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

Student’s Name: Priyanka Kumari **Mobile No: 8328354314**

Roll Number: B20307 **Branch: ELECTRICAL ENGINEERING**

1

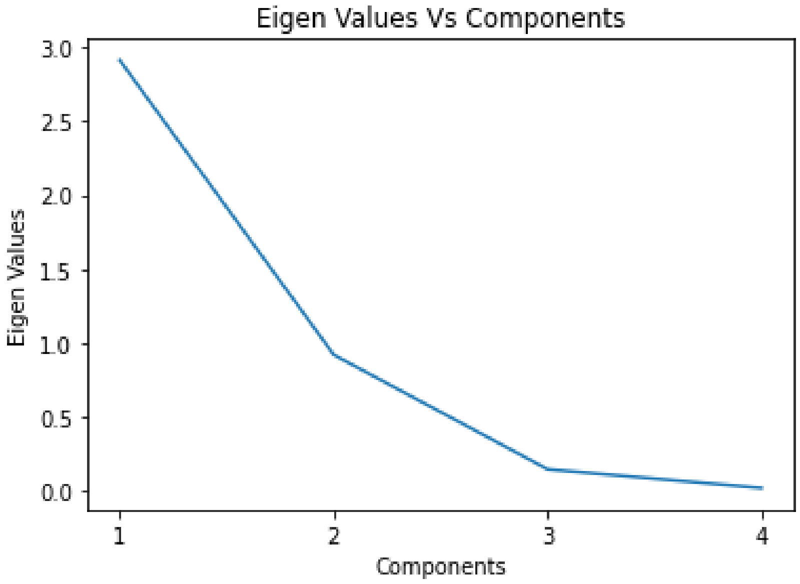


Figure 1 Eigenvalue vs. components

Inferences:

1. Eigenvalue decrease on increasing component successively.
2. Because the attributes are more dependent on the first eigen value so it has more spread around it.

2 a.

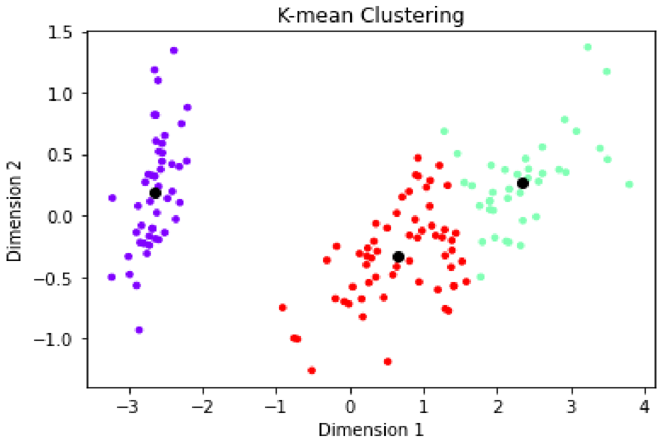


Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. Clustering looks quite good and accurate as K-Means is an unsupervised algorithm, K-Means can provide well-formed clusters.
2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, the boundary doesn't seem to be clearly circular but it is fairly circular.

b. The value for distortion measure is 64.3

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

3

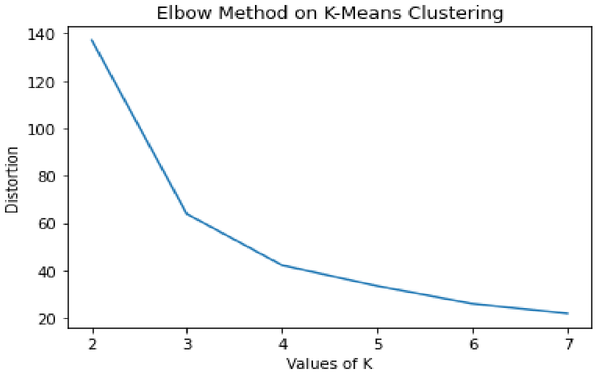


Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure decreases with an increase in K.
2. Because number of species in our sample is 3 so the distortion measure decreases drastically for k=2 to k=3 and then decreases very gradually.
3. The number of species in the given dataset, intuitively k=3 should be the number of optimum clusters. The elbow and distortion measure plot closely 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	0.693
5	0.68
6	0.507
7	0.507

Inferences:

1. The highest purity score is obtained with K = 3 .
2. Purity score increases from k=2 to k=3 and then decreases with increasing value of K.
3. Because the number of species in our data is 3 so purity score for k=3 comes out to be the highest.
4. Yes, except k=3, as the distortion measures decreases purity score increases.

4 a.

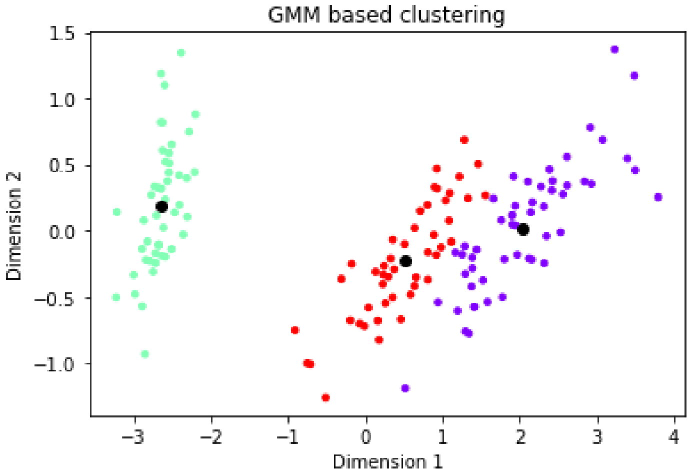


Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. GMM looks quite accurate as the predicted results are very close to the actual ones.
2. GMM algorithm assumes cluster boundaries to be elliptical in 2D. From the output, the boundary seem fairly elliptical not clearly though.
3. Yes from the graphs we can see that the boundaries in K-means were circular while in GMM they are elliptical.

b. The value for distortion measure is -281.3

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

5

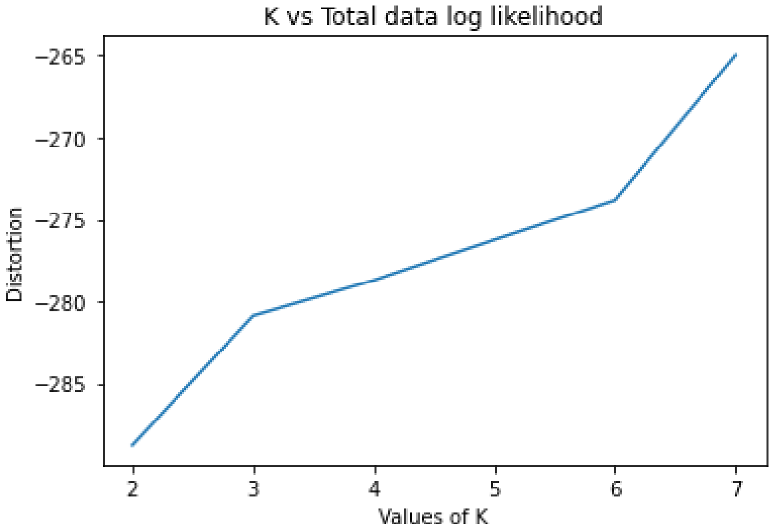


Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increase with an increase in K.
2. As the number of species is 3 so the distortion measure relatively has more slope between k=2 and k=3 and then it increases gradually till k=6 and abruptly increases after that.
3. From the number of species in the given dataset, intuitively k=3 be the number of optimum clusters? The elbow and distortion measure plot follow the intuition closely.

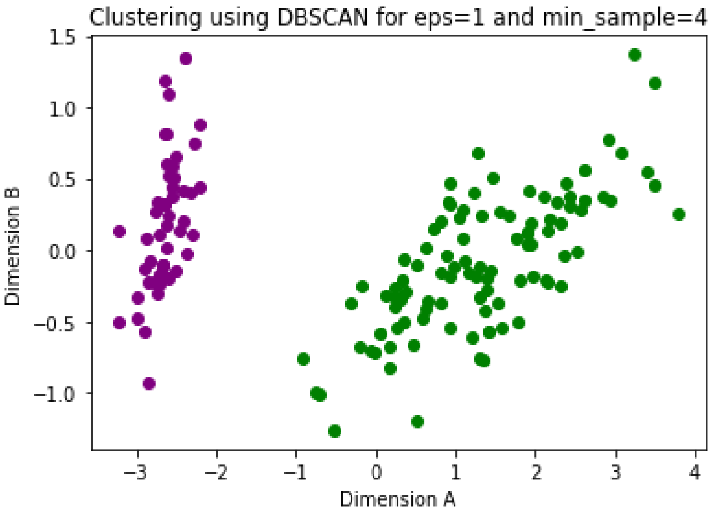
Table 2 Purity score for K value = 2,3,4,5,6 & 7

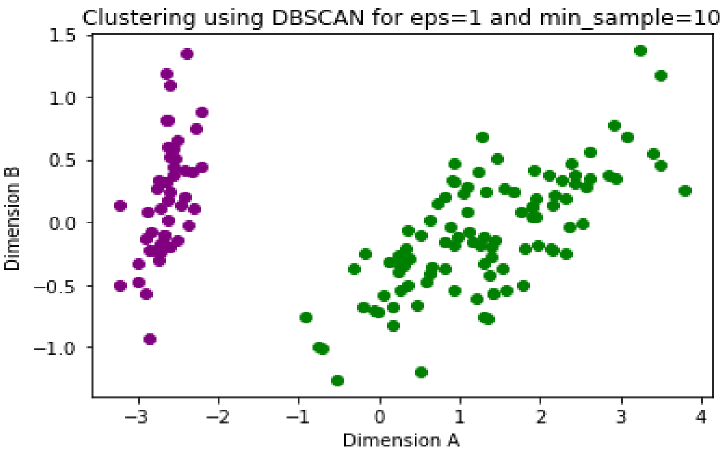
K value	Purity score
2	0.667
3	0.98
4	0.833
5	0.767
6	0.64
7	0.627

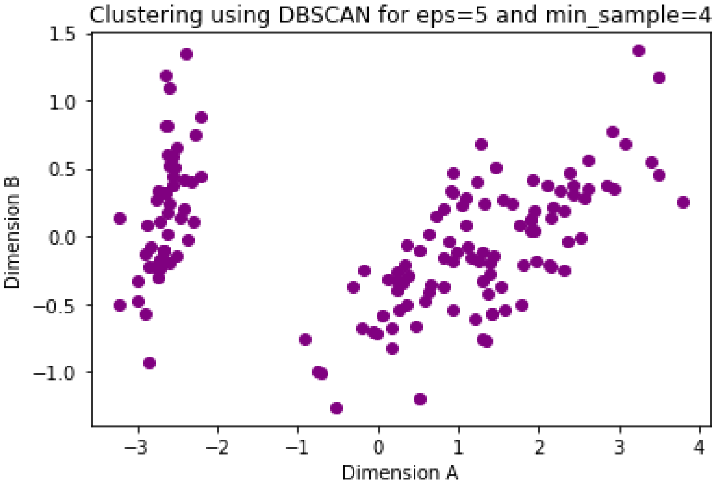
Inferences:

1. The highest purity score is obtained with K = 3 .
2. Purity score increases from k=2 to k=3 and then decreases with increasing value of K.
3. Because the number of species in our data is 3 so purity score for k=3 comes out to be the highest.
4. Yes, except k=3, as the distortion measures decreases purity score increases.
5. By the inferences we can see that GMM is more accurate than K-means.

6.




Clustering using DBSCAN for eps=1 and min_sample=10
1.5
1.0
0.5
0.0
−0.5
−1.0
Dimension B
−3
−2
−1
0
1
2
3
4
Dimension A


Clustering using DBSCAN for eps=5 and min_sample=4
1.5
1.0
0.5
0.0
−0.5
−1.0
Dimension B
−3
−2
−1
0
1
2
3
4
Dimension A

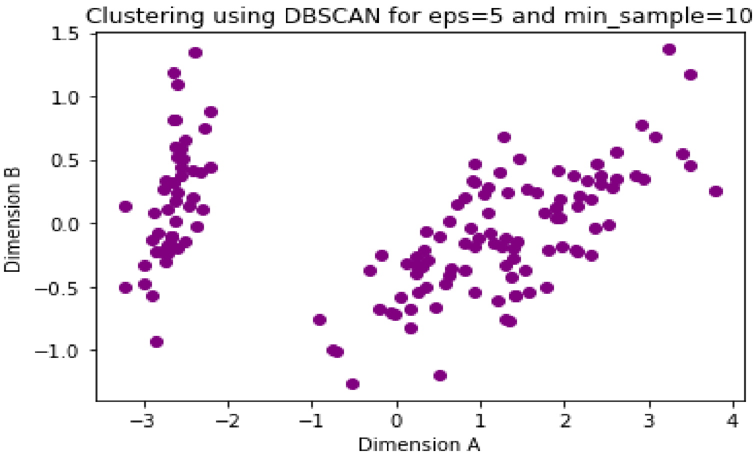

Clustering using DBSCAN for eps=5 and min_sample=10
1.5
1.0
0.5
0.0
−0.5
−1.0
Dimension B
−3
−2
−1
0
1
2
3
4
Dimension A

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Here the accuracy is not very good one reason might be our choice of value of eps.
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
4	4	0.333
	10	0.333

Inferences:

1. For the same eps value, increasing min samples don’t change purity score.
2. For the same min samples, increasing eps value decrease purity score.