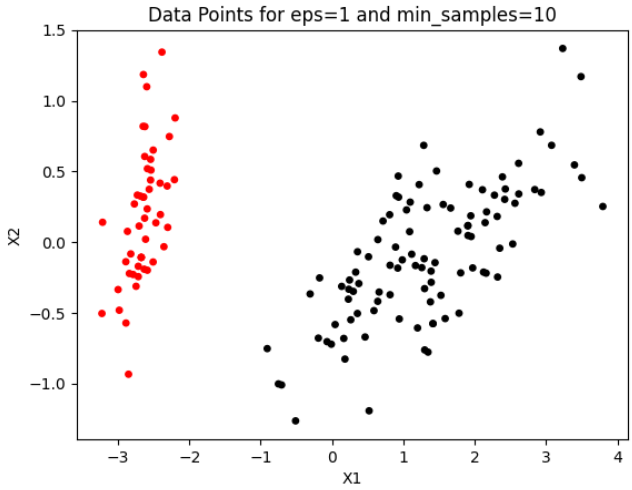

Figure 7 DBSCAN clustering on Iris flower dataset

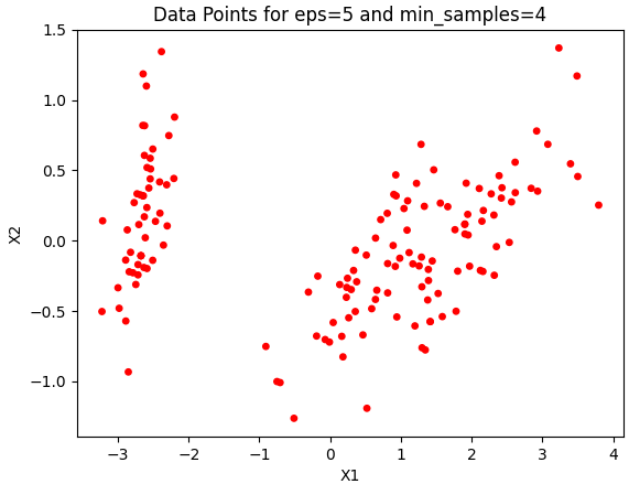

Figure 8 DBSCAN clustering on Iris flower dataset

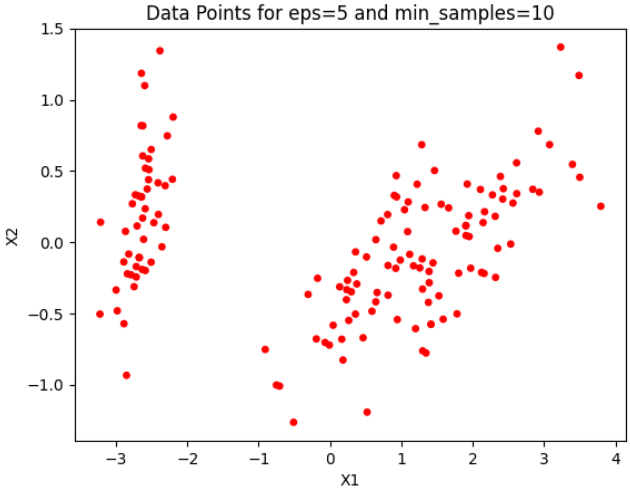

Figure 9 DBSCAN clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, the clustering prowess of the algorithm is good.
2. The observable difference between clusters formed using K-means in 2.a, GMM in 4.a and DBSCAN in 6.a is that the DBSCAN algorithm decides the no. of clusters by itself whereas GMM doesn't.

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 doesn't change purity score.
2. For the same min_samples, increasing eps value decreases purity score.