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

Student's Name: Akshar Singh **Mobile No: 7428357700**

Roll Number: B20147 **Branch:** CSE

1

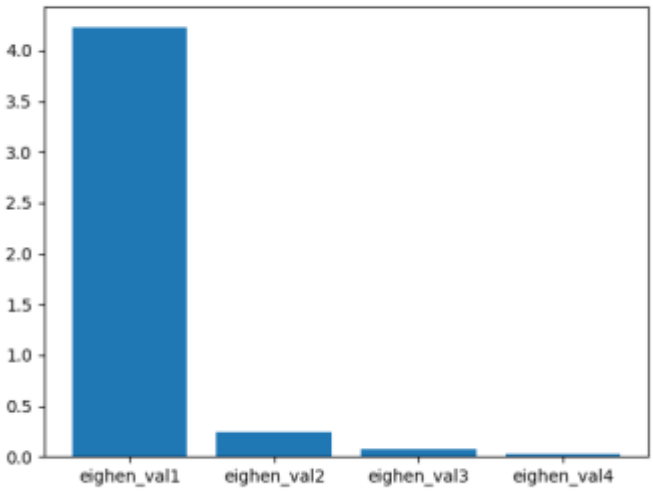

Figure 1 Eigenvalue vs. components

Inferences:

1. The eigenvalue decreases with increase in component.
2. There is a decrease in eigen values since the variance of the reduced data decreases with increase in components

2 a.

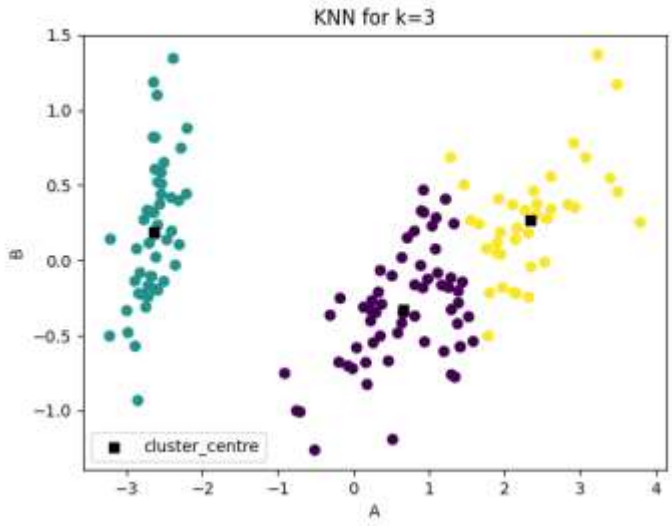

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. There are 3 clusters formed by recomputing mean and measuring distance of data points iteratively.
2. The boundary seems to be linear as Euclidean distance formula is used to calculate the distance.

b. The value for distortion measure is 63.873

c. The purity score after examples are assigned to the clusters is 0.886

3

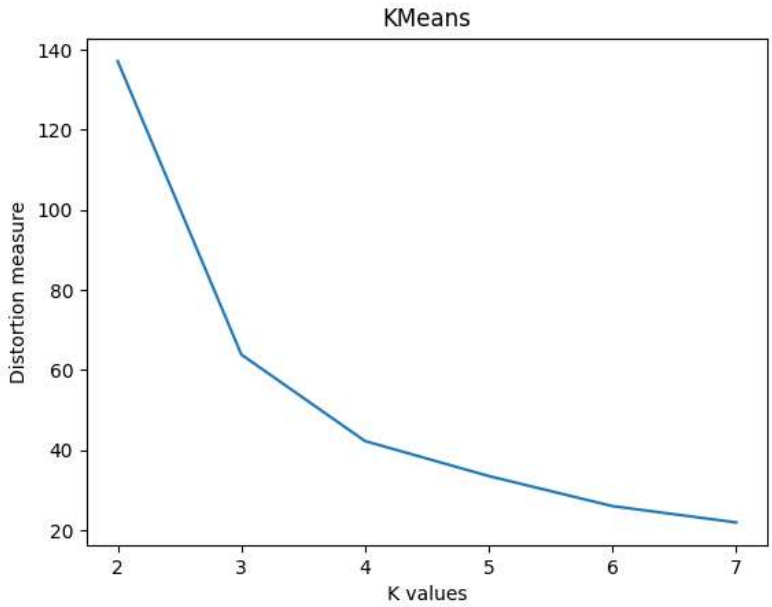

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with increase in K values.
2. This happens because as the number of groups increases the square distances decreases because of compactness.
3. The number of optimum clusters should be 3 and 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	0.693
5	0.68
6	0.526
7	0.52

Inferences:

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

2. Purity score increases upto the optimal value of K=3 and then decreases because at the optimal value the number of clusters is same as that given in the dataset.
3. There is no such observable relationship but one can deduce that at the elbow the purity score would be highest.

4 a.

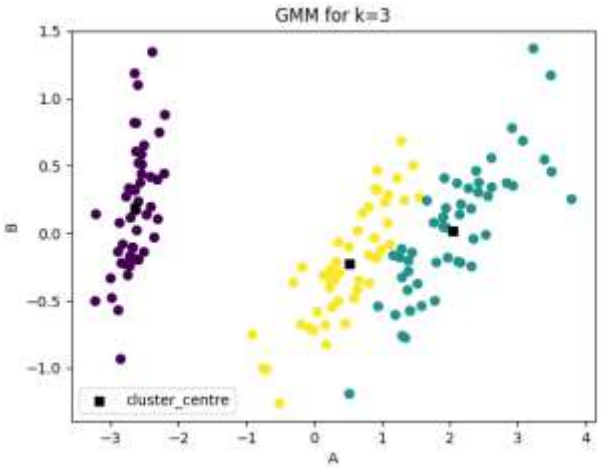

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. The clustering process seems to be soft clustering since each element has a probability to be in each cluster.
2. The boundary seems to be circular
3. There is a difference in the shape and boundary of clusters between K-means and GMM.

b. The value for distortion measure is -288.960

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

5

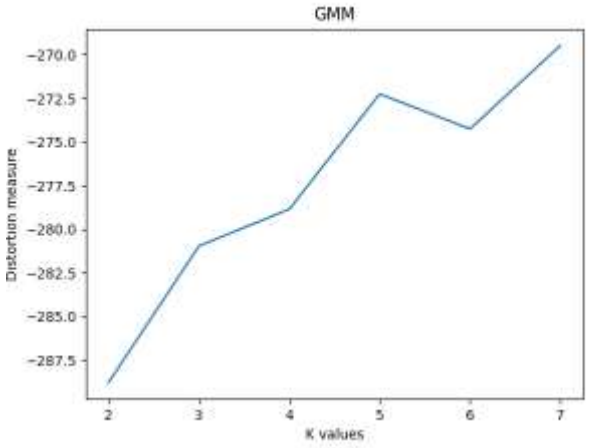

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The magnitude of distortion measure decreases with increase in K.
2. This happens because as the number of groups increases the square distances decreases because of compactness.
3. The number of optimum clusters should be 3 and 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	0.76
6	0.693
7	0.6

Inferences:

1. The highest purity score is obtained with K = 3.
2. Purity score increases upto the optimal value of K=3 and then decreases because at the optimal value the number of clusters is same as that given in the dataset.
3. There is no such observable relationship but one can deduce that at the elbow the purity score would be highest.

4. The priority score of GMM is greater than that of KNN for the optimal value because of soft clustering which tells us that GMM is more accurate.

6

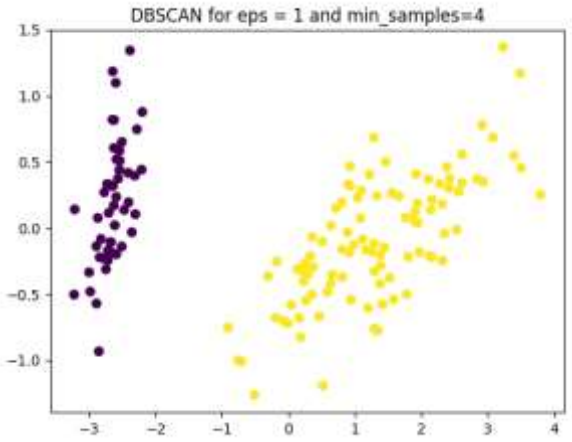

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. There are two clusters formed and the process used in this method is finding the number of connected components based on the radius and min values.
2. The number of clusters formed in this method is less than K-means and GMM

b.

Eps	Min_samples	Purity Score
1	5	0.667
	10	0.667
4	5	0.333
	10	0.333

Inferences:

1. For the same eps value, increasing min_samples doesn't increase or decrease purity score
2. For the same min_samples, increasing eps value decreases purity score