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

Student's Name: Aryan Garg **Mobile No: 8219383122**

Roll Number: B19153 **Branch: EE**

1 a.

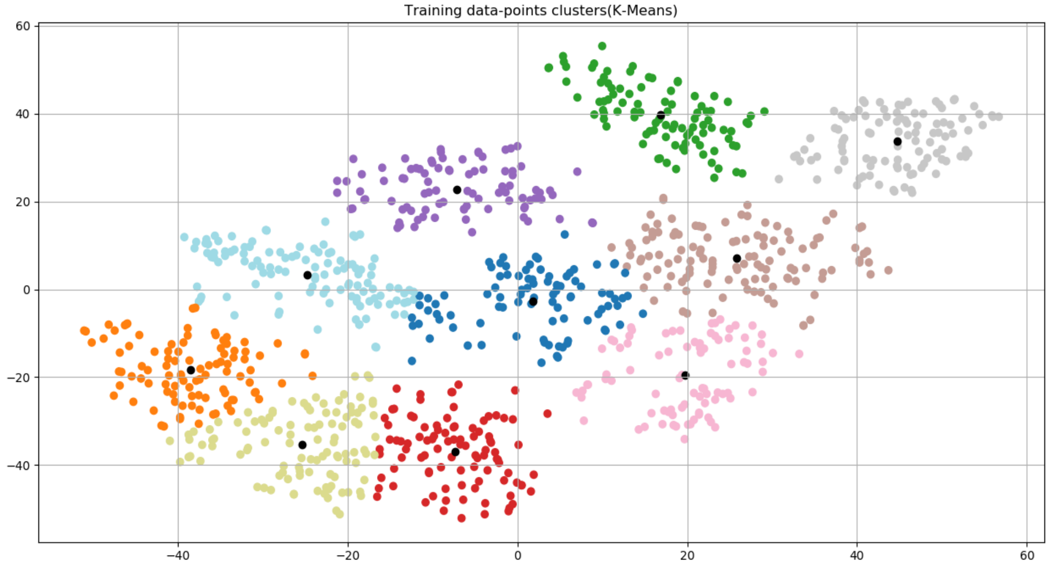

Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. K-Means guarantees us convergence and easily generalizes complex clusters; however, we are at some loss as we have to initialize the number of clusters and the time complexity of the algorithm is too high for practical purposes. Overall, K-Means usually performs well.
2. The boundaries for most clusters appear to be elliptical.

b.

The purity score after training examples are assigned to the clusters is **0.689**

c.

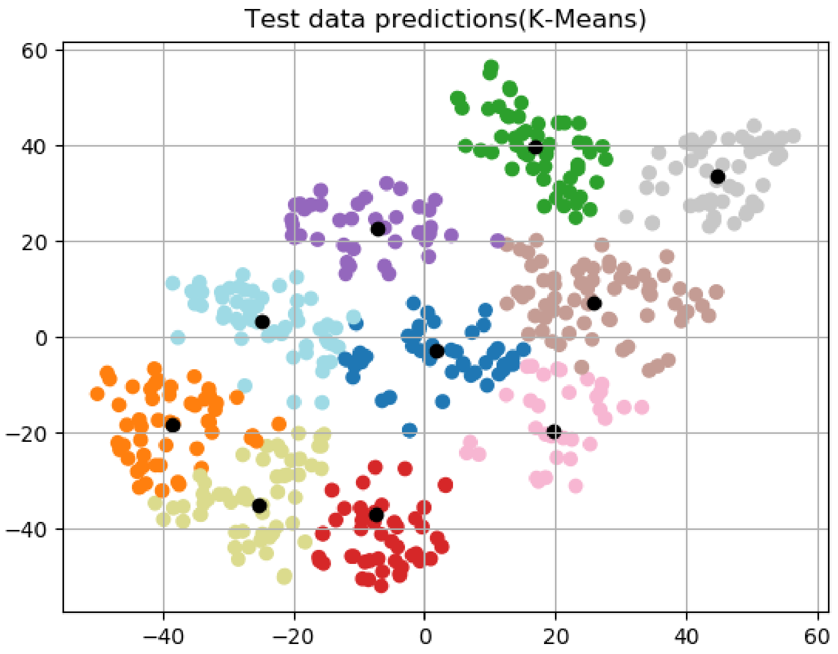

Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. Test data clusters appear to be more circular and closer than the training data. There appears to be the chance of miss-classifying some of the test-points very easily.

d.

The purity score after test examples are assigned to the clusters is **0.676.**

Inferences:

1. The purity scores of training and testing (training >~ testing) are almost similar.
2. K-Means uses lazy propagation and takes a lot of time to run at execution. Poor time complexity.

2 a.

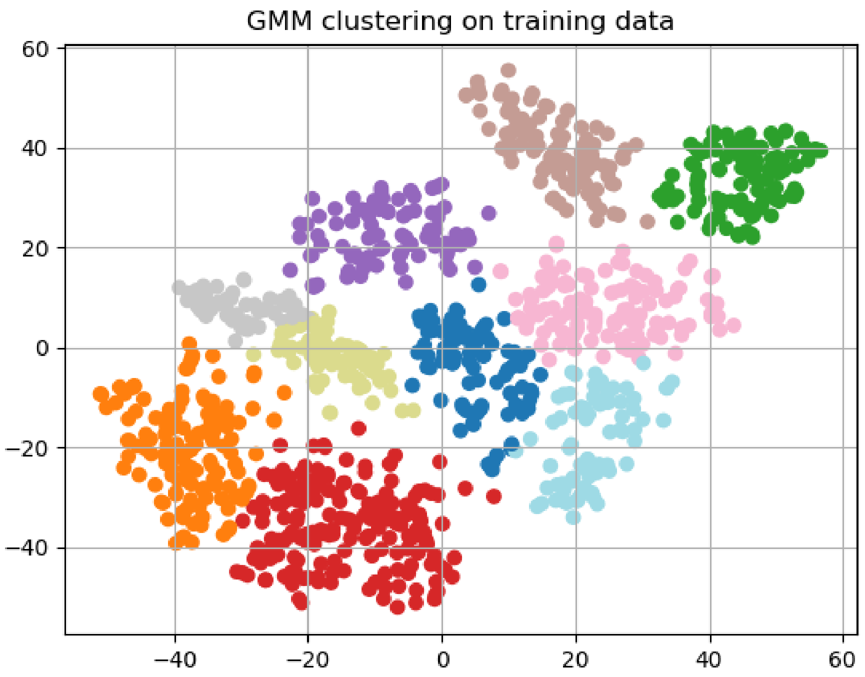

Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. Just because GMM doesn't assume data to be circular, this algorithm takes precedence over the former. It uses mean and variance to find the centroids and the respective clusters.
2. It is evident from the plot that most clusters are elliptical in nature.
3. There is a stark difference in 1a and 2a as some clusters have been merged and some unmerged from 1a.

b.

The purity score after training examples are assigned to the clusters is **0.635**

c.

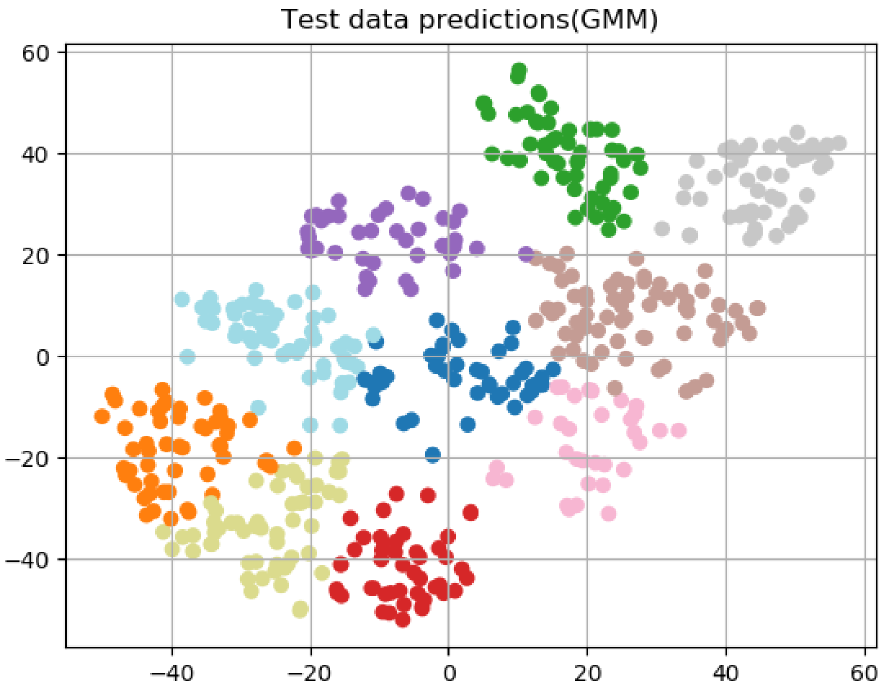

Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. Training clusters are vastly different from the test clusters.

d.

The purity score after test examples are assigned to the clusters is **0.638**

Inferences:

1. Test purity is higher in magnitude.
2. Construction of this model and expectation maximization is not easy. This is a big limitation of this model.

3 a.

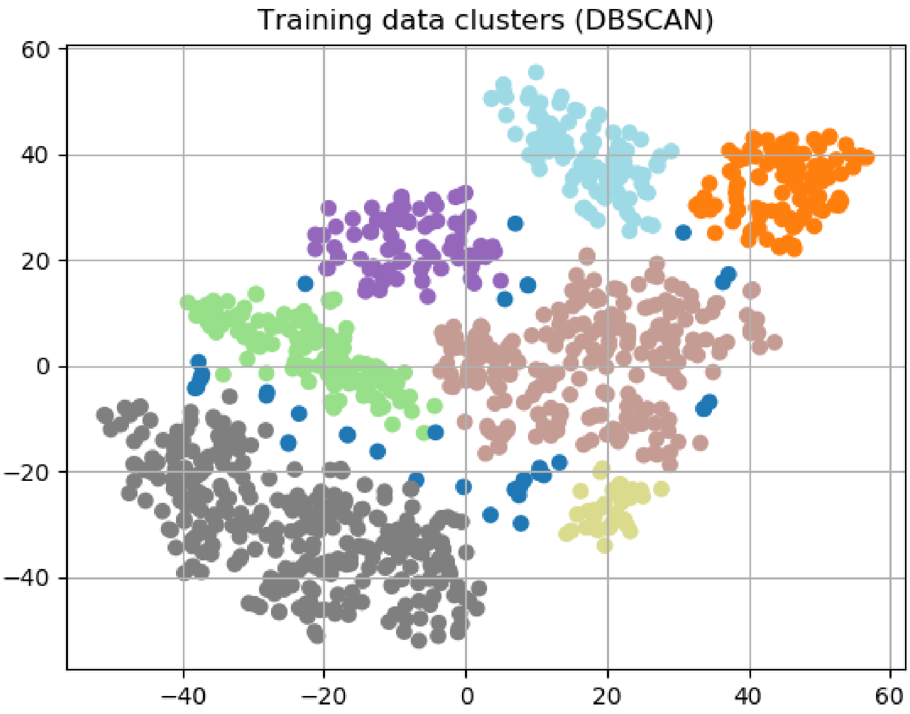

Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. DBSCAN is very good at identifying the number of clusters using the math that it uses and is robust to outliers. It is considered to be one of the finest algorithms for clustering in practice till this date as well.
2. DBSCAN's clusters are less in number as compared to K-Means and GMM algorithm's outputs.

b.

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

c.

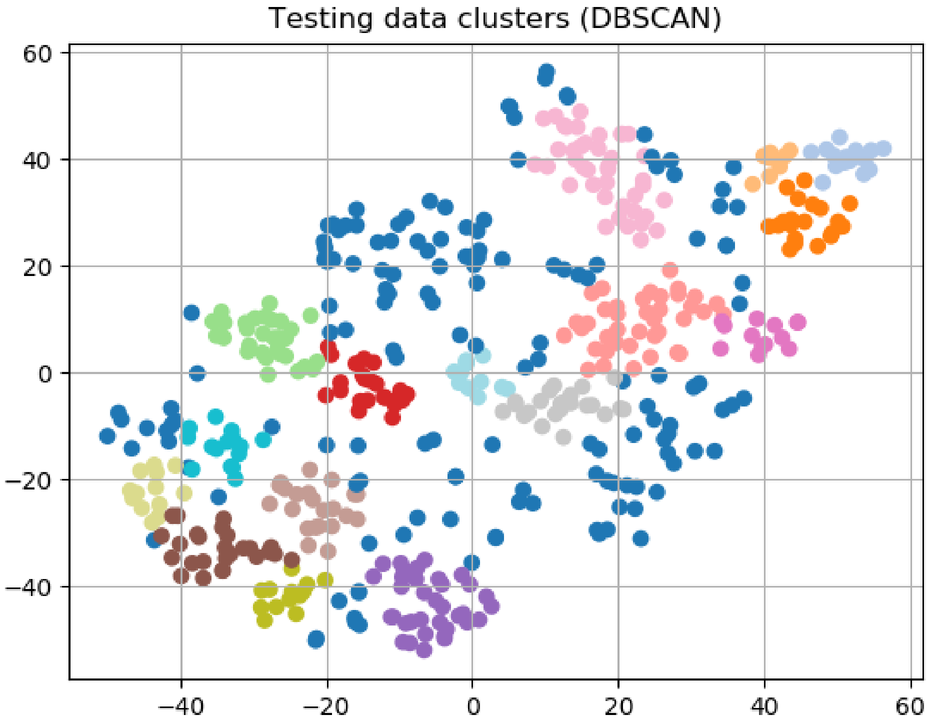

Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. DBSCAN does a poor job in making clusters for test data as we have certainly taken the radius for identifying outliers from border-points as way smaller than it should have been. Also the test data is very close for some clusters and hence leads to bad outcome.

d.

The purity score after test examples are assigned to the clusters is **0.484**

Inferences:

1. Training purity is way higher than testing purity due to the intrinsic difference (Euclidean) in the data-sets
2. The only limitation of DBSCAN is choosing the correct radius parameter for good results.