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

Student's Name: Nikhil **Mobile No: 8949463760**

Roll Number: B20219 **Branch:**CSE

1

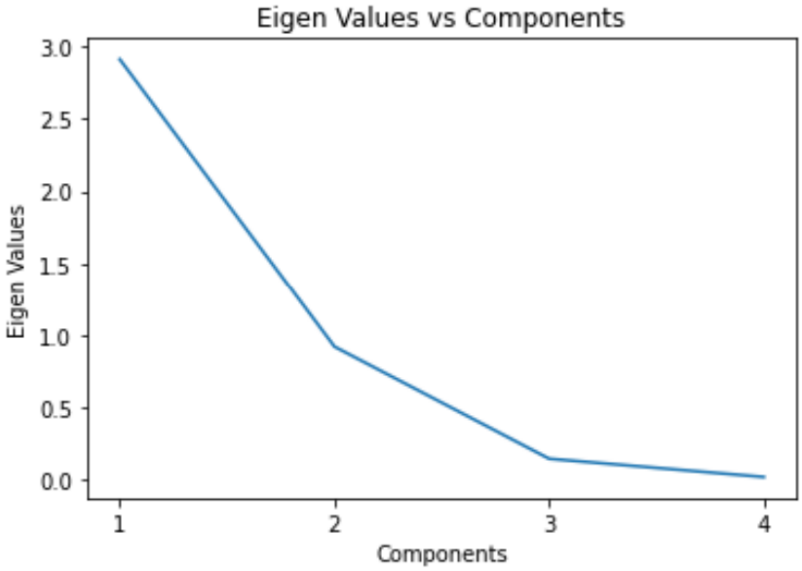

Figure 1 Eigenvalue vs. components

Inferences:

1. The eigenvalues show a decreasing trend with respect to increase in the component number.
2. The above trend is a result of the output we get by using the inbuilt function for PCA: The function gives the eigenvalues in the decreasing order. This trend has no relation with the mathematics behind calculation of the eigen values but only with how the inbuilt function is designed to give the output.

2 a.

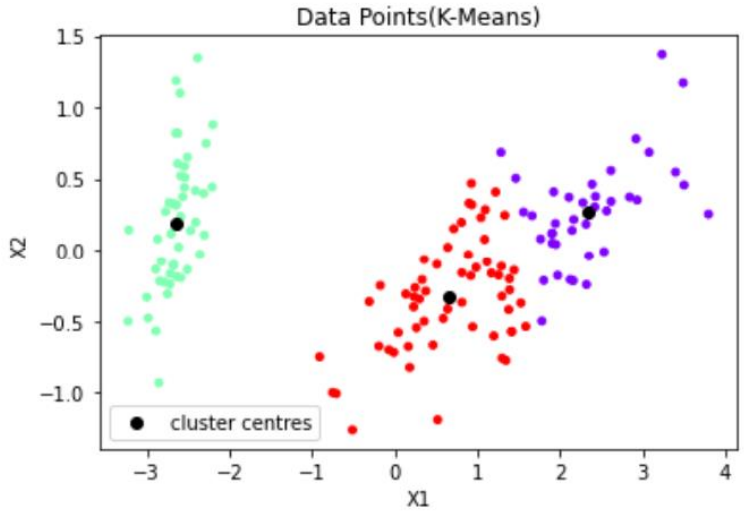

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering seems to be quite accurate, by inferring from the plot. The purity score is 0.887 which is quite high, so it works quite well.
2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, the boundaries do not seem to be that much circular for the purple cluster, but it seems to be fairly circular for other clusters.

b. The value for distortion measure is 63.874

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

3

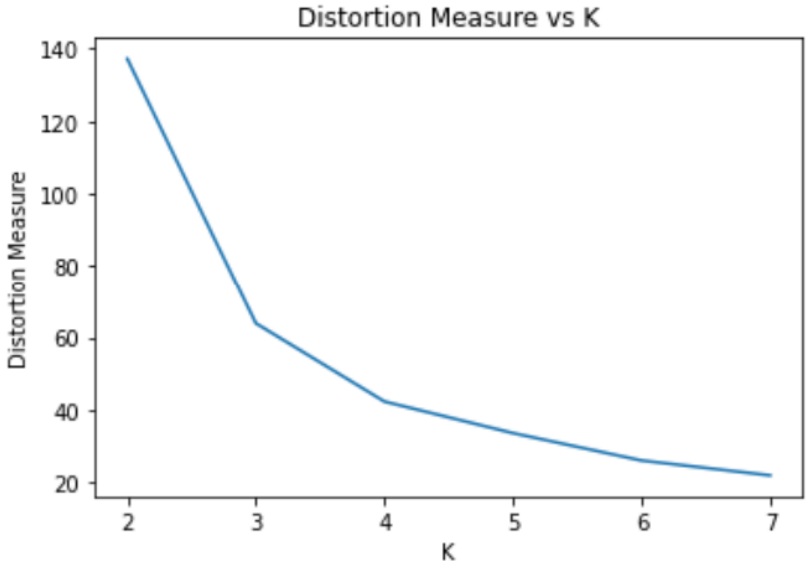

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with an increase in K.
2. This is because for the distortion measure here, we are just calculating the squared distances so as the number of clusters increase, most of the points which were far away from their cluster centers in a smaller number of total clusters now have a new center which is closer to them than in previous centers. Therefore, the overall distortion measure shows a decreasing trend.
3. From the number of species in the given dataset, intuitively the number of optimum clusters must be 3. From the given plot we see an elbow at k=3 so, the elbow measure also tells the optimum number of clusters to be 3.

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	0.667
6	0.520
7	0.493

Inferences:

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

2. The purity score increases till k=3 and then starts to decrease.
3. By the elbow measure, the optimum number of clusters must be 3. So, the purity score increases till we reach k=3 but then starts to decrease as we increase k.
4. There is a relationship between the purity score and the distortion measure, as the purity score decreases after we reach the elbow in the plot of distortion measure but increases before it.

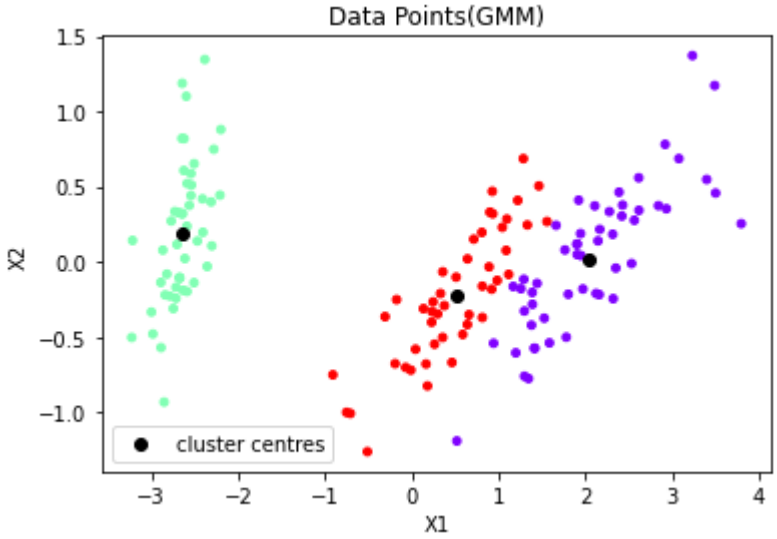

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering seems to be quite accurate, by inferring from the plot. The purity score is 0.98 which is so high, so it works very well.
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D. From the output, we can observe that the cluster boundaries are in fact elliptical.
3. The clustering by using GMM gives a different result for red and purple clusters than the red and purple clusters made in the k means algorithm previously.

b. The value for distortion measure (log likelihood) is -280.87

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

4

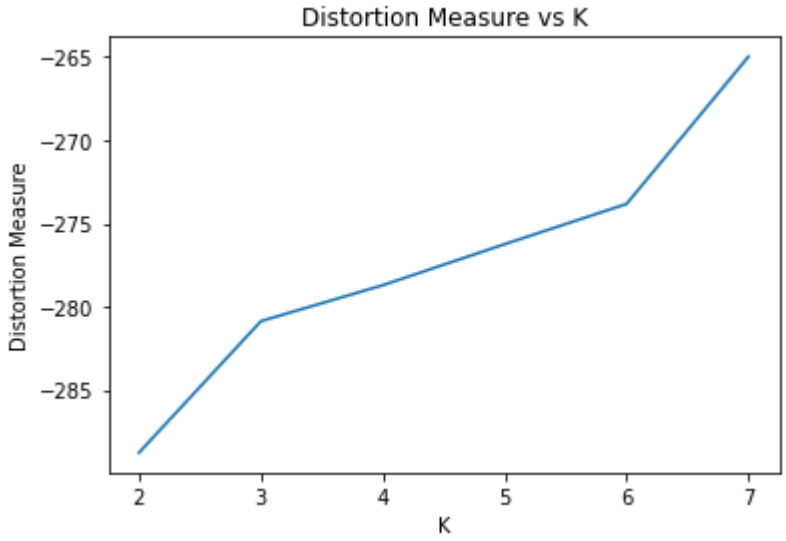

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The log likelihood (distortion measure) increases with increase in the k value.
2. This is because as we start to increase the number of clusters, the algorithm starts to assume the data coming from higher and higher modal densities, so it starts to give best approximation of the distribution of the data as we keep increasing k. Increasing number of modes can fit arbitrary kind of data.
3. From the number of species in the given dataset, intuitively the number of optimum clusters must be 3. From the given plot we see an elbow at k=3 so, the elbow measure also tells the optimum number of clusters to be 3.

[0.667, 0.98, 0.833, 0.767, 0.64, 0.627]

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

K value	Purity score
2	0.667
3	0.980
4	0.833
5	0.767
6	0.640
7	0.627

Inferences:

1. The highest purity score is obtained with K =3.
2. The purity score increases till k=3 and then starts to decrease.
3. By the elbow measure, the optimum number of clusters must be 3. So, the purity score increases till we reach k=3 but then starts to decrease as we increase k.
4. There is a relationship between the purity score and the distortion measure, as the purity score decreases after we reach the elbow in the plot of distortion measure but increases before it.
5. On the given dataset, the GMM model proved to be better as it gives the purity score to be 0.98 while the k means algorithm gives it to be 0.877.

5

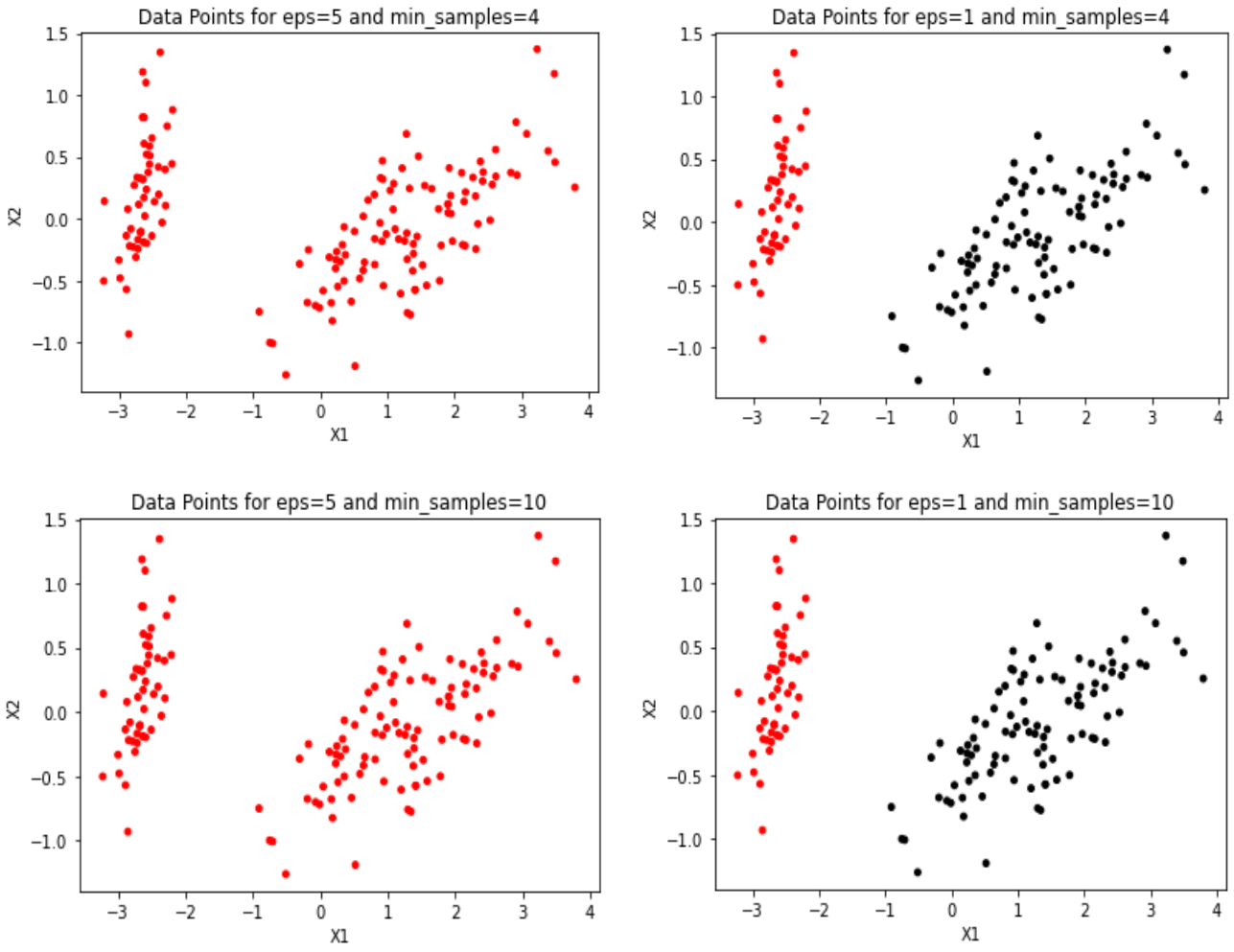

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Inferring from the clusters formed in the above plot, the clustering does not seem to be that accurate. For some plots it shows just 1 cluster which is a bad result.
2. The number of clusters is less than that in kmeans or GMM. Also, the boundaries are neither circular nor elliptical.

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 does not change the purity score.
2. For the same min_samples, increasing eps value decreases the purity score.