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

Student's Name: Abhay Vijayvargiya **Mobile No: 6377967485**

Roll Number: B20176 **Branch:DSE**

1

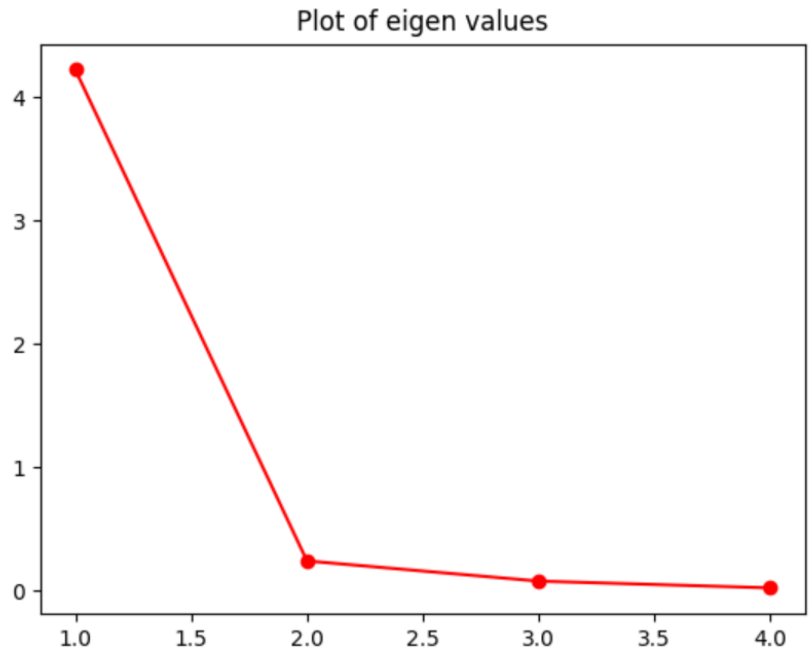


Figure 1 Eigenvalue vs. components

Inferences:

1. Eigen value decreases as the components increases.
2. Magnitude of Eigen value is low if the distribution of the data in the corresponding eigen vector is less.

2 a.

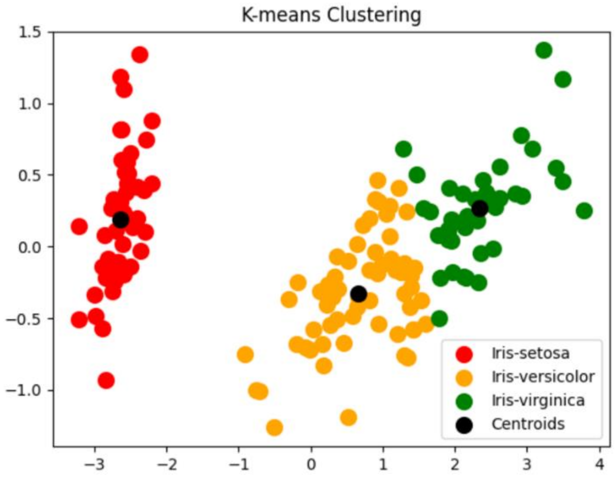


Figure 2 K-means (K=3) clustering on Iris flower dataset

Inferences:

1. K-means algorithm is an iterative algorithm that tries to partition the dataset into K (pre-defined) clusters. It is a distanced-based measurement algorithm. It iteratively measures distance with the cluster center and then modifies the cluster boundary until it doesn't change.
2. From the above plot, the boundary seems to be linear as we can see between Iris-versicolor and Iris-virginica.

b. The value for distortion measure is 63.87

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

3

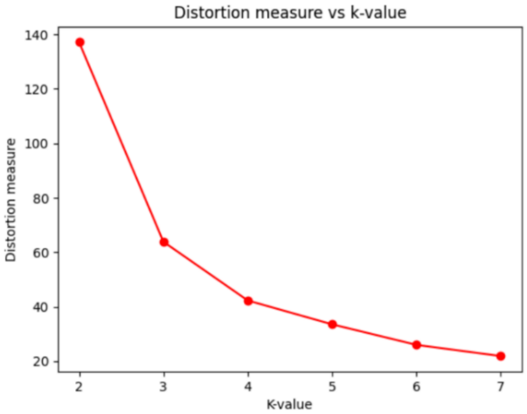


Figure 3 Number of clusters(K) vs. distortion measure

Inferences:

1. Distortion measures decreases with increase in value of K.
2. As the value of K increases, there will be fewer elements in the cluster. So average distortion will decrease. The lesser number of elements means closer to the centroid.
3. By intuition from number of species in the given dataset, the value of K should be 3. Moreover, the elbow method also predicts the optimum value of K as 3.

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

K value	Purity score
2	0.667
3	0.887
4	0.687
5	0.680
6	0.513
7	0.500

Inferences:

1. The highest purity score is obtained with K = 3.
2. Increasing the value of k from 2 to 3 increases the purity score and then further increase in k results in decrease of purity score.
3. Below the optimum value of K, the plot will hide the required information because more clusters are needed. However, increasing above the optimum value of K will cause overfitting. More than required cluster will be there.
4. Purity score is at maximum when the graph takes the shape of the elbow.

4 a.

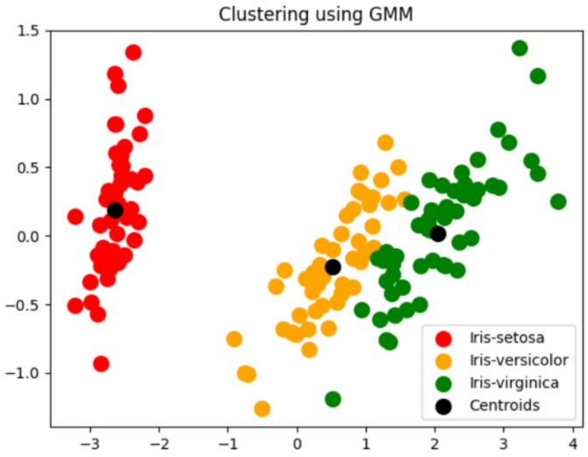


Figure 4 GMM (K=3) clustering on Iris flower dataset

Inferences:

1. GMM unlike the K-means clustering is a distribution-based clustering model. It computes the probability of each data-point belonging to each cluster. It is a soft clustering model.
2. From the output above, the shapes of the clusters are observed to be elliptical.
3. The clusters formed using GMM are elliptical and in K-means those clusters are more circular. K-means uses probabilistic model while k-means uses distance-based model.

b. The value for distortion measure is -280.96

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

5

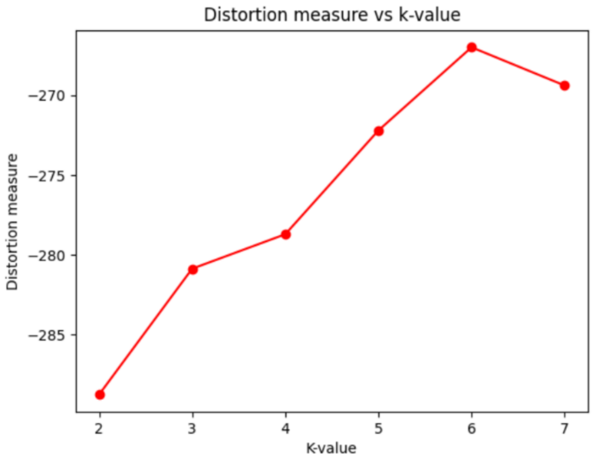


Figure 5 Number of clusters(K) vs. distortion measure

Inferences:

1. Magnitude of Distortion measure increases with increase in value of K.
2. As K increases the there will be more clusters and likelihood of data-point belonging to each cluster will increase.
3. By intuition from number of species in the given dataset, the value of K should be 3. Moreover, the elbow method also predicts the optimum value of K as 3.

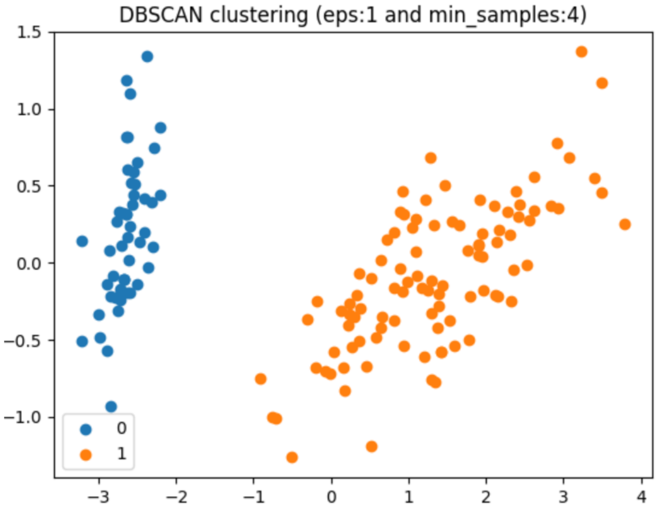
Table 2 Purity score for K value = 2,3,4,5,6 & 7

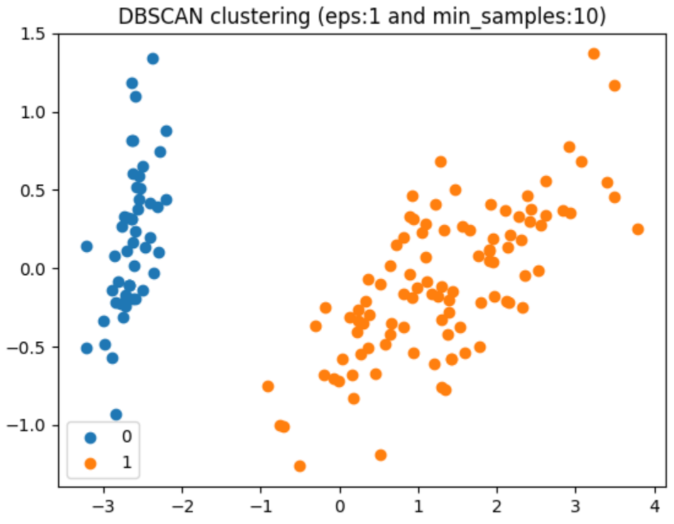
K value	Purity score
2	0.667
3	0.980
4	0.833
5	0.773
6	0.693
7	0.600

Inferences:

1. The highest purity score is obtained with K = 3.
2. Increasing the value of k from 2 to 3 increases the purity score and then further increase in k results in decrease of purity score.
3. Below the optimum value of K, the plot will hide the required information because more clusters are needed. However, increasing above the optimum value of K will cause overfitting. More than required cluster will be there.
4. Purity score is at maximum when the graph takes the shape of the inverted elbow.
5. GMM clustering gives more accuracy than K-means because unlike k-means it uses two parameters for assigning the centroid of the cluster. So, the accuracy increases. Also, k-means is suitable for data having circular clusters but GMM can applied to any distribution. K-means is susceptible to outliers.

6





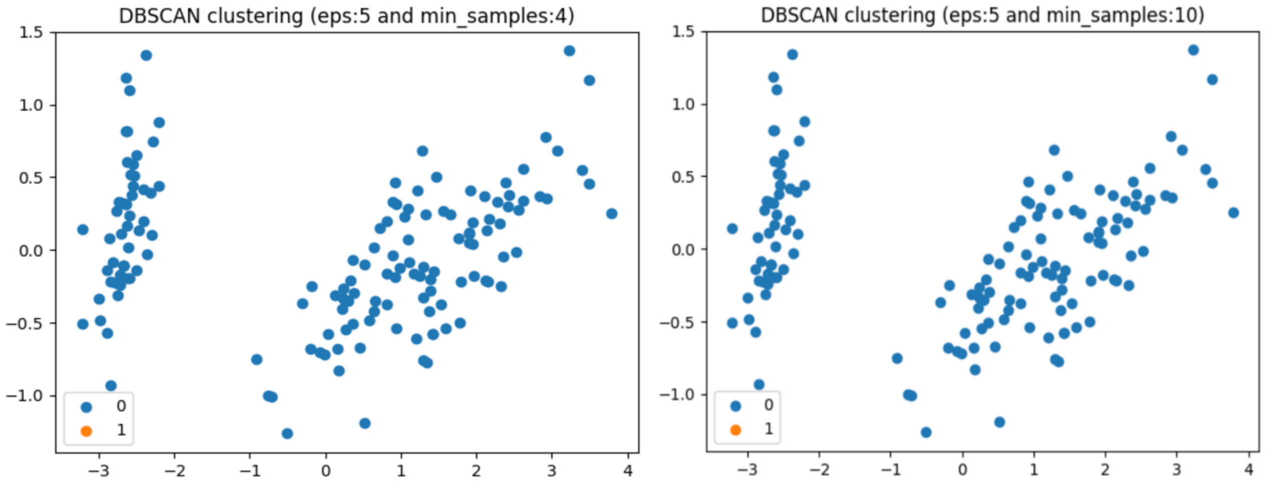


Figure 6 DBSCAN clustering on Iris flower dataset

Inferences:

1. Accuracy of the model is not very good it may be due to our bad choice of eps value.
2. The no. of Clusters is 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	5	0.667
	10	0.667
4	5	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 decreases purity score,