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

Student's Name: Prashant Kumar **Mobile No: 8700350173**

Roll Number: B19101 **Branch: CSE**

1 a.

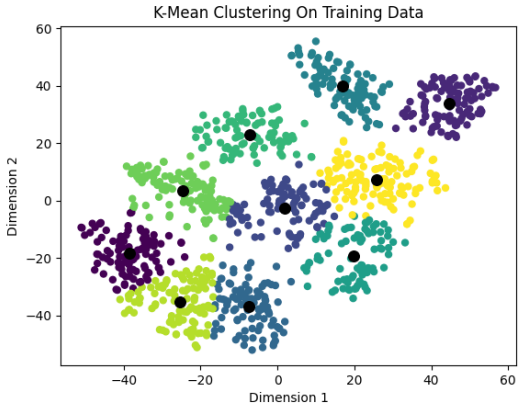


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. From the clusters formed in the above plot, we can say that the clustering prowess of the algorithm is quite good.
2. K-means algorithm assumes cluster boundaries to be circular in 2D. From the output, some of the boundaries are seem circular shape.

b.

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

c.

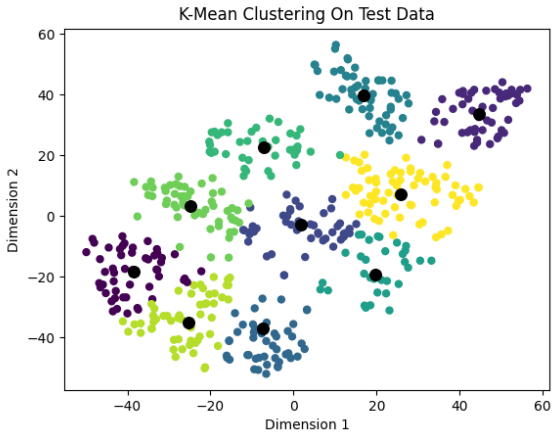


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. Test data seems to be less concentrated as compared to the train data. But there is not much difference in the distribution of data.

d.

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

Inferences:

1. Purity score of the train data is higher in magnitude as compared to the test data. There is not any specific reason, test data may have high purity score.
2. Limitation of this clustering approach is that the user has to specify k in the beginning.

2 a.

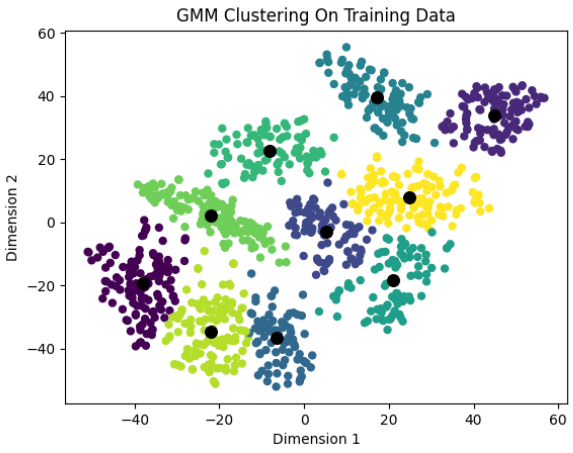


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. From the clusters formed in the above plot, we can say that the clustering prowess of the algorithm is quite good and the GMM is feebler in terms of cluster covariance.
2. GMM algorithm constraints cluster boundaries to be elliptical in 2D. From the output, the boundary also seems to be elliptical in shape.
3. GMM clustering is cleaner around boundaries as compared to K-Means clustering.

b.

The purity score after training examples are assigned to the clusters is **0.708.**

c.

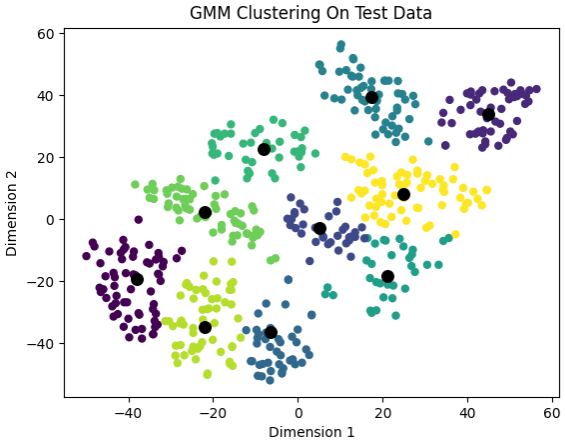


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. There is not much difference in the distribution of training and test data. Only the train data is more concentrated as compared to the test data.

d.

The purity score after test examples are assigned to the clusters is **0.704.**

Inferences:

1. Purity score of the train data is higher in magnitude as compared to the test data. There is not any specific reason, test data may have high purity score. It also depends on the fact that the model is trained from the training data which can also be one of the reasons for quite high purity score.
2. Limitation of this clustering approach is that we have specify k in the beginning.

3 a.

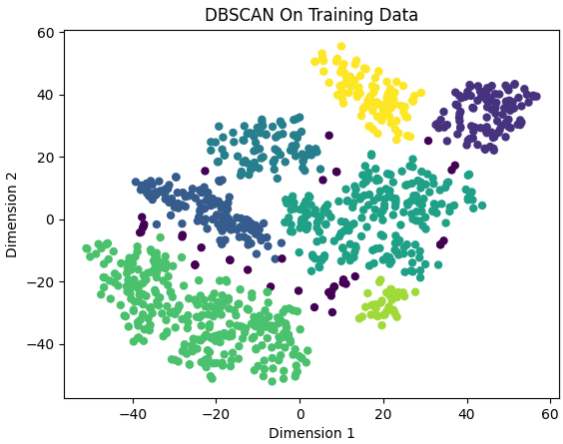


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. AS DBSCAN does not depend on any random value it just follows the path that have similar density. It is not affected by outliers.
2. DBSCAN formed 8 cluster. Here we not have to specify number of cluster but in K-means and GMM we have to specify k in the beginning.

b.

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

c.

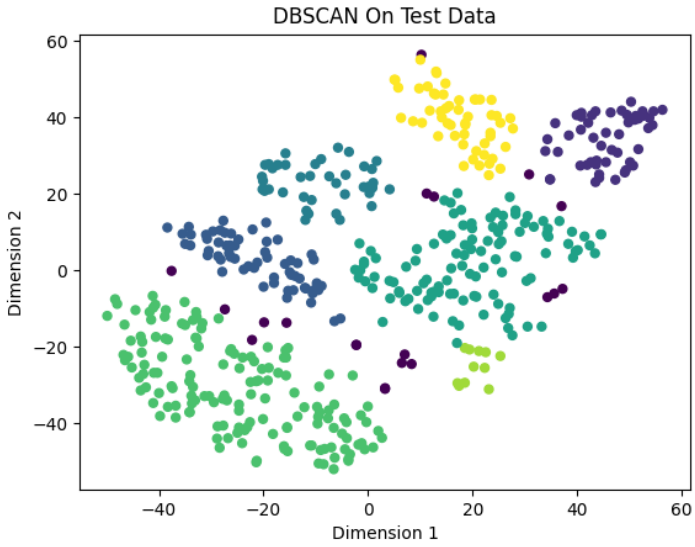


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. Cluster of test case are almost similar to training data as the data used for test and train data are

d.

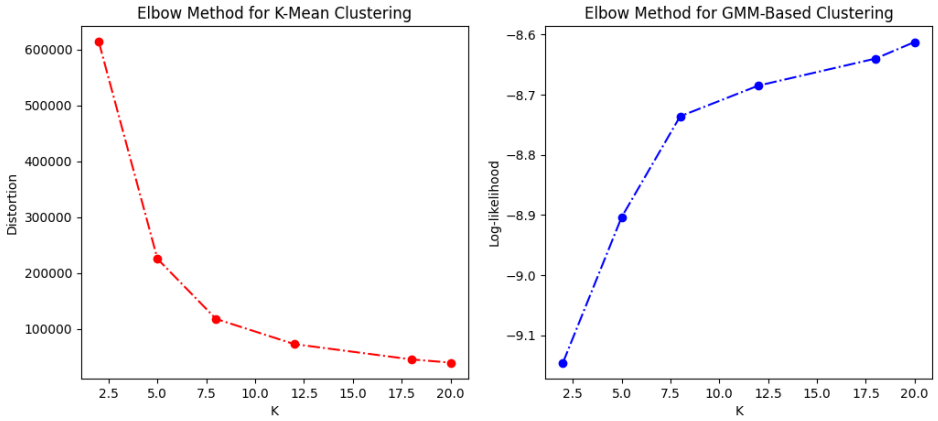
The purity score after test examples are assigned to the clusters is **0.584.**

Inferences:

1. The train purity score (**0.585**) is just slightly greater than the test purity score (**0.584**). Here the data is almost same that's why the purity score doesn't differ more.
2. Limitation of this clustering approach is that this was not good for highly dense data.

Bonus Questions

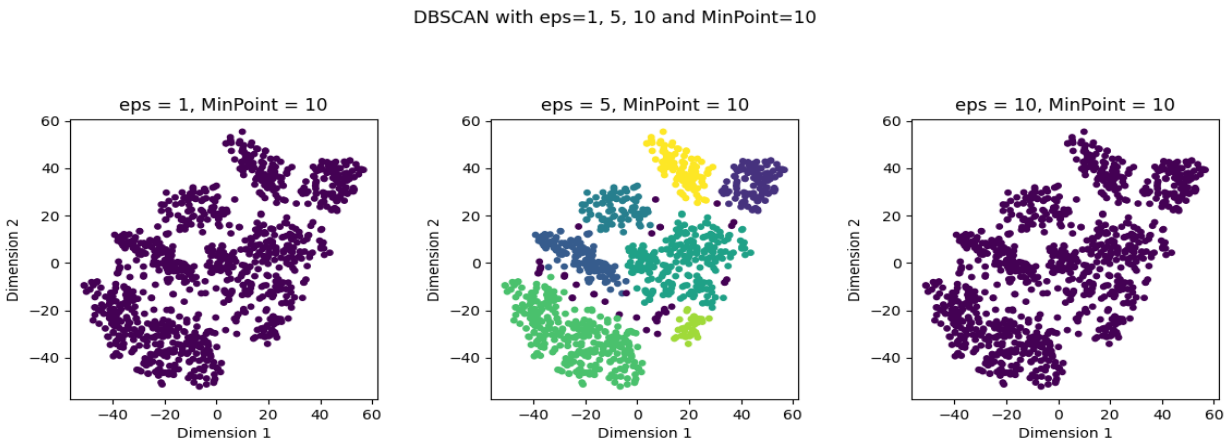
1)



K	Purity Score (K-Mean)	Purity Score (GMM)
2	**0.2**	**0.2**
5	**0.393**	**0.46**
8	**0.630**	**0.629**
12	**0.611**	**0.66**
18	**0.481**	**0.508**
20	**0.432**	**0.455**

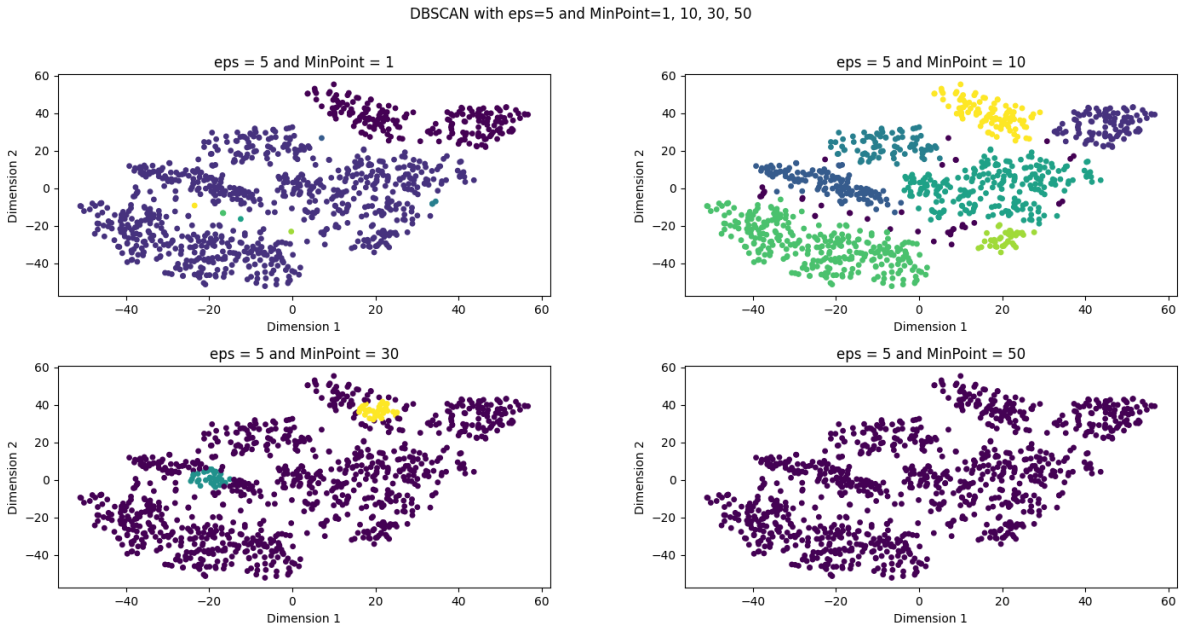
To determine the optimal number of clusters, we must select the value of k at the “elbow” i.e., the point after which the distortion/inertia start decreasing in a linear fashion. Thus, in the given data, we can say that the optimal number of clusters for the data is 8 that we used as a k in our data. For different value of K we can see that the purity score increased and then decreased. Hence the **optimal purity score experimentally found for K=8**.

2)



Purity score table for different epsilon:

Eps	Minimum samples	Purity score of training data
1	10	0.1
5	10	0.585
10	10	0.1



Purity score table for different minimum samples:

Eps	Minimum samples	Purity score of training data
5	**1**	**0.208**
5	**10**	**0.585**
5	**30**	**0.158**
5	**50**	**0.1**

Best epsilon for data clustering = **5**

Best minimum sample for data clustering = **10**

The maximum Purity score is for eps = 5 and min samples = 10. For rest of the value, no good clustering can be found, and we can also see the purity value are very less. This can be attributed that the data is very dense for some values of eps and in some cases no distinct boundary can be found.