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

Student's Name: Rajeev Kumar **Mobile No: 9341062431**

Roll Number: B20124 **Branch:CSE**

1

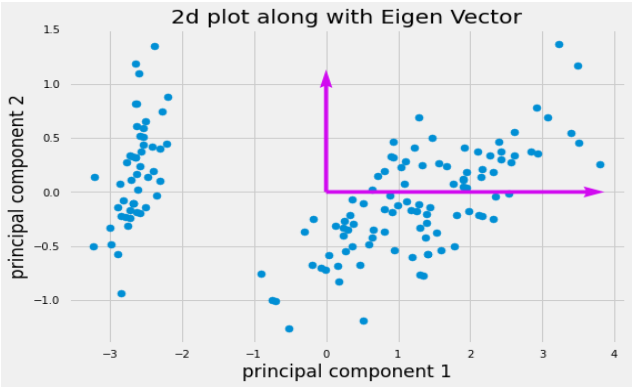

Figure 1 Eigenvalue vs. components

Inference

1. Eigen values are decreasing corresponding to each component increase or decrease successively. Since the eigen values contains the characteristic of whole data then therefore eigen value(principle comp 1) > eigen value(principle comp 2) > eigen value(principle comp 3) >…….> eigen value(principle comp n)

2 a.

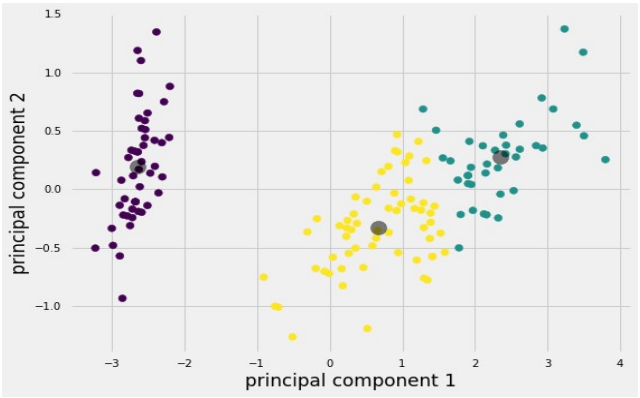

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. From the plot we can see that any point is at minimum distance with respect to allotted centre than the other two centers.
2. No here it does not seem to be. Because it only searches at a perticluar distance over all the direction from the centers. So it may not possible always such that all point lie on circumference.

b. The value for distortion measure is 63.87.
c. The purity score after examples are assigned to the clusters is 0.88.

3

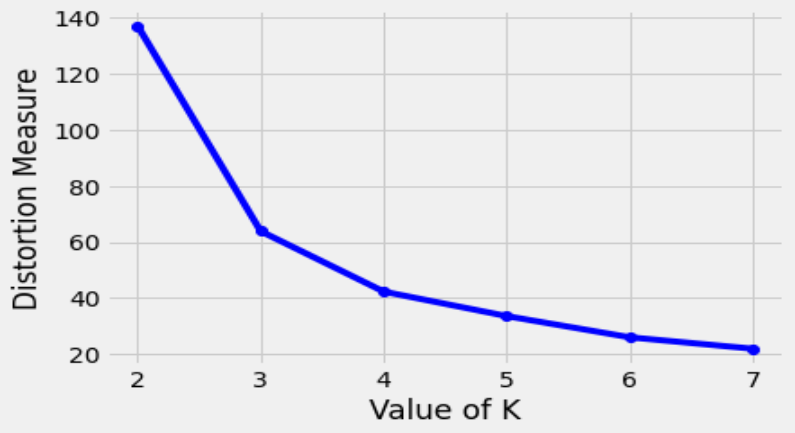

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measure decreases with an increase in K.
2. If the number of cluster increased than it the model will have more number of choices to get fit into.
3. From the plot we can take as K = 3 or 4 for more precise model with help of elbow method.

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

K value	Purity score
2	0.66
3	0.88
4	0.84
5	0.90
6	0.89
7	0.96

Inferences:

1. The highest purity score is obtained with K = 7.
2. Increasing the value of K it increases the purity score.
3. If the number of cluster increased than it the model will have more number of choices to get fit into.
4. Yes as the distortion measure decreases then the purity score increases.

4 a.

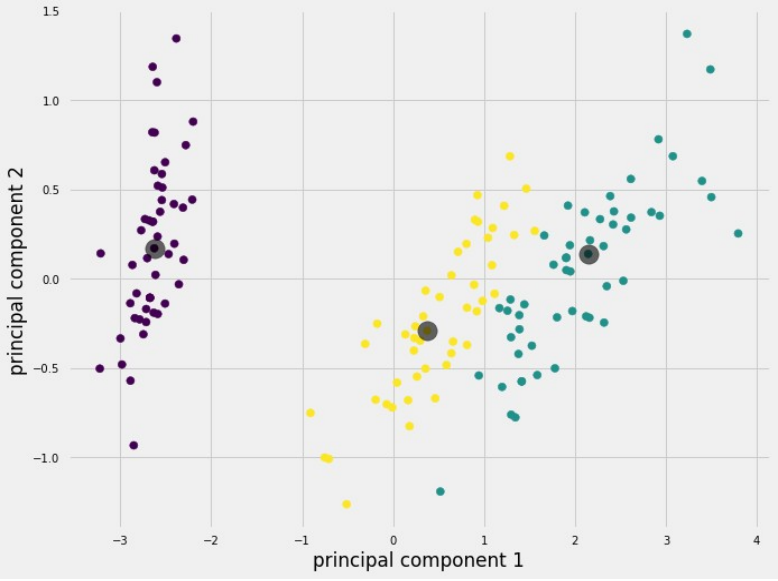

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. From the plot we can see that any point is at minimum distance with respect to allotted centre than the other two centers
2. No here this is not elliptical.GMM only uses elliptical area only to find the number the point lying in a cluster.
3. GMM is much better in prediction than K-Means as purity score of GMM is higher.

b. The value for log likelihood is -0.87.

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

5

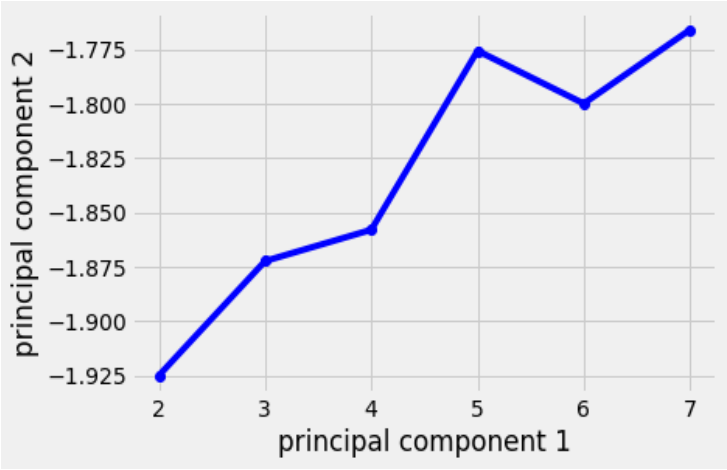

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure decreases with an increase in K.
2. As the more number of cluster is given then it become easier to divide the data.
3. The number of optimum clusters is 3 or 4.Yes since between 3 to 4 it converges as difference become negligible.

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

K value	Purity score
2	0.66
3	0.98
4	0.98
5	0.94
6	0.90
7	0.96

Inferences:

1. The highest purity score is obtained with K =.3 and 4
2. The increament is random on increasing the number of K.
3. As the distortion measure decreases, the purity score increases.
4. Since GMM is soft clustering techniques. Hence GMM has better purity score than K Means.

6

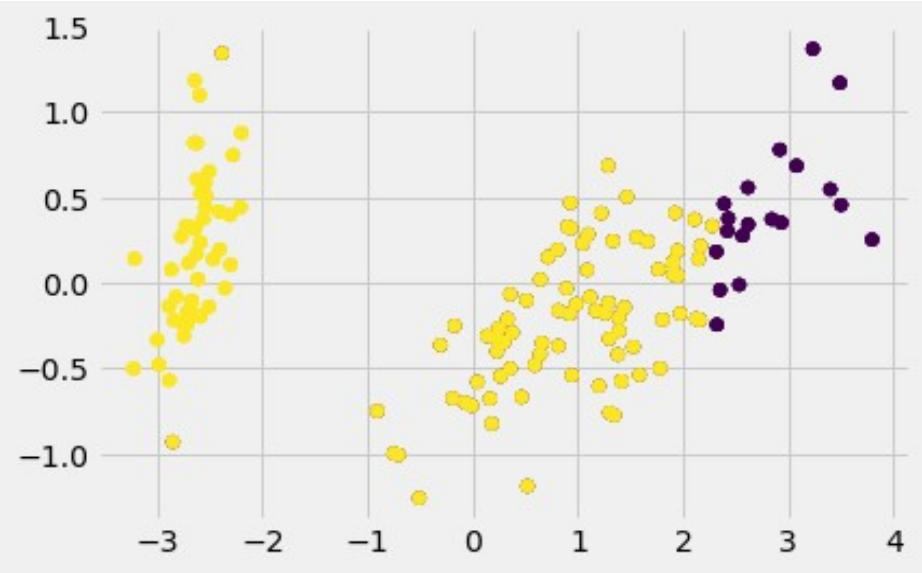

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. DBSCAN only take two cluster to divide the whole set of data.
2. Yes in GMM, we will have to give the prior information about number of clusters while that is not in the case of DBSCAN.

b.

Eps	Min_samples	Purity Score
1	4	0.66
	10	0.66
5	4	0.33
	10	0.33

Inferences:

1. For the same eps value, does increasing min_samples does not really increases purity score.
2. For the same min_samples, increasing eps value the purity score decreases.