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

Student's Name: Ankit Pal Singh **Mobile No: 9149024234**

Roll Number: B20181 **Branch:**EE

1

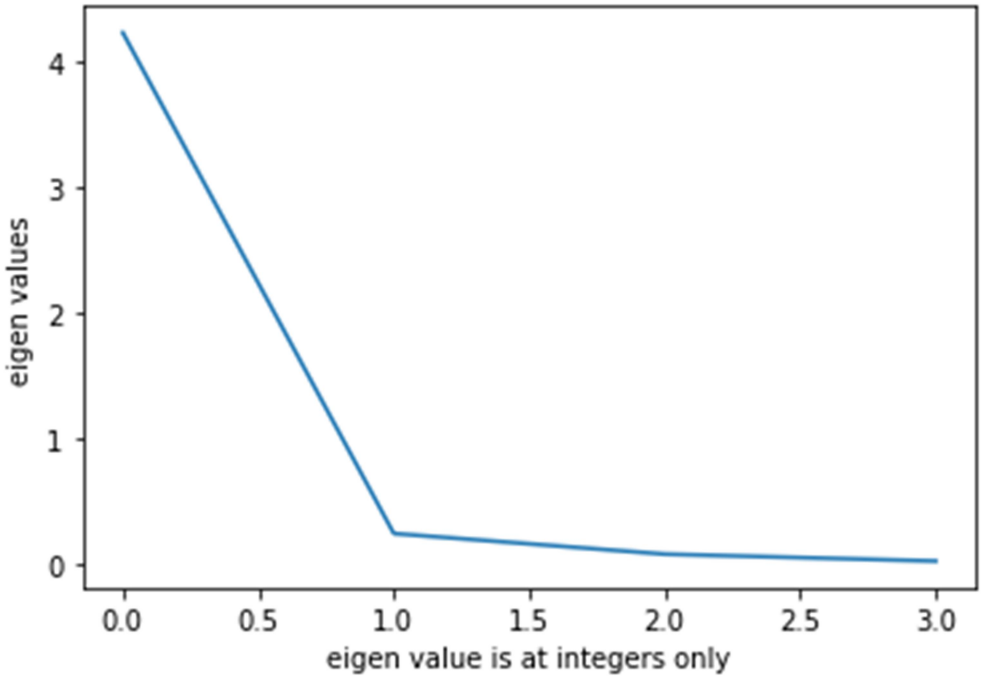

Figure 1 Eigenvalue vs. components

Inferences:

1. As component increases the value of eigen values decreases.
2. Since eigen value corresponding to each PCA component tells that how much variation in data is accounted by that eigen value and PCA is done in such a way that the eigen values corresponding to PCA component 1 Is maximum and decreases further.

2 a.

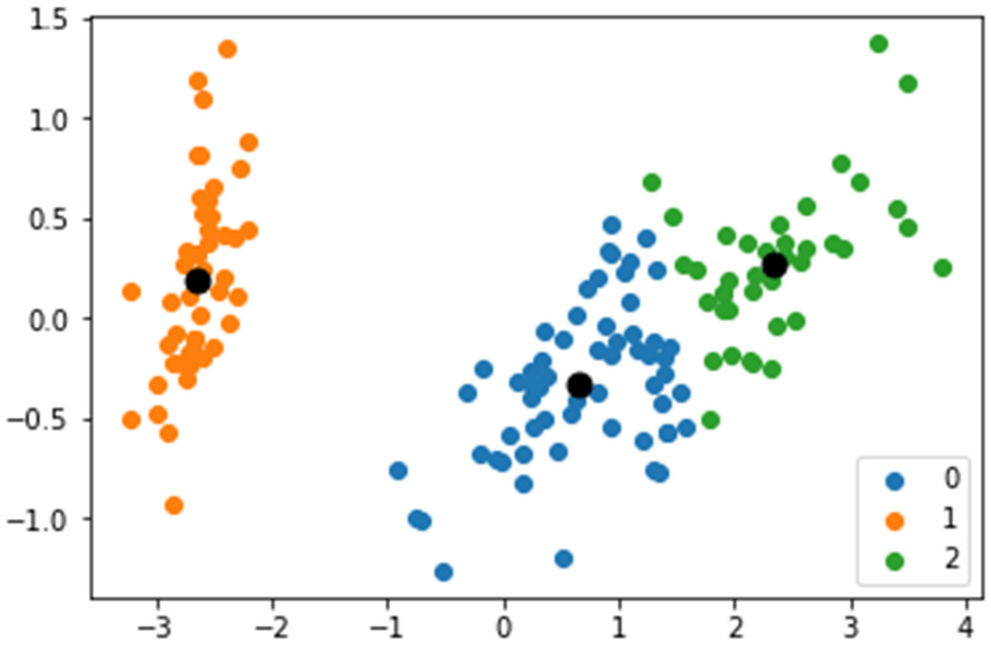

Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. As It can be seen that clusters formed are at considerable distance from each other so k means model predicts very well.
2. Yes , there are circular boundaries for the clusters (since the radius of boundary is the distance between centroid and farthest point in the cluster).

b. The value for distortion measure is 63.87

c. The purity score after examples are assigned to the clusters is 0.88

3

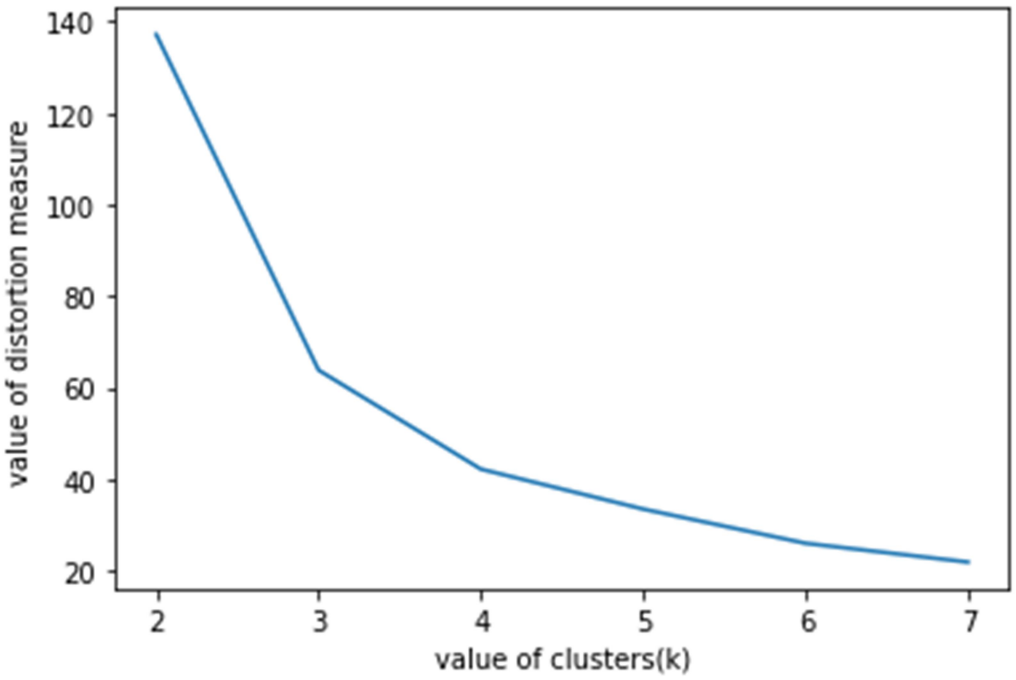

Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. As the value of k increases , the value of distortion measure decreases.
2. As distortion measure is **sum of squared distances of points from cluster centers and** it **Decreases with an increasing number of clusters because as there are more number of clusters then the square of distance decreases as they are not that much apart within that constant space**.
3. The number of clusters should be 3 and elbow and distortion follows the intuition. (as the number of samples in the training data is 150 , 50 for each of three) so 1 cluster for 50 samples.

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

K value	Purity score
2	0.66
3	0.88
4	0.68
5	0.67
6	0.50
7	0.52

Inferences:

1. The highest purity score is obtained with K = 3
2. First upto the optimal value of k purity score increases then It decreases further.
3. If value of k increases above the optimal value of k then purity score value decreases because as there are more number of clusters and then very often the data points falls into wrong clusters.
4. Purity score and distortion measure have positive correlation , as one decreases other also decreases and vice versa.

4 a.

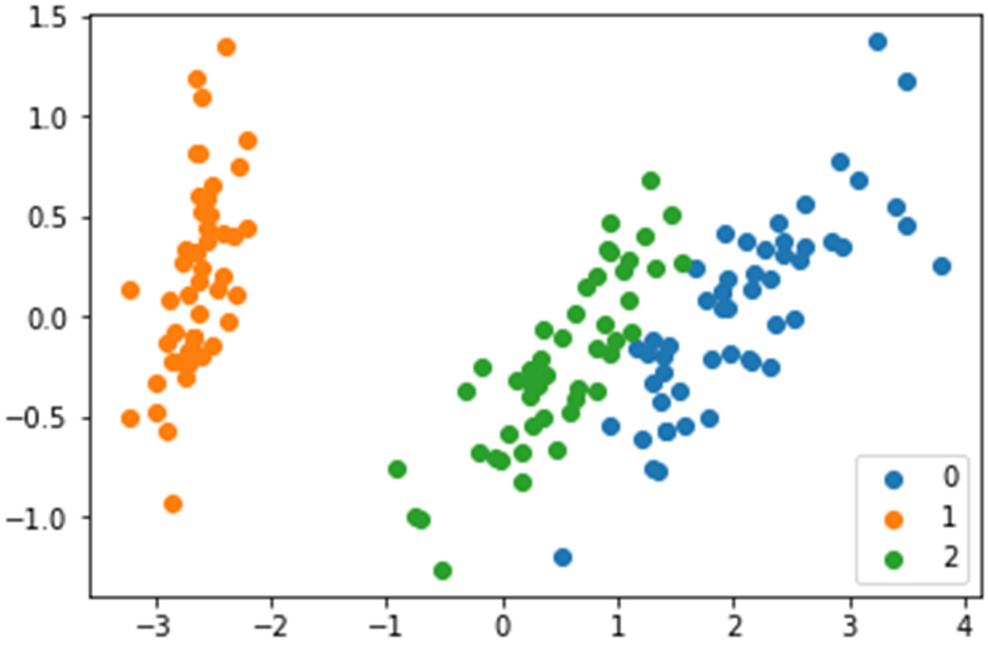

Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. As It can be seen that clusters formed are at considerable distance from each other so gmm model predicts very well .
2. Yes , the boundary seems to be circular .
3. In k means the clusters 0 and 2 are not that much similar to the 0 and 2 clusters in gmm.

.

b. The value for distortion measure is -280.96

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

5

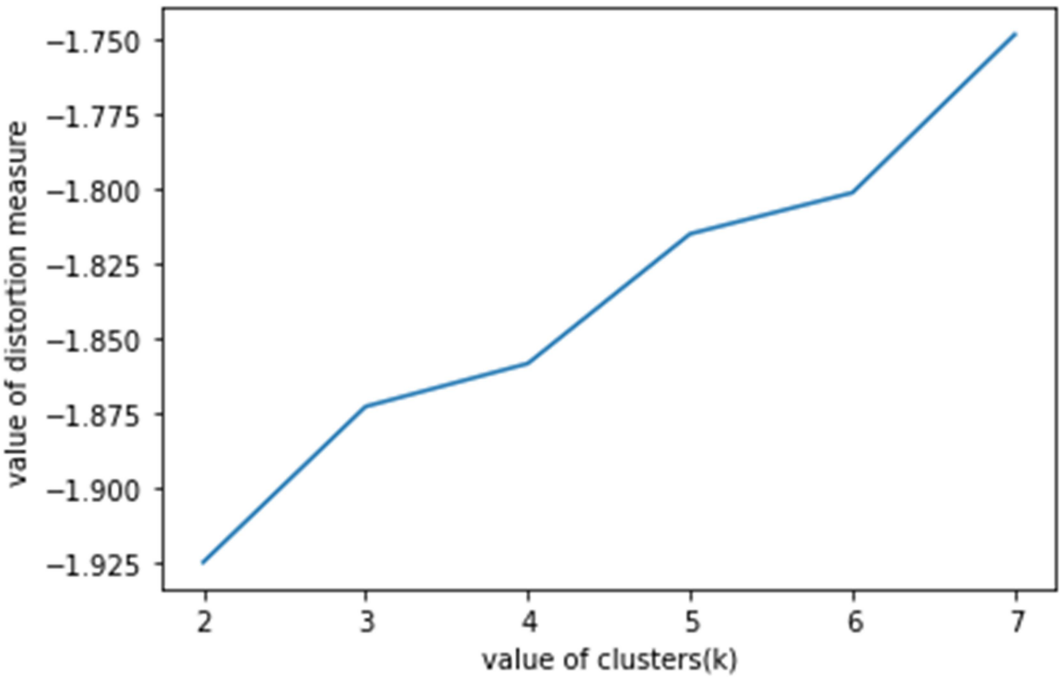

Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. The distortion measure value increases with increase in value of K.
2. As distortion measure is sum of lower bound of log likelihood so with increase in value of k increase more and more so that's why distortion measure value increases.
3. The number of clusters should be 3 and elbow and distortion follows the intuition. (as the number of samples in the training data is 150 , 50 for each of three) so 1 cluster for 50 samples.

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

K value	Purity score
2	0.66
3	0.98
4	0.83

5	0.77
6	0.61
7	0.64

Inferences:

1. The highest purity score is obtained with K =3.
2. First upto the optimal value of k purity score increases then It decreases further.
3. If value of k increases above the optimal value of k then purity score value decreases because as there are more number of clusters and then very often the data points falls into wrong clusters.
4. Purity score and distortion measure have positive correlation , as one decreases other also decreases and vice versa
5. If we consider the optimal number of clusters then GMM is much better than k means for predicting.

6

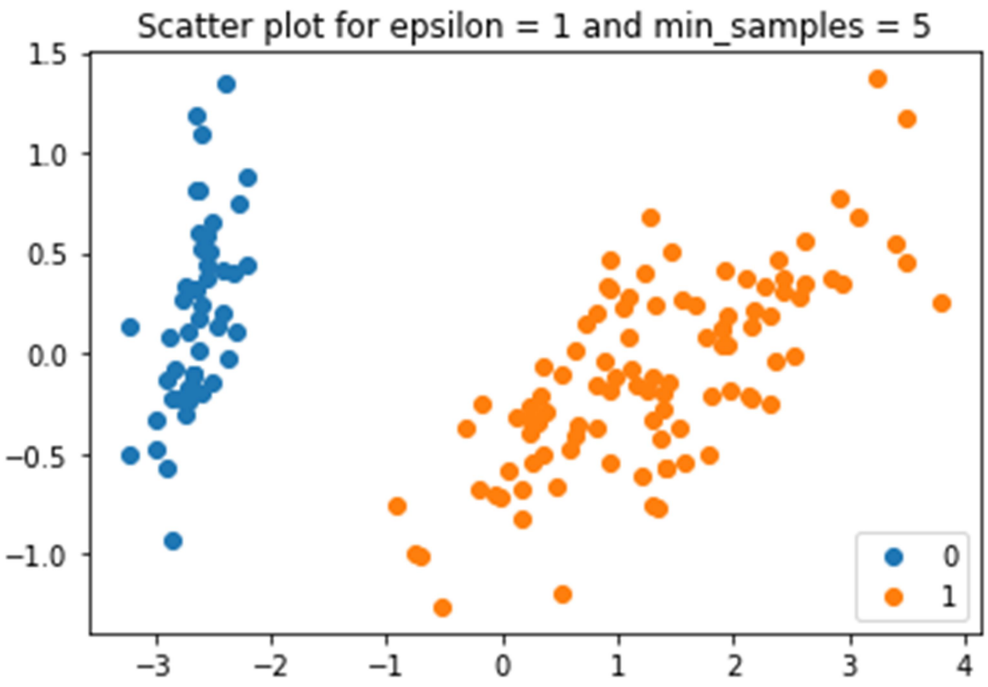

Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Here in DBSCAN clusters form are 2 and we have our data of 50 samples for each class so DBSCAN model can't predict much well.
2. In DBSCAN there are only 2 clusters while in k means and gmm there are 3 clusters available.

b.

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

Inferences:

1. For the same eps value, increasing minimum samples don't have any affect on purity score.
2. For same min_samples , increase in eps value decreases the purity score.