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

Student's Name: Khushi Ladha **Mobile No: 7665519043**

Roll Number: B20013 **Branch:CSE**

1 a

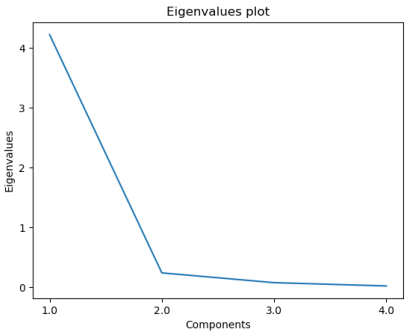

Figure 1 Eigenvalue vs. components

Inferences:

1. Eigenvalue decrease corresponding to each component increase. There is high decrease till component = 2. After that decrease is less.
2. Eigenvalues decreases corresponding to increase in component because the attributes are more dependent on the first Eigen value so they have more spread around it.

1 b.

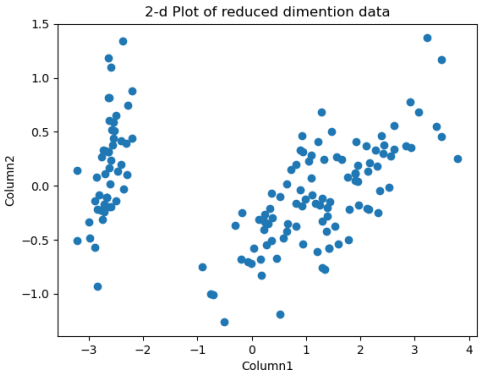

Figure 2 plot of reduced dimenetional data

Inferences:

1. 2 distinct clusters are visible.

2 a.

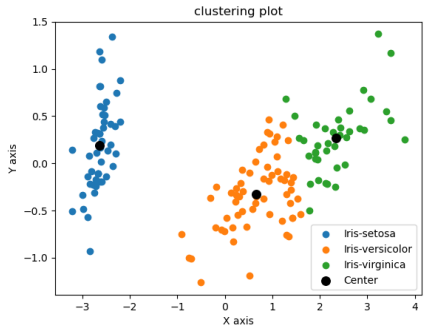

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering has been done fairly well with purity score of 0.887 and is able to make fairly well formed clusters.
2. No, all boundaries do not seem circular.

b. The value for distortion measure is 63.874

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

3

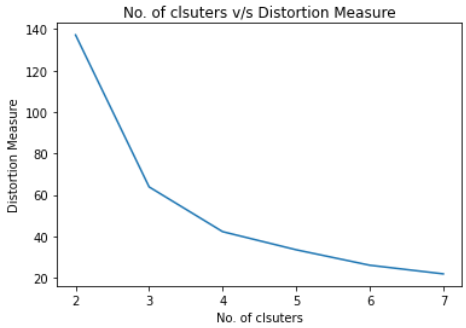

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. distortion measure decreases with an increase in K.

2. As more clusters are created, more centers are created. Distortion measure is sum of distance of cluster points form centers and hence distortion measure also decreases with increasing k.
3. From the number of species in the given dataset, intuitively k = 3 be the number of optimum clusters. Yes, 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	0.66
6	0.527
7	0.493

Inferences:

1. The highest purity score is obtained with K =3.
2. Initially, purity increases with increasing k value. After that, it starts to decrease. Purity is highest at the optimum k value.
3. Closer to the optimum k value, purity score gets higher, while away from it, it starts to decrease. This is because lesser or more number of clusters than optimally required are present.
4. Distortion measure decreases for all increasing value of K, while same trend isn't for purity score. They follow similar trend after purity score's k value has reached optimum k, both start decreases after that. Also, the elbow point of distortion measure gives the k value which would give maxim purity score.

4 a.

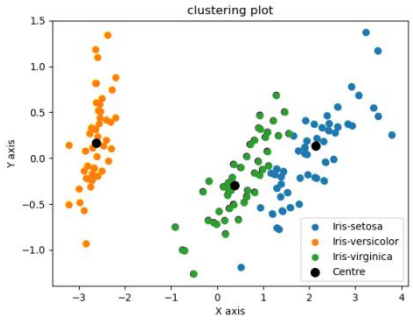

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. the clustering prowess of the algorithm is fairly good as the clusters are separated in elliptical boundaries.
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D and yes the boundaries are eleptical fairly enough.
3. Yes, the clusters are different. The GMM clustering works better as it assumes clusters to be elliptical and the clusters are elliptical as well, whereas in k means some clusters were not circular.

b. The value for distortion measure is -280.87

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

5

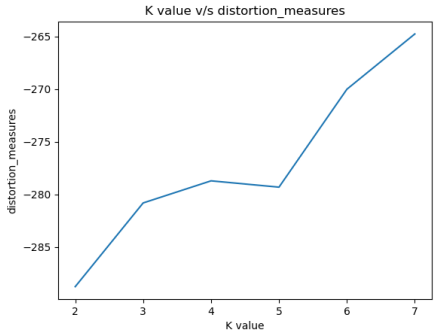

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increase with an increase in K.
2. hence the distortion measure increases fast from k=2 to k =3 . The rate of increase decreases from k = 3 to k = 6 and then it increases fast again after k = 6 .
3. From the number of species in the given dataset, intuitively what should be the number of optimum clusters , k = 3 . Yes, 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	0.827
5	0.6
6	0.633
7	0.613

Inferences:

1. The highest purity score is obtained with K = 3 .
2. Initially, purity increases with increasing k value. After that, it starts to decrease. Purity is highest at the optimum k value.
3. Closer to the optimum k value, purity score gets higher, while away from it, it starts to decrease. This is because lesser or more number of clusters than optimally required are present.
4. Distortion measure decreases for all increasing value of K, while exactly same trend isn't for purity score. They follow similar trend after purity score's k value has reached optimum k, both start decreases after that. Also, the elbow point of distortion measure gives the k value which would give maxim purity score.
5. For all k values including optimal and non optimal k value, higher purity score is obtained in GMM clustering only. Hence, for this sample data GMM is the better technique.

6

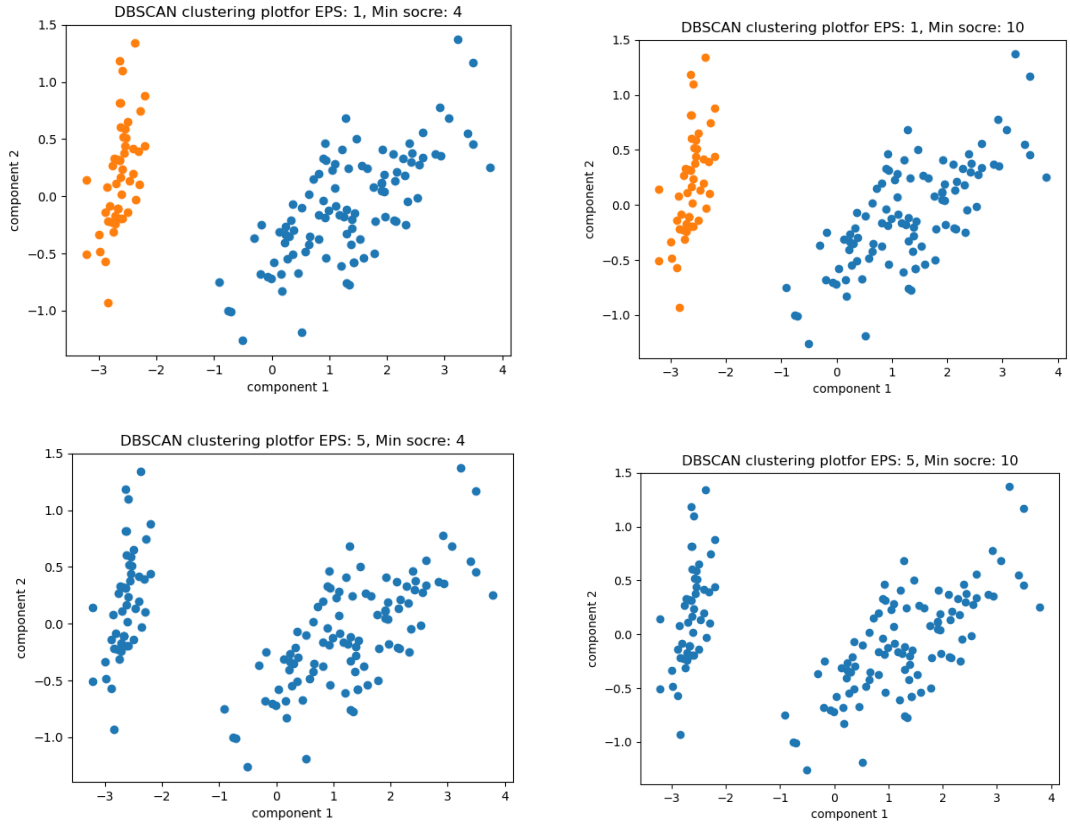

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. The accuracy isn't good. The reason might be that optimum vaue of eps isn't taken.
2. The number of clusters are less than that those in K-means and GMM and also the boundaries are neither circular nor elliptical in DBSCAN.

b.

Eps	Min_samples	Purity Score
1	4	0.667
	10	0.667
6	4	0.333
	10	0.333

Inferences:

1. Min_samples doesn't affect purity scores value.
2. Increasing EPS_Value decreases purity score for same min_samples.