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

Student's Name: Aryan Tiwari **Mobile No: 8982562898**

Roll Number: B20187 **Branch:** Electrical Engineering

1

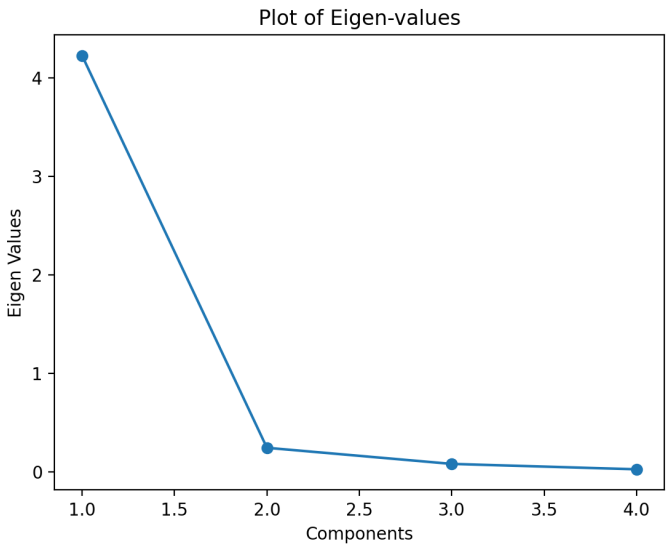


Figure 1 Eigenvalue vs. components

Inferences:

1. he eigenvalue decreases corresponding to each component increase.
2. Eigen values tell the variation along each principal component and the component with highest eigen values carry the most data.

2 a.

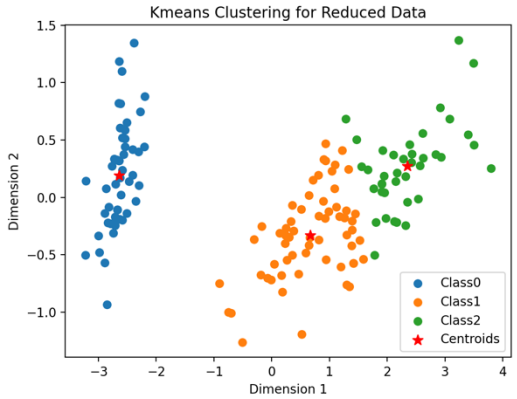


Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. As we specified K=3, the K-means model has given 3 distinct clusters.
2. The boundary seem circular (non linear) around the centroids although there is some distortion

b. The value for distortion measure is 63.87

c. The purity score after examples are assigned to the clusters is 0.89

3

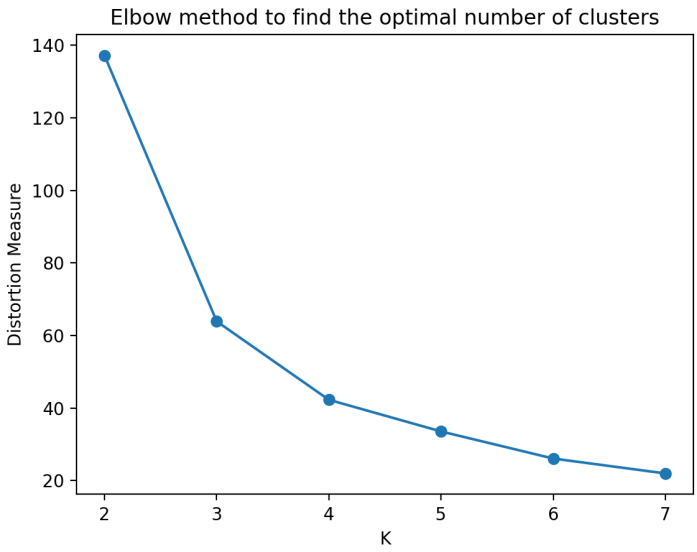


Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion decreases with increasing K.
2. Justify the observed trend.
3. The elbow is formed at K = 3, this shows that 3 is the optimal number of clusters. It follows our data as there were three varieties in the label column.

Table 1 Purity score for K value = 2,3,4,5,6 & 7

K value	Purity score
2	0.667
3	0.886
4	0.686
5	0.667
6	0.513
7	0.526

Inferences:

1. The highest purity score is obtained with K =3.
2. The purity score increase upto k=3, then decreases with increasing k.
3. The data had 3 class originally and so the purity is highest at 3 cluster.
4. The purity is highest at elbow of distortion plot.

4 a.

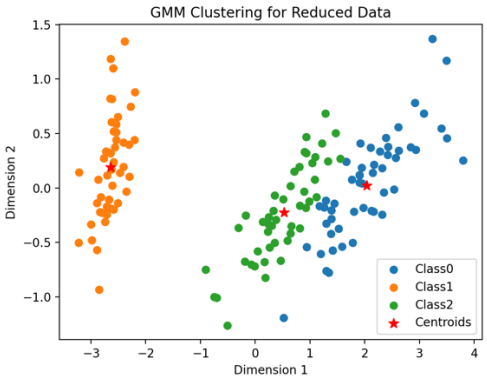


Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. 3 clusters (k=3) are formed by GMM clustering. GMM performs soft clustering as it gives the probability of sample point to belong to a particular cluster.
2. The clusters are elliptical with a little distortion.
3. There is an observable difference in shape and boundary of class 1 and class 2 in K-means and GMM clustering.

b. Total Log Likelihood for K=3 clusters = -280.87.

c. Purity score for data = 0.98

5

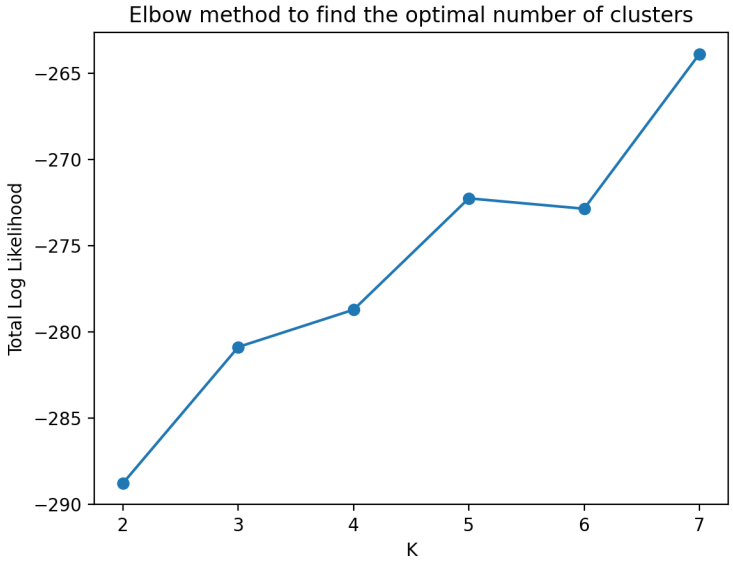


Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure increase with an increase in K.
2. With more clusters, probability of belonging to each cluster increases as

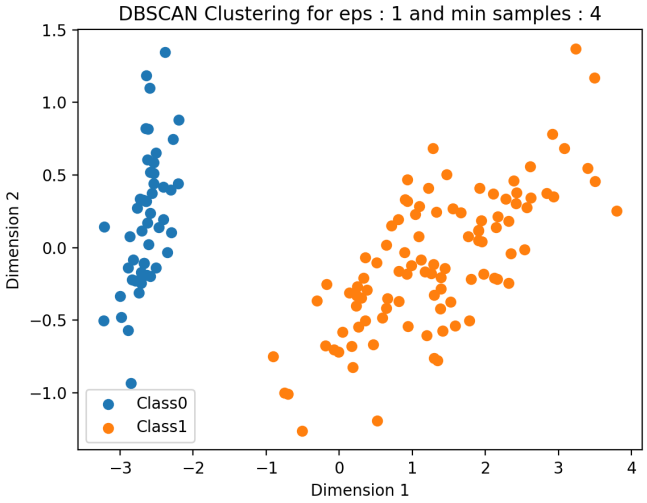
Table 2 Purity score for K value = 2,3,4,5,6 & 7

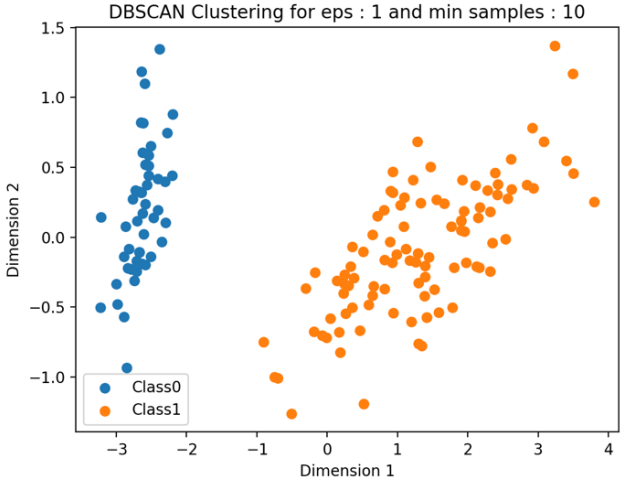
K value	Purity score
2	0.667
3	0.98
4	0.82
5	0.77
6	0.73
7	0.70

Inferences:

1. The highest purity score is obtained with K =3.
2. The purity score increase upto k=3, then decreases with increasing k.
3. The original data had 3 clusters and hence purity is highest at k =3.
4. The maximum purity score for GMM is 0.98 whereas for KMeans is 0.886.

6


DBSCAN Clustering for eps : 1 and min samples : 4
Dimension 2
Dimension 1
Class0
Class1


DBSCAN Clustering for eps : 1 and min samples : 10
Dimension 2
Dimension 1
Class0
Class1

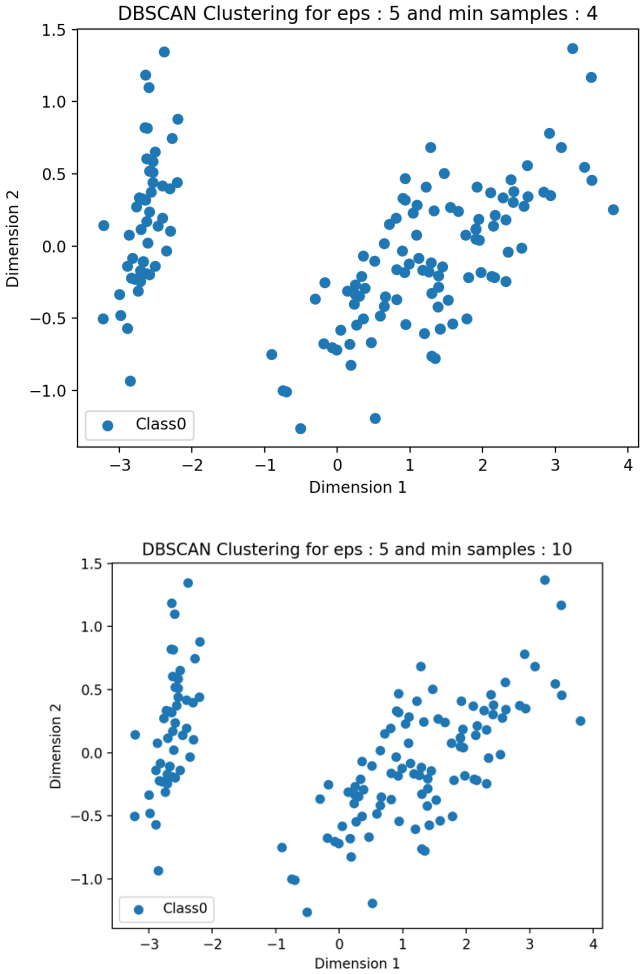


Figure 6 to 9 DBSCAN clustering on Iris flower dataset

Inferences:

1. We cannot specify the number of clusters in DBSCAN. It is robust to outliers and can also form arbitrarily shaped clusters. It is not suitable for too dense or too sparse data with connected boundaries.
2. There are 3 clusters in gmm (as specified), but 2 clusters in eps 1 and only 1 cluster in eps 5.

b.

Eps	Min_samples	Purity Score
1	5	0.667
	10	0.667
4	5	0.33
	10	0.33

Inferences:

1. For the same eps value, the Purity score remains the same on increasing the min samples.
2. For the same min_samples, increasing the value of eps decreases the value of purity score for the sample.