LAB ASSIGNMENT – VII

Harsh Arya (B20043)

1 a.

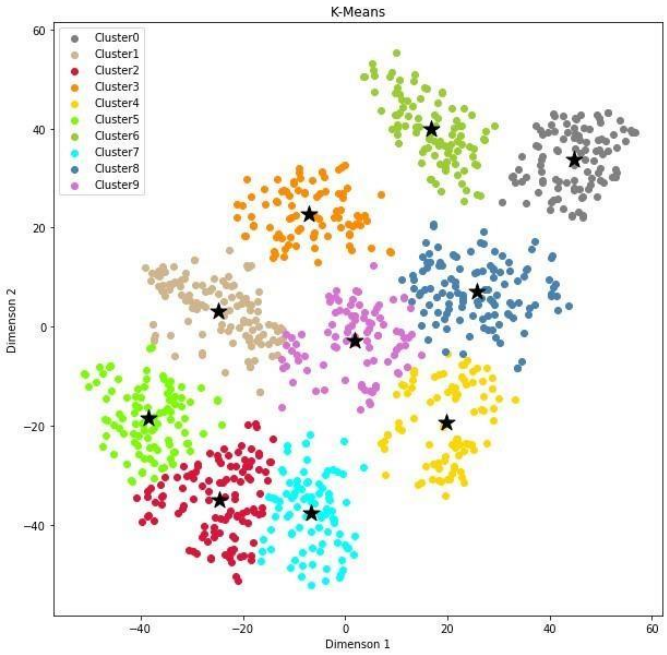

Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. K-Means uses hard clustering. Means are updated after each iteration. The algorithm converges after a certain number of iterations. Boundary between clusters are linear (Euclidean distance is used here).

2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, the boundary seems to be almost circular only for few clusters. **b.**

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

c.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

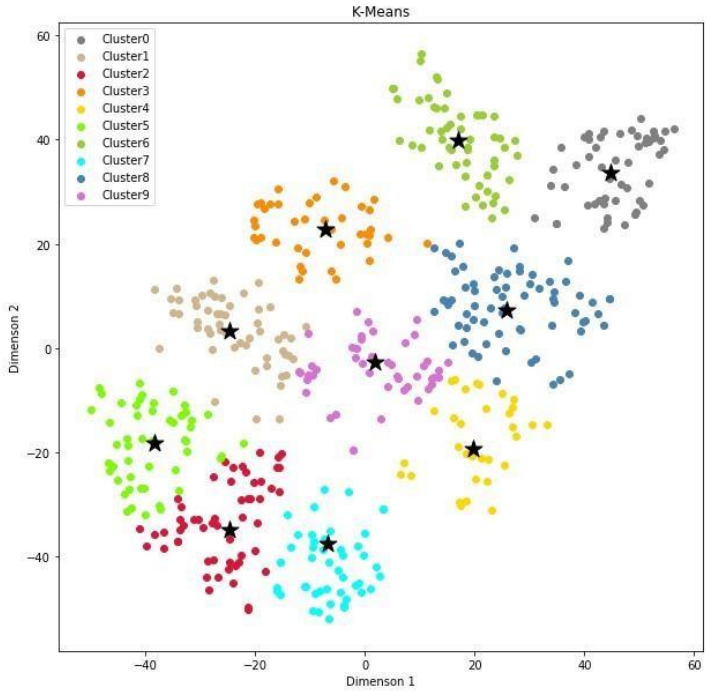

Figure 2 K-means (K=10) clustering on the mnist tsne test data

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 not circular.

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

Inferences:

1. The purity score of train data is slightly higher than that of test data. The K-Means model is built on train data, so the cluster means are that of the clusters obtained from train data. The test data is nearly same as that of train data that is why the purity scores are similar.
2. The user must specify the value of K. K-Means assumes that we deal with spherical clusters and each cluster has roughly equal numbers of observations. K-means clusters depends on initial values. Sensitive to Outliers.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

2 a.

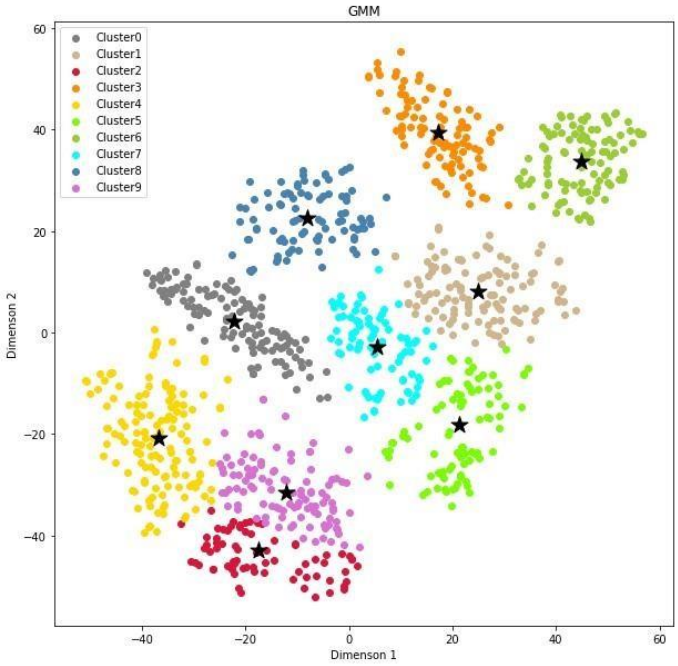

Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. GMM uses soft clustering. Means and Covariance matrices are updated after each iteration. The algorithm converges after a certain number of iterations. Each cluster is represented as a Gaussian distribution. Each cluster has an elliptical shape
2. GMM algorithm constraints cluster boundaries to be elliptical in 2D. From the output, the boundary seems to be almost elliptical for most of the clusters.
3. The cluster boundary in K-Means was spherical whereas in GMM it is elliptical. GMM also considers covariance of data.

b.

The purity score after training examples are assigned to the clusters is 0.712.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

c.

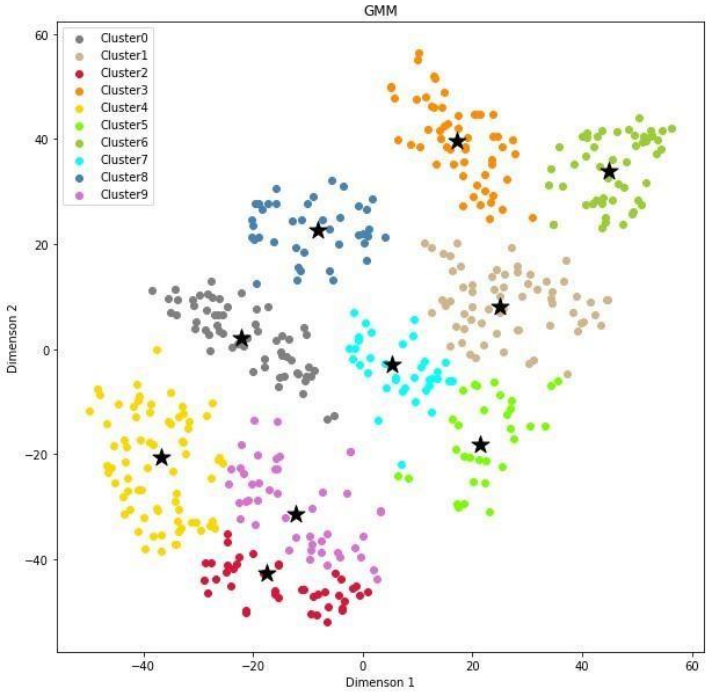

Figure 4 GMM clustering on the mnist tsne test data

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 not elliptical.

d.

The purity score after test examples are assigned to the clusters is 0.688.

Inferences:

1. The purity score of train data is slightly higher than that of test data. The GMM model is built on train data so the cluster parameters are that of the clusters obtained from train data. The test data is nearly same as that of train data that is why the purity scores are similar.
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.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

3 a.

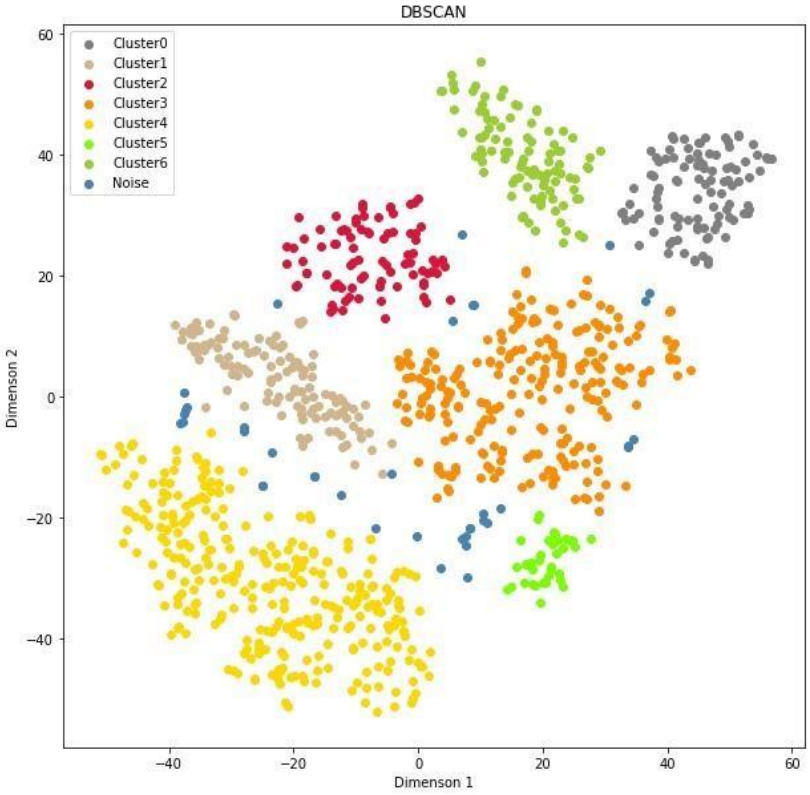

Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. Clusters are made based on density of the data. There is no specific shape for clusters. The clusters are robust to outliers.
2. Unlike K-means and GMM, DBSCAN clusters doesn't have any specific cluster shapes. The clustering is done based on density of data points. Also, noises are not assigned to clusters unlike K-Means and GMM. The number of clusters are not manually set. **b.**

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

c.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

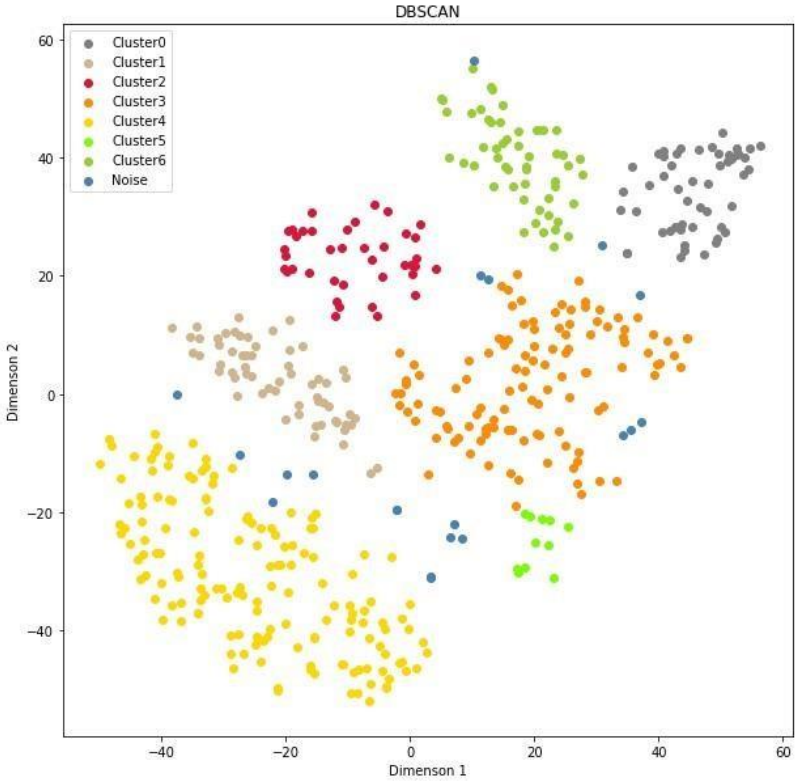

Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. Test data and train data clusters are almost similar except the facts that the test clusters are sparse.

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

Inferences:

1. The purity score of train data is slightly higher than that of test data. The DBSCAN is built on train data so the cluster parameters are that of the clusters obtained from train data. The test data is nearly same as that of train data that is why the purity scores are similar.
2. DBSCAN is not suitable when the data is completely dense and there is no low dense area to separate.

Bonus Questions: A.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

K-Means:

K	Purity Score for Train Data	Purity Score for Test Data
2	0.2	0.2
5	0.392	0.402
8	0.63	0.624
12	0.61	0.62
18	0.478	0.458
20	0.453	0436

Table 1: Purity Score for K-Means

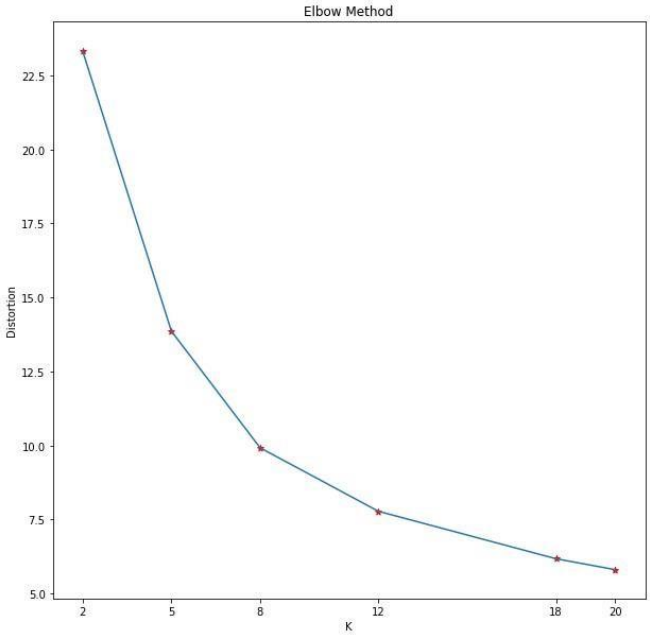

Figure 7: Elbow method for K-Means

Inference:

K = 8 is a suitable value for clustering as per elbow method.

We can also be observed that it acts as a threshold for Distortion and produces maximum purity score.

LAB ASSIGNMENT – VII

Harsh Arya (B20043)

GMM:

K	Purity Score for Train Data	Purity Score for Test Data
2	0.2	0.2
5	0.471	0.466
8	0.588	0.59
12	0.63	0.61
18	0.472	0.45
20	0.424	0402

Table 2: Purity Score for GMM

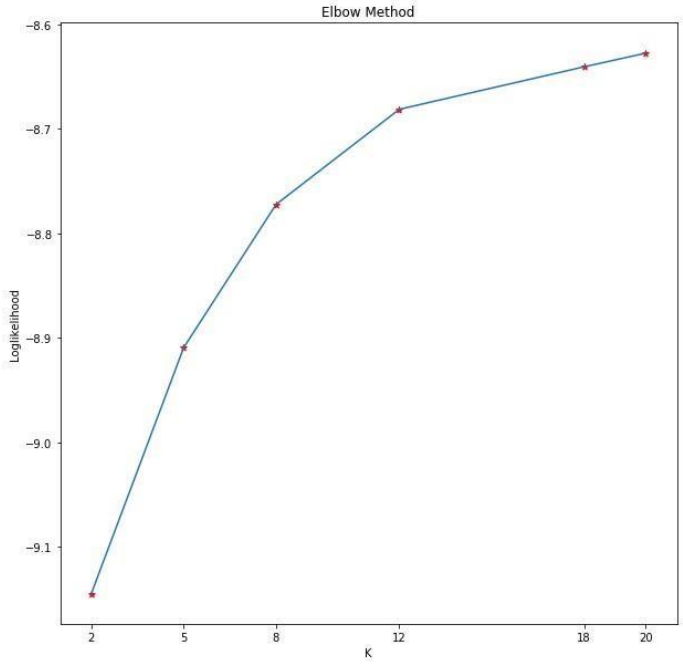

Figure 8: Elbow method for K-Means

Inference:

K = 12 is a suitable value for clustering as per elbow method.

We can also be observed that it acts as a threshold for Log-Likelihood and produces maximum purity score.