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

Student's Name: Sumit Kumar **Mobile No: 7549233722**

Roll Number: B19118 **Branch:CSE**

1 a.

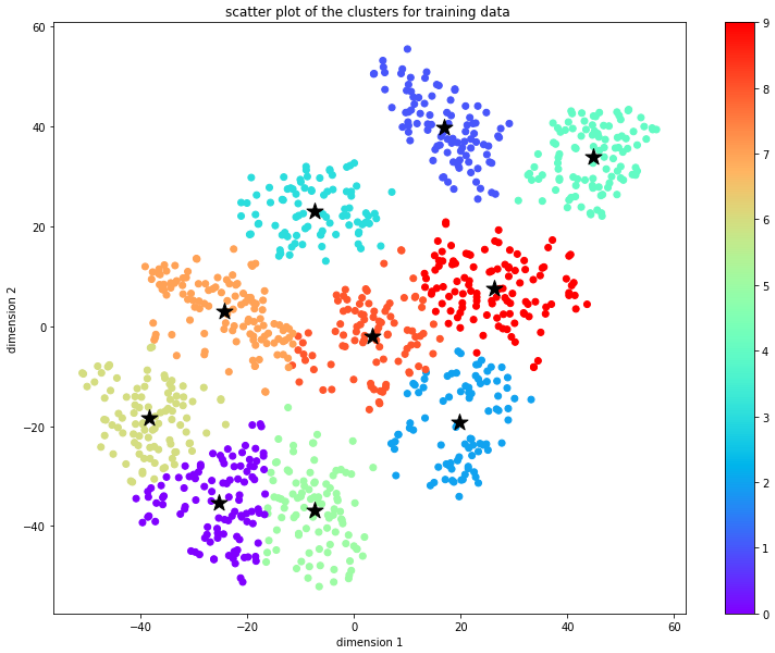


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. In K means clustering, the value of means are updated after each iterations until it converges. It is a hard clustering method because each example must belong to a particular class. Here Euclidean distance is used for measuring the dissimilarity.
2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, the boundary seems to be circular for most of the clusters.

b.

The purity score after training examples are assigned to the clusters is 0.691

c.

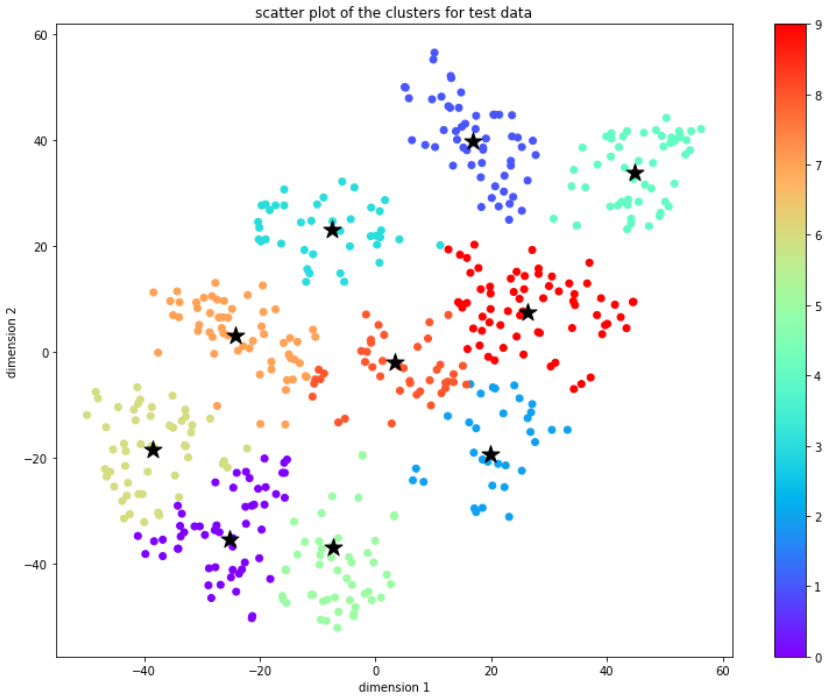


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. The clusters in the training data is very dense as compared to the test data while the test data is sparse. The cluster boundaries are circular for most of the clusters. Form the scatter graph, it seems that the class of the test data is correctly predicted.

d.

The purity score after test examples are assigned to the clusters is 0.678

Inferences:

1. The purity score of the train data is higher than the test data although it is not much higher. The K means model is built on train data so the cluster means are that of the clusters obtained from the train data. The test data is nearly same as that of the train data that is why purity score is quite similar.

2. In k means we have to specify the value of K. This method is very sensitive to the outliers and it is hard clustering method. It also assumes that each clusters has roughly equal number of observations and we are dealing with spherical clusters.

2 a.

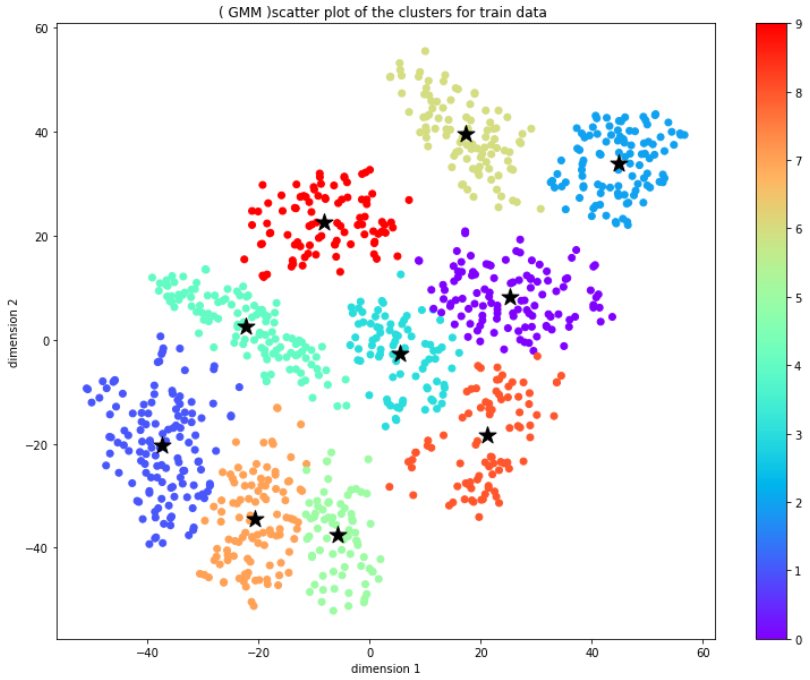


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. GMM is a soft clustering method because it uses probability to predict the class of a given test case. Here also the mean and the convergence matrices get updated after every iterations. Each cluster assumes the Gaussian distribution model.
2. GMM algorithm constraints cluster boundaries to be elliptical in 2D. For most of the clusters it seems elliptical but if we increase the number of training examples then it will be more clear.
3. The clusters formed by K means and GMM methods are similar only difference is in their boundary. K means is of spherical boundary whereas GMM is of elliptical boundary.

b.

The purity score after training examples are assigned to the clusters is 0.713

c.

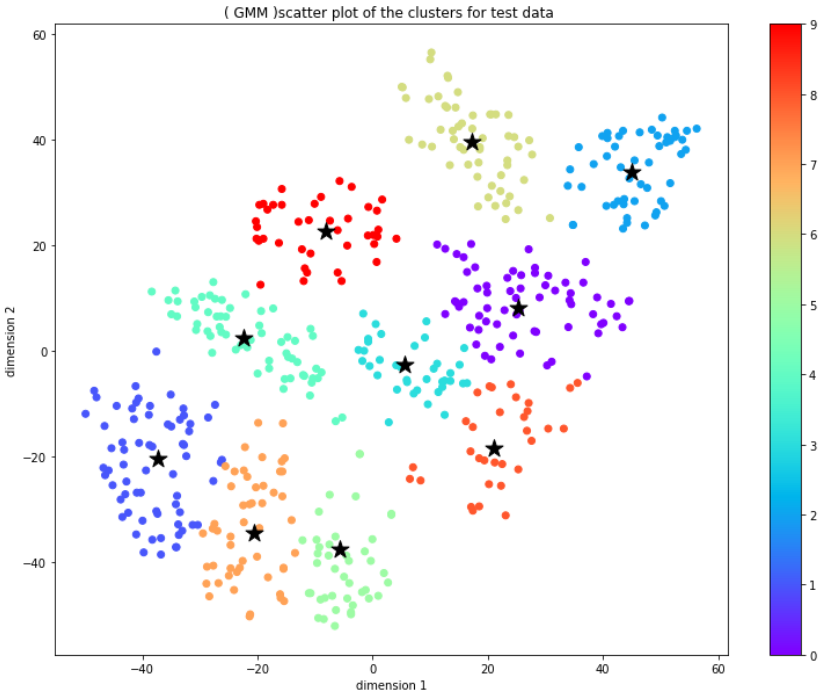


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. Test data and train data clusters are almost similar except the facts that the test clusters are sparse, the cluster mean is not same as mean of train data cluster and the clusters are elliptical for most of the cases.

d.

The purity score after test examples are assigned to the clusters is 0.7

Inferences:

1. Train purity score is higher than test purity score. This is because the model is based on training examples or is learned from training examples but test data points are just assigned classes on the basis of this model.

2. The major limitations are that GMM is computationally expensive for data with higher dimensions. It fails when the covariance matrix of a cluster is singular. Also, the number of clusters is chosen manually.

3 a.

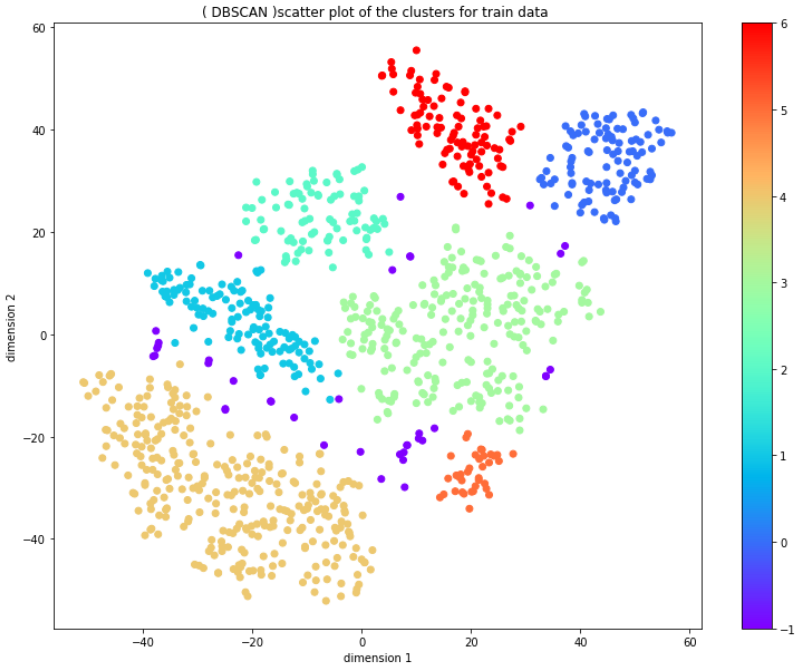


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. Clusters are made on the basis of density of the data. There is no specific shape for clusters. The clusters are robust to outliers. The shapes of the cluster boundary is arbitrary.
2. In K means and GMM, there is a specific shape of the cluster boundary but in DBSCAN the cluster boundary is arbitrary in shape. This clustering method is done on the basis of density of the data points also we do not have to specify the number of clusters in this method.

b.

The purity score after training examples are assigned to the clusters is 0.585

c.

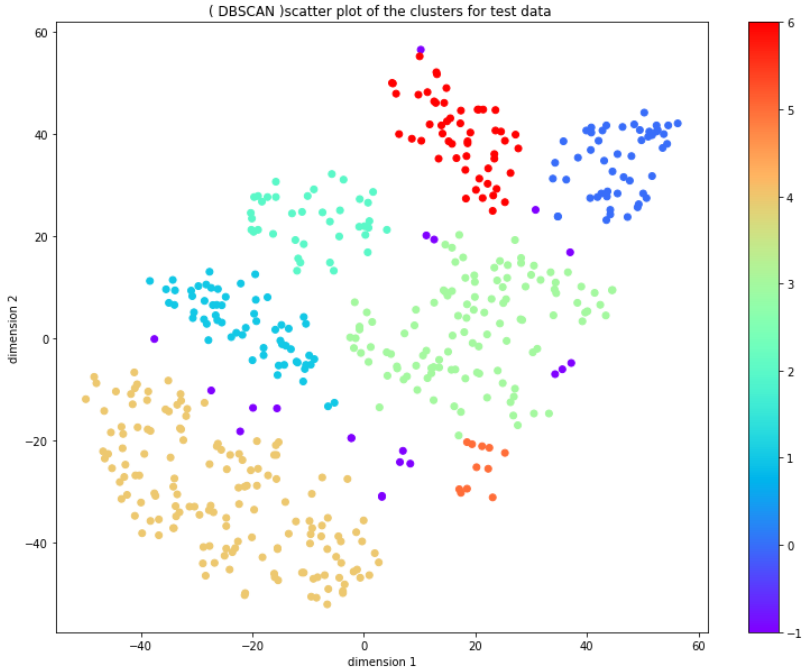


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. The clusters in the training data is very dense as compared to the test data while the test data is sparse. The cluster boundaries are arbitrary in shape. Form the scatter graph, it seems that the class of the test data is correctly predicted.

d.

The purity score after test examples are assigned to the clusters is 0.584

Inferences:

1. The purity score of the train data is very slightly greater than the test data, this is because we fit the training data into the model and predicting the class for train data and test data. The clusters parameters are that of the training data and therefore the purity score is almost similar.
2. DBSCAN uses the density of the data points as the classification algorithm. So, if the data is very danse and there is no any low dense data to separate then this model is not suitable.

Bonus Questions :

K Means Method :

Table of Purity Score on training data and test data for K means method

Value of K	Purity Score on training data	Purity Score on test data
2	0.2	0.2
5	0.392	0.402
8	0.63	0.624
12	0.632	0.628
18	0.468	0.452
20	0.416	0.416

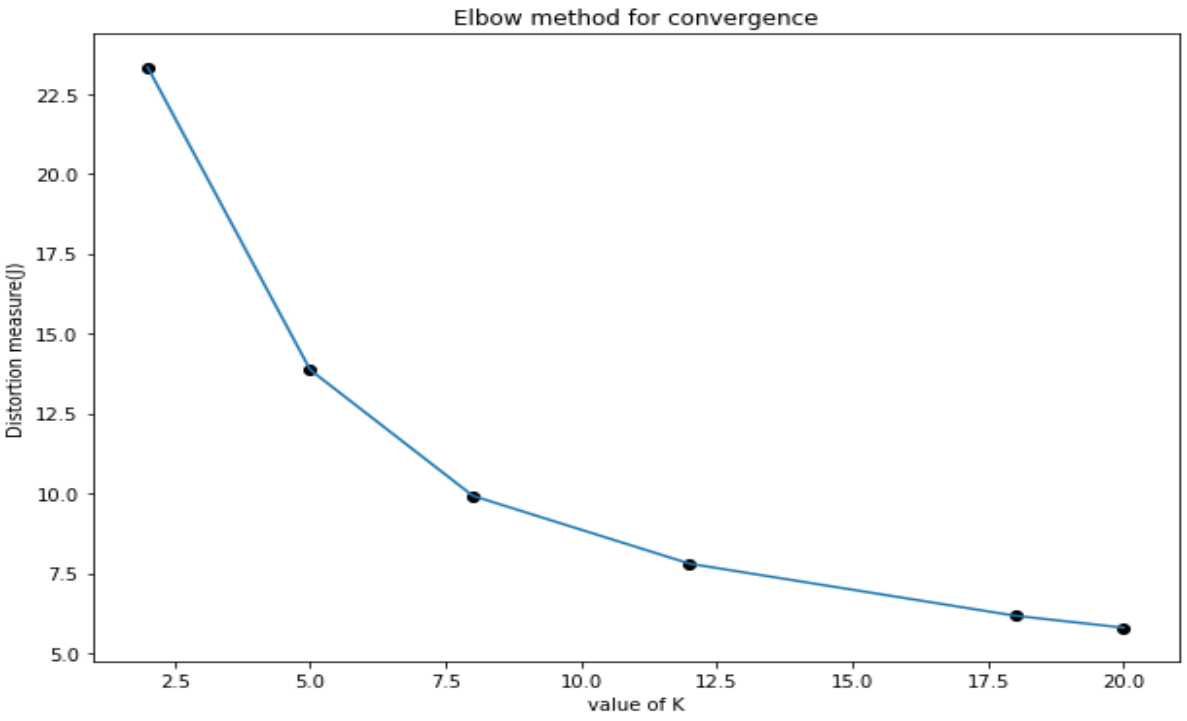


Figure 7 elbow method in K means method

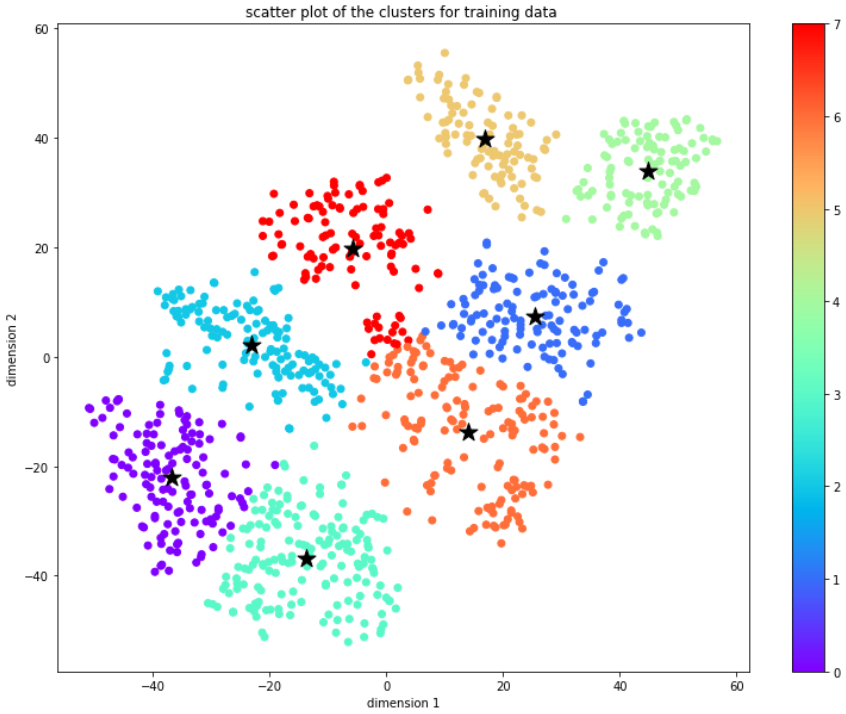


Figure 8 scatter plot of cluster for K means (K = 8) on training data

Inferences:

1. From the elbow method we find that the value of K which is most suitable in this method is for K = 8. This is because after that value the graph of the distortion vs K is linear. This is the convergence criteria for the elbow method.
2. The purity score for K = 8 on training data is 0.63 and the purity score on the test data is 0.624.
3. It can also be observed that for K = 8, the purity score is maximum.

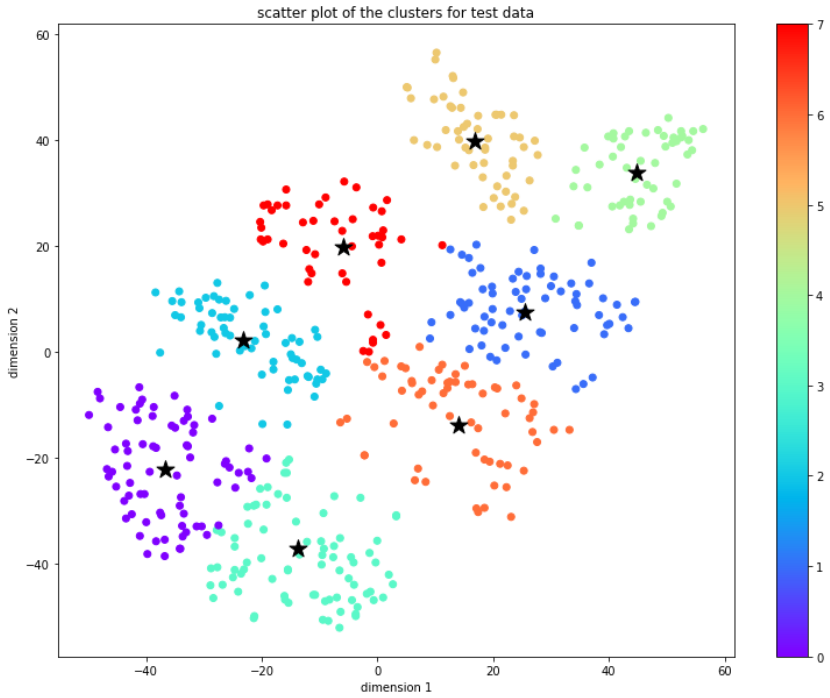


Figure 9 scatter plot of cluster for K means (K = 8) on test data

Inferences:

1. The purity score on the test data for K = 8 is 0.624
2. The clustering of the test data is similar to that of the train data but it is less dense due to the less number of examples. The cluster boundary is spherical in shape for most of the clusters.

GMM method:

Table of Purity Score on training data and test data for GMM method

Value of K	Purity Score on Training data	Purity score on Test data
2	0.2	0.2
5	0.47	0.466
8	0.627	0.638
12	0.567	0.576
18	0.506	0.482
20	0.426	0.42

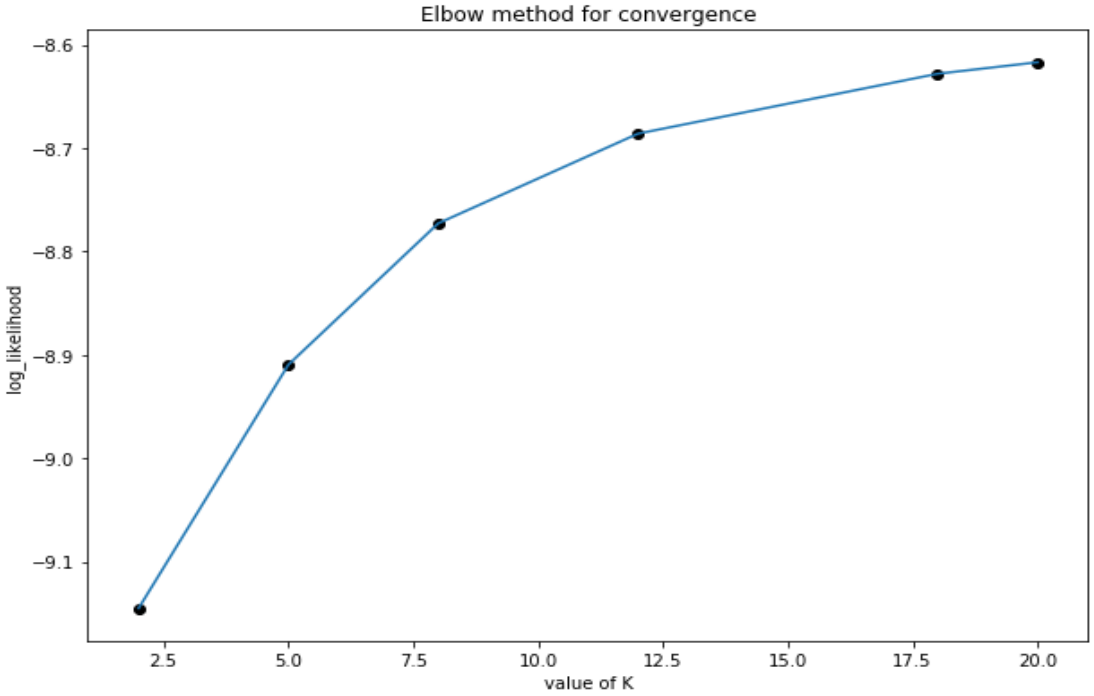
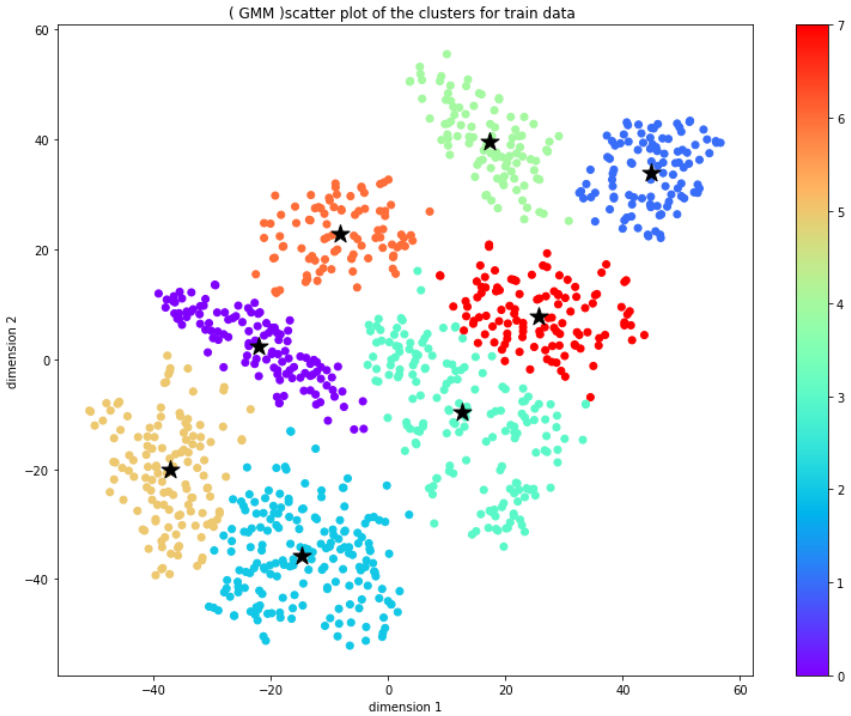


Figure 10 Elbow method on GMM



Inferences:

1. From the elbow method we find that the value of K which is most suitable in this method is for K = 8. This is because after that value the graph of the log likelihood vs K is linear. This is the convergence criteria for the elbow method.
2. The purity score for K = 8 on training data is 0.627 and the purity score on the test data is 0.638.
3. It can also be observed that for K = 8, the purity score is maximum.

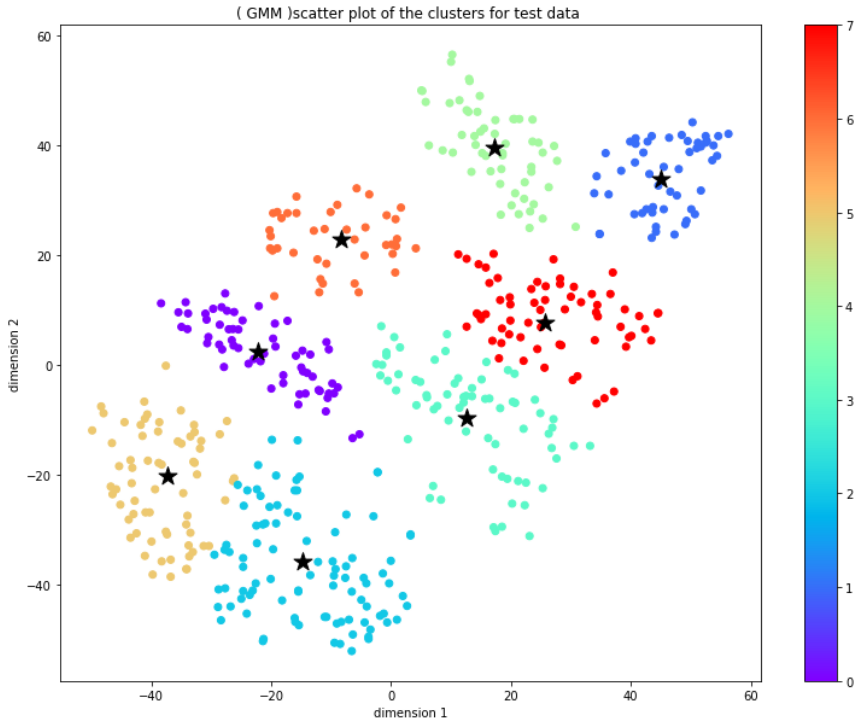


Figure 11 scatter plot of cluster for GMM (K = 8) on test– data

Inferences:

1. The purity score on the test data for K = 8 is 0.638
2. The clustering of the test data is similar to that of the train data but it is less dense due to the less number of examples. The cluster boundary is elliptical in shape for most of the clusters.

DBSCAN Method :

Table of Purity score on training and test set for DBSCAN method

Value of epsilon	Min - samples	Purity score on training data	Purity score on test data
1	10	0.1	0.1
1	30	0.1	0.1
1	50	0.1	0.1
5	1	0.208	0.212
5	10	0.585	0.584
5	30	0.158	0.14
5	50	0.1	0.1
10	1	0.1	0.1
10	10	0.1	0.1
10	30	0.1	0.1
10	50	0.503	0.5

Inferences:

1. The best purity score is for epsilon = 5 and the min-samples = 10
2. Finding the value of epsilon and min – samples is an experimental process. We can the value of epsilon and min-samples and observe that what values works well on the given data.