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

Student's Name: Smriti Srivastava **Mobile No: 6388490594**

Roll Number: b19116 **Branch: C.S.E**

1 a.

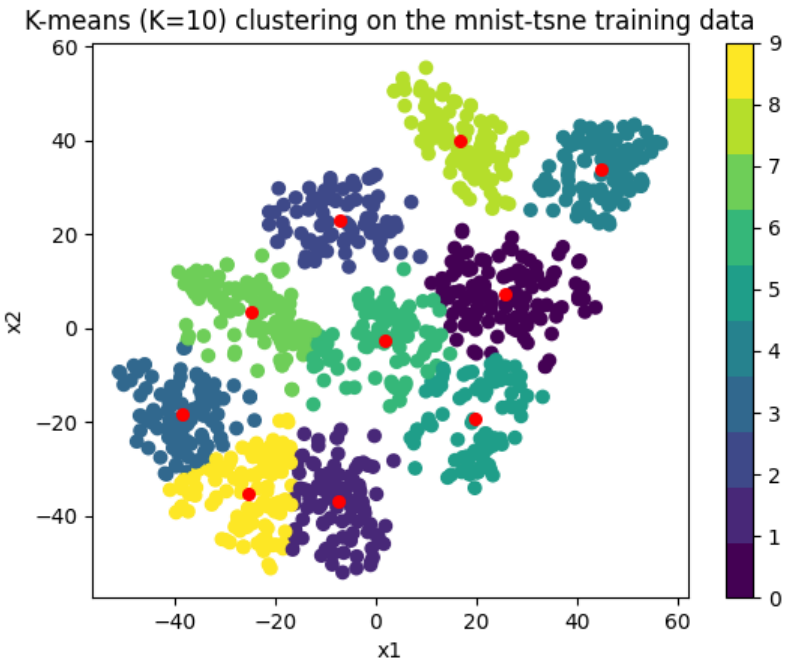


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. K-Means is a clustering algorithm of unsupervised learning. Looking at the above graphs, it has formed circular clusters. The grouping looks accurate though the purity score is 0.69. Being an unsupervised algorithm, K-Means can provide well-formed clusters and is a good clustering algorithm.
2. Yes, K-Means form circular boundaries. Though some cluster boundaries look like oval shaped but majority have circular boundaries.

b.

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

c.

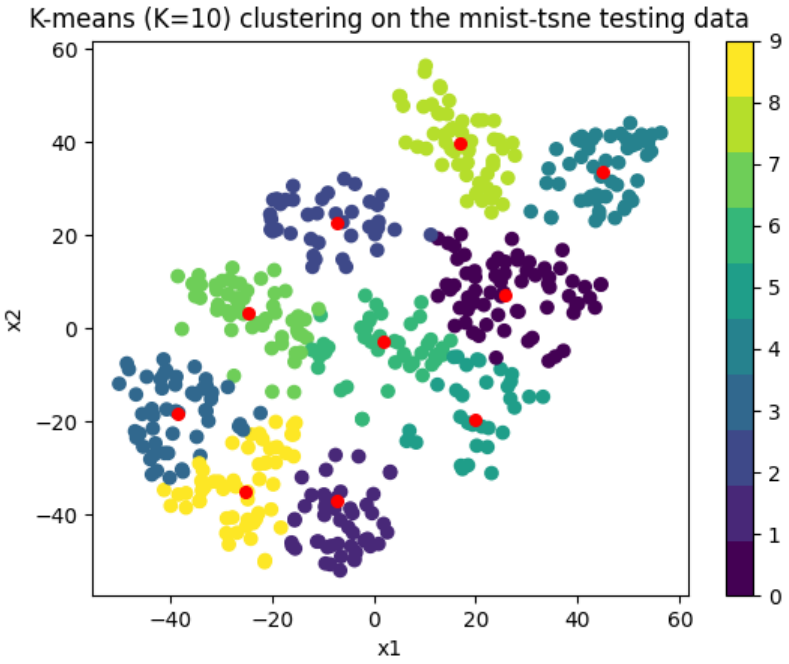


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. As the data in test in less in number the clusters are not that dense. Else, both the graphs are almost similar, the clusters and the cluster centers. The purity scores are also similar.

d.

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

Inferences:

1. Train purity is slightly higher than test purity. This may be because the model is fit on train data and hence gives more accurate results on the train data.

2. The value of K is to be chosen after experimenting. Small K and high values of K does not give accurate results. Elbow method can be used for calculating the optimum K. K-Means also does not perform well when the data consists of several outliers.

2 a.

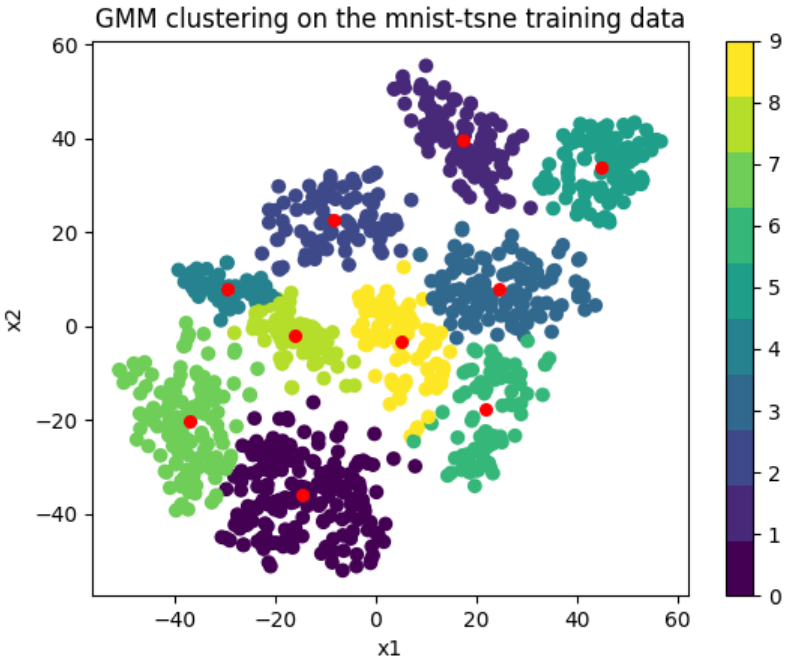


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. GMM is an algorithm of unsupervised learning. Looking at the above graphs, it has formed elliptical clusters. The grouping looks accurate though the purity score is 0.708. Being an unsupervised algorithm, GMM can provide well-formed clusters and is a good clustering algorithm.
2. Yes, GMM form elliptical boundaries. Though some cluster boundaries look like circular shaped but majority have elliptical boundaries.
3. Surprisingly, both the graphs don't differ much but the boundaries are more circular in K-means and elliptical in GMM.

b.

The purity score after training examples is assigned to the clusters is **0.612.**

c.

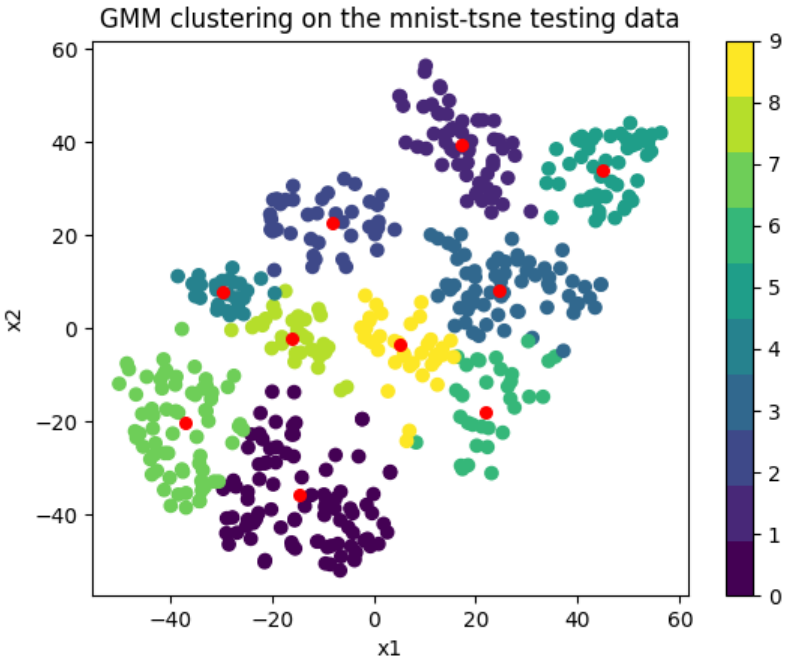


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. Both the graphs are almost similar, the clusters and the cluster centers. The purity scores are also similar.

d.

The purity score after test examples is assigned to the clusters is **0.61.**

Inferences:

1. Train purity is slightly higher than test purity. This may be because the model is fit on train data and hence gives more accurate results on the train data.

2. The main limitation of the GMM algorithm is that, for computational reasons, it can fail to work if the dimensionality of the problem is too high.

3 a.

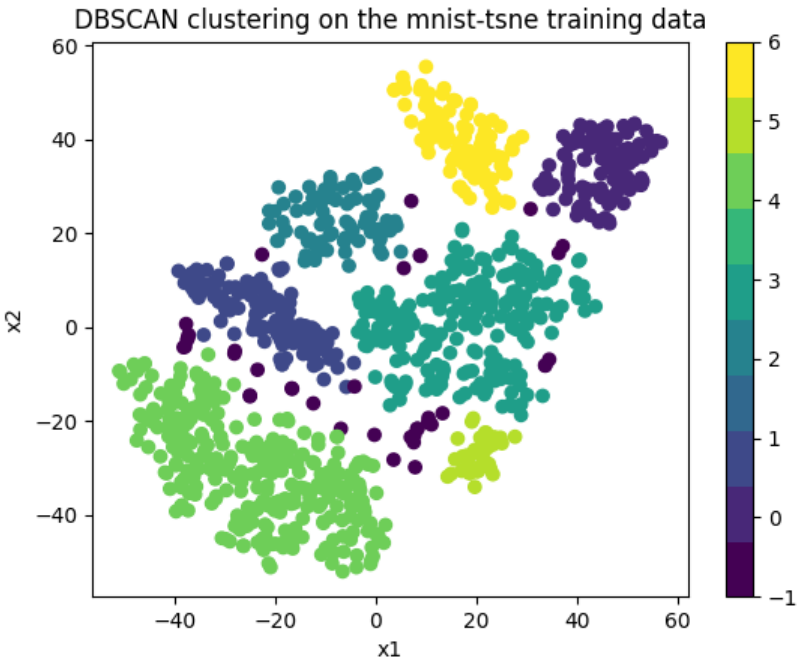


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. DBSCAN does not assume any shape or boundary hence it can discover arbitrarily shaped clusters. IT can find clusters completely surrounded by other clusters. It is not affected by outliers.
2. The major difference is between the number of clusters formed. There are 8 clusters but while using K-Means or GMM we found 10 clusters. This is because DBSCAN finds the clusters solely based on the data distribution and we do not have to manually define the number of clusters.

b.

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

c.

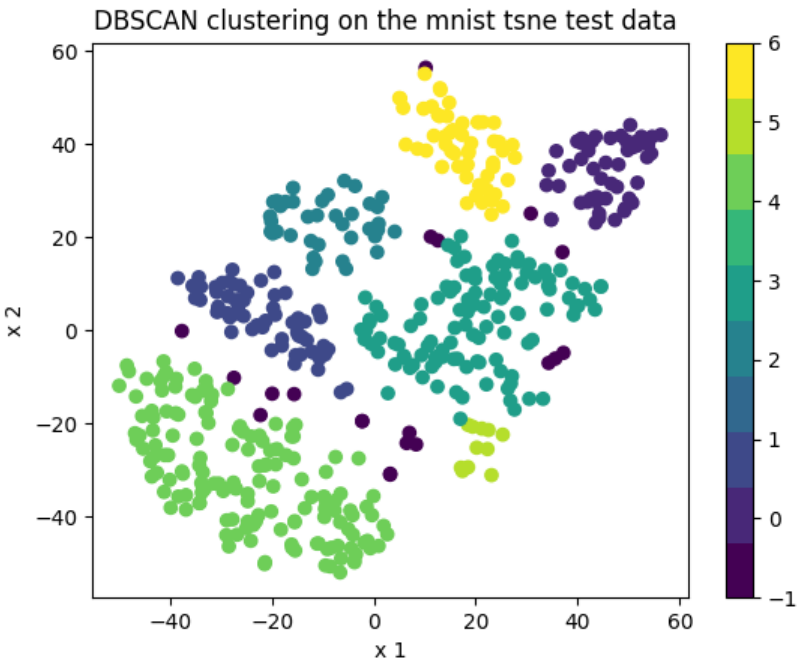


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. The test case is clusters are not condensed, also there seems to be in the cluster points assigned especially in the black class. Overall there is not much difference in the clustering produced in both train and test cases.

d.

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

Inferences:

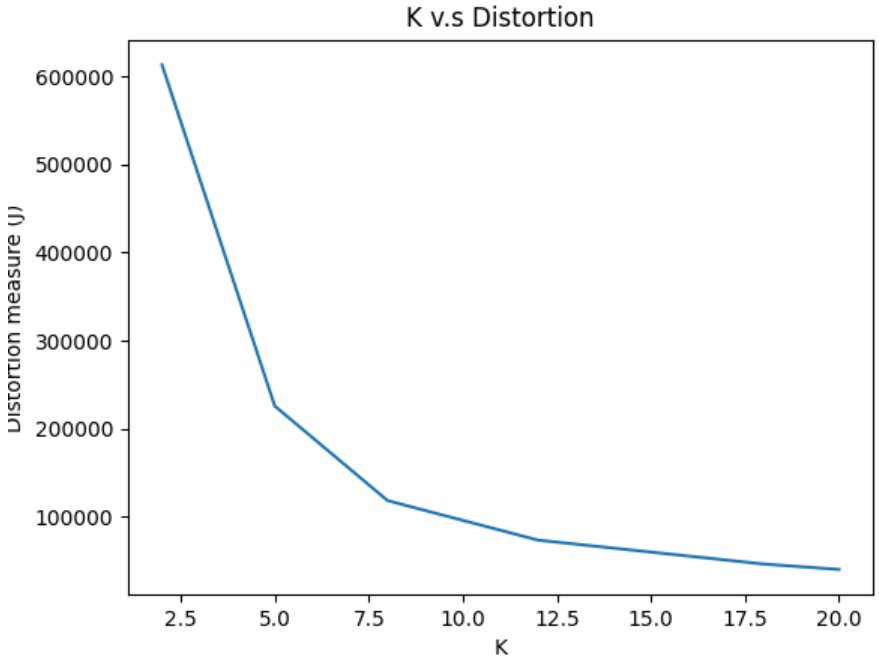
1. 1. Train purity score is higher than test purity score. This is because the model is based on training examples or is learned from training examples but test data points are just assigned classes on the basis of this model.
2. DBSCAN cannot cluster data sets well with large differences in densities. It is because the min Pts and epsilon values cannot be chosen appropriately for all clusters. If the data and scale is not well understood choosing these values can be very difficult.

Bonus Questions

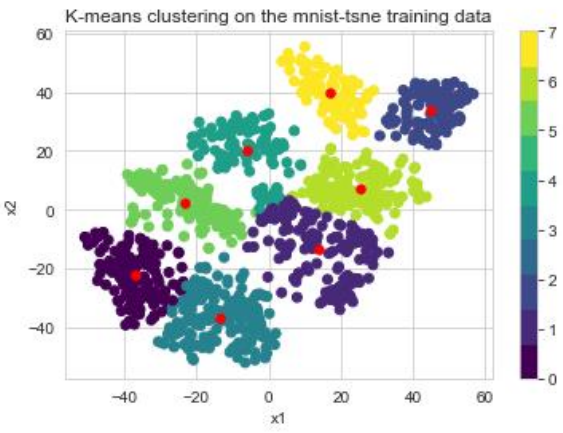
Question A

K-Means:

K	Purity Train	Purity Test
2	0.200	0.200
5	0.393	0.398
8	0.630	0.624
12	0.611	0.612
18	0.493	0.480
20	0.440	0.410

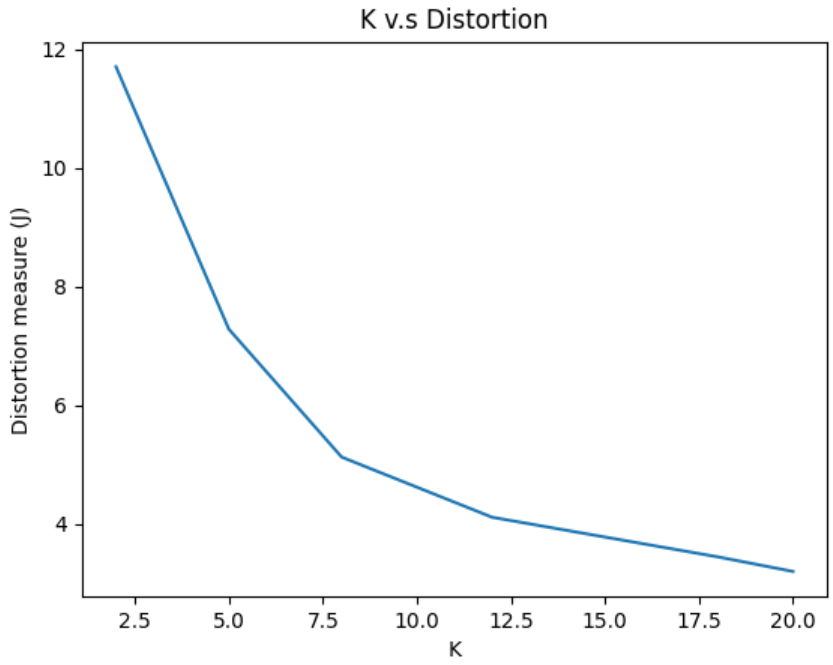


Optimal value of K using K-Means = 8

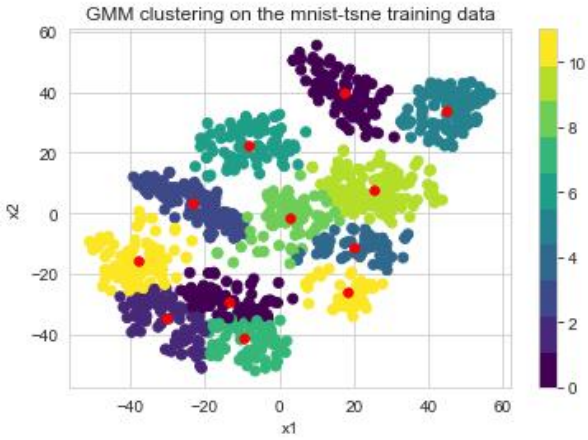


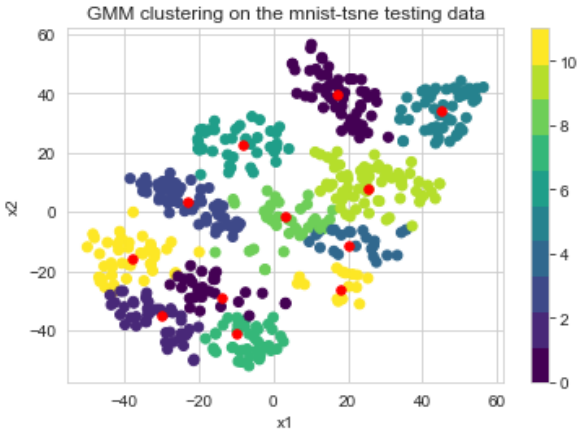
GMM:

K	Purity Train	Purity Test
2	0.200	0.200
5	0.471	0.466
8	0.628	0.640
12	0.638	0.660
18	0.527	0.506
20	0.472	0.456



Optimal value of K using GMM = 12





Question B

esp	Min_samples	Purity Train	Purity Test
1	10	0.100	0.100
1	30	0.100	0.100
1	50	0.100	0.100
5	10	0.585	0.584
5	30	0.158	0.140
5	50	0.100	0.100
10	10	0.100	0.100
10	30	0.100	0.100
10	50	0.503	0.500

The max purity score is for eps = 5 and Min_Samples = 10

The graph of the most optimized pair has been pasted in Q3