**Student’s Name: Srishti Ginjala**

**Roll Number: b19084**

**Mobile No: 9440000900**

**Branch:**CSE

# a.

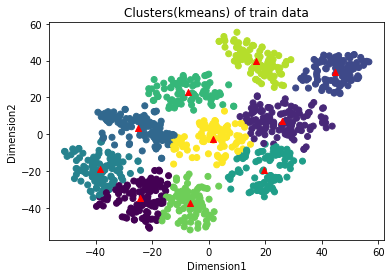


Figure 1 K-means (K=10) clustering on the mnist tsne training data

# Inferences:

1. The clusters formed are good enough to use for test data.
2. Yes, from the output , the boundary seems to be circular.

**b.**

The purity score after training examples are assigned to the clusters is 0.69.

**c.**

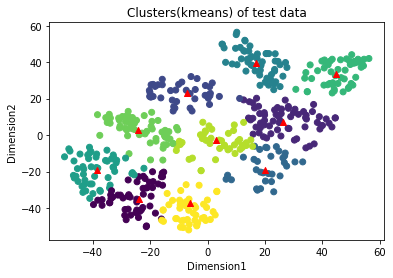


Figure 2 K-means (K=10) clustering on the mnist tsne test data

# Inferences:

1. Compared to the train data clusters, the test data clusters are less dense.

**d.**

The purity score after test examples are assigned to the clusters is 0.684.

# Inferences:

1. Train purity score is more than test because as there are more samples, the error reduces.
2. The user has to specify k (the number of clusters) in the beginning. K-means can only handle numerical data. K-means assumes that we deal with spherical clusters and that each cluster has roughly equal numbers of observations.

# a.

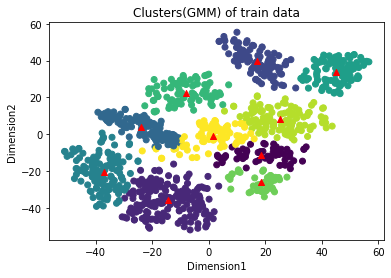


Figure 3 GMM clustering on the mnist tsne training data

# Inferences:

1. The clustering prowess of the algorithm is moderate.
2. Yes, from the output, the boundary seems to be elliptical.
3. Clusters formed using K-means in 1.a are circular and more accurate in classifying data vectors than 2.a.

**b.**

The purity score after training examples are assigned to the clusters is 0.634.

**c.**

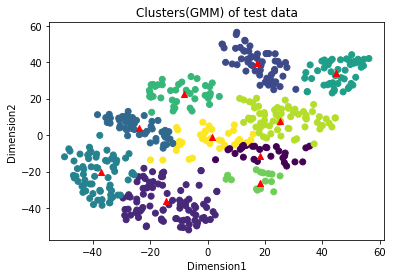


Figure 4 GMM clustering on the mnist tsne test data

# Inferences:

1. Compared to the train data clusters, the test data clusters are less dense.

**d.**

The purity score after test examples are assigned to the clusters is 0.634.

# Inferences:

1. Both are same.
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 ( greater than 6 dimensions for instance).

# a.

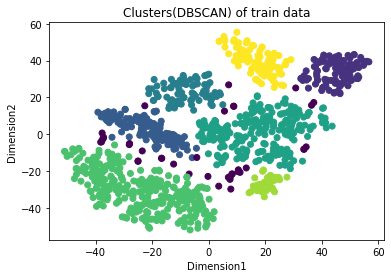


Figure 5 DBSCAN clustering on the mnist tsne training data

# Inferences:

1. The clustering prowess of the algorithm is not very good as it leaves out points which are not densely packed.
2. Clusters formed using K-means in 1.a are more accurate than GMM in 2.a and DBSCAN in 3.a. There are less number of clusters in DBSCAN compared to previous two.

**b.**

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

**c.**

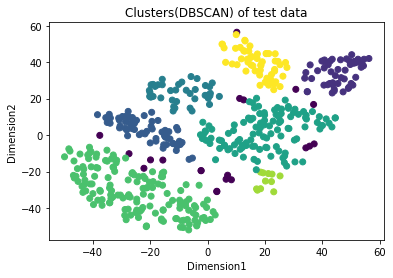


Figure 6 DBSCAN clustering on the mnist tsne test data

# Inferences:

1. Compared to the train data clusters, the test data clusters are less dense.

**d.**

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

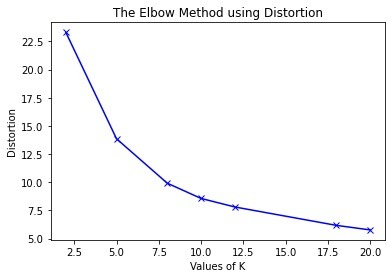
# Inferences:

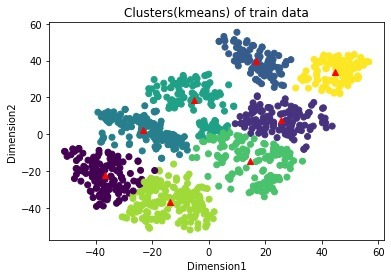
1. Both are almost same.
2. Does not work well when dealing with clusters of varying densities. While DBSCAN is great at separating high density clusters from low density clusters, DBSCAN struggles with clusters of similar density.
3. Struggles with high dimensionality data.

**Bonus questions:**

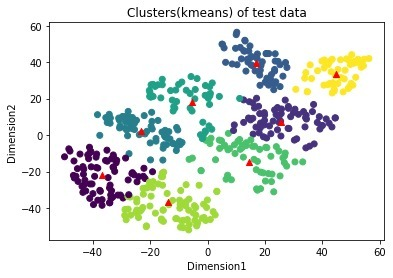
We observed that the optimum number of clusters using elbow method for kmeans clustering and GMM based clustering is 8.The corresponding plots are included below.

For Kmeans



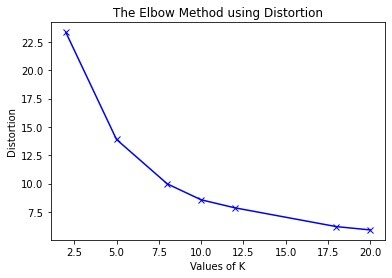


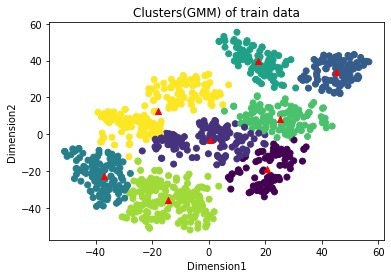
Purity score(kmeans) for train data is 0.63



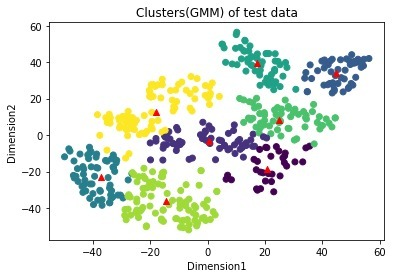
Purity score(kmeans) for test data is 0.624

For GMM





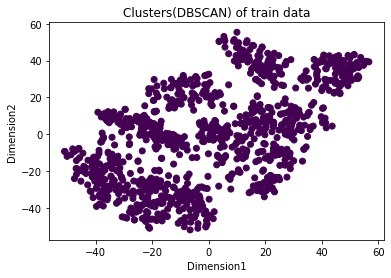
Purity score(GMM) for train data is 0.592



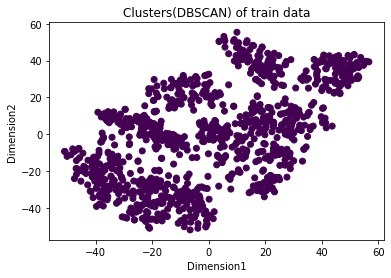
Purity score(GMM) for test data is 0.598

**PART-2**

Let eps=1, samples=10

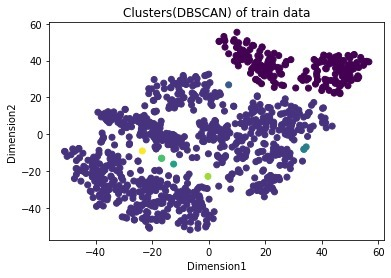


Here we observed that as the radius is small and the number of samples are large ,No point is considered as a core point and we get only one cluster in 1 color. Purity score(DBSCAN) for train data is 0.1

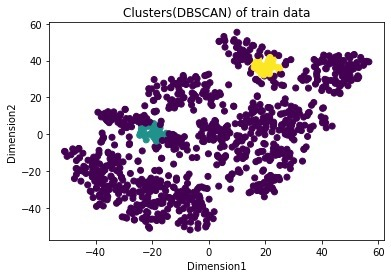


Now eps=10,samples=10

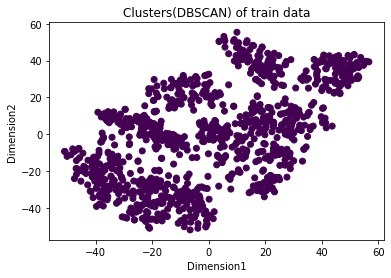
In this case samples are large and also the radius is large so all points are core points and we get only one color. Hence the purity score remains same. Purity score(DBSCAN) for train data is 0.1



Now eps=5,samples=1. This time we get two clusters as the radius is 5 and samples are only 1. So mostly all points will belong to one cluster except for those which are little farther and they result in a different color. Purity score(DBSCAN) for train data is 0.208

Now eps=5,samples=30

In this case, many points are not core points( indicated by violet color ) because the no. of samples is huge except for few densely packed points( indicated by different colors). Purity score(DBSCAN) for train data is 0.158



Now eps=5, samples=50

Here also 50 is a huge number for epsilon of 5 and hence there are no core points and all of them result in a single cluster. Purity score(DBSCAN) for train data is 0.1

So, we observe that optimal value of eps=5 and no.of min samples is 10 to get ideal no.of clusters.

