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

Student's Name: Rashmi

Mobile No: 7015331137

Roll Number: B19218

Branch: Engineering Physics

1 a.

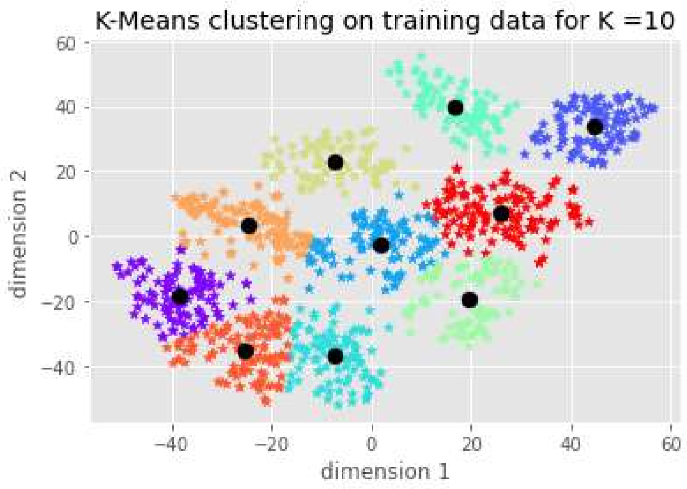


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. First assign random centers and then assign data points to nearest center then update center with mean and repeat until centers does not change.
2. Here boundary seems to be linear.

b.

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

c.

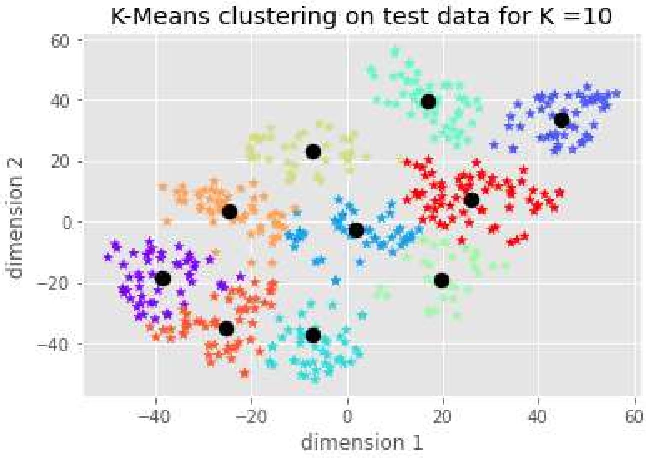


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. There is no difference in the distribution of the data as such. The only difference is that there is less number of data points.

d.

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

Inferences:

1. Train purity score is higher than test purity score. This is because the model is based on training examples but test data points are just assigned classes on the basis of this model.
2. It is sensitive to outliers.

2 a.

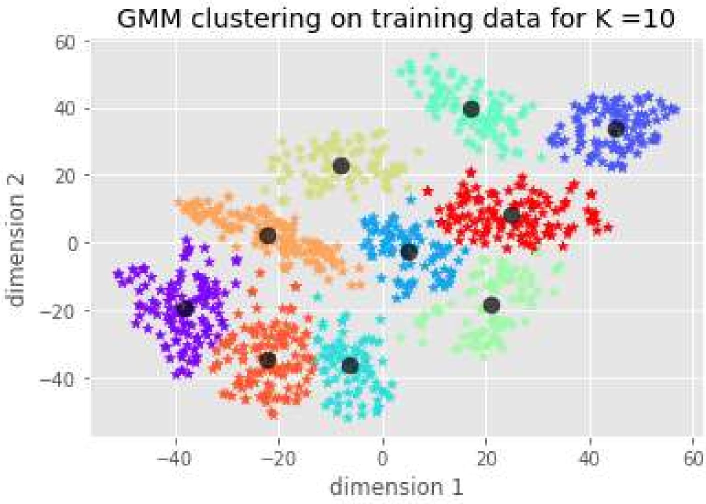


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. In this we use mean and covariance to represent cluster and Expectation maximum is used to predict parameters.
2. The boundary seems to be elliptical.
3. Both type of clustering yield almost same clusters.

b.

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

c.

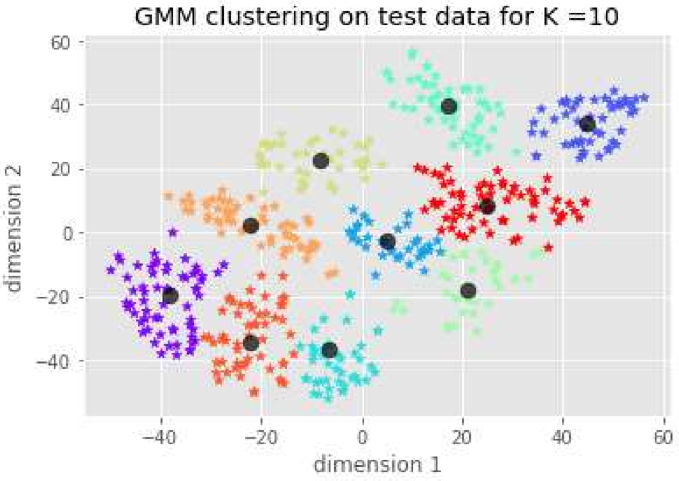


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. There is no difference in the distribution of the data as such. The only difference is that there is less number of data points.

d.

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

Inferences:

1. Train purity score is higher than test purity score. This is because the model is based on training examples but test data points are just assigned classes on the basis of this model.
2. It assumes gaussian distribution.

3 a.

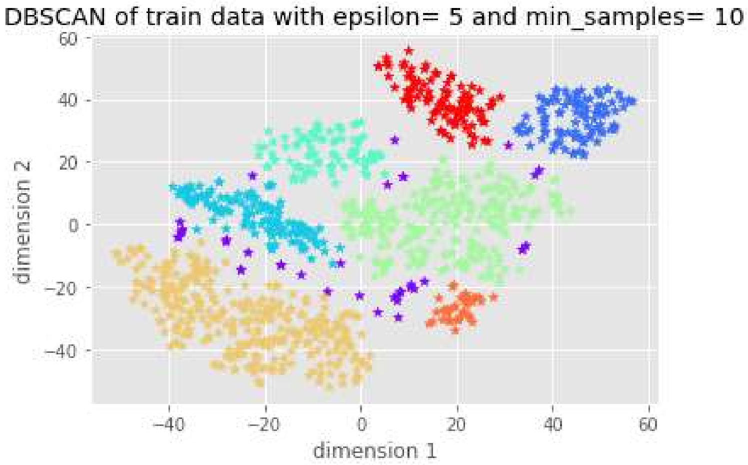


Figure 5 DBSCAN clustering on the mnist tsne training data

Inferences:

1. It first finds out the connected components based on core and border points and rest are outliers. And one component is assigned one cluster.
2. In DBSCAN the number of clusters are less as compared to other as it does not consider outliers and forms cluster of arbitrary shape.

b.

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

c.

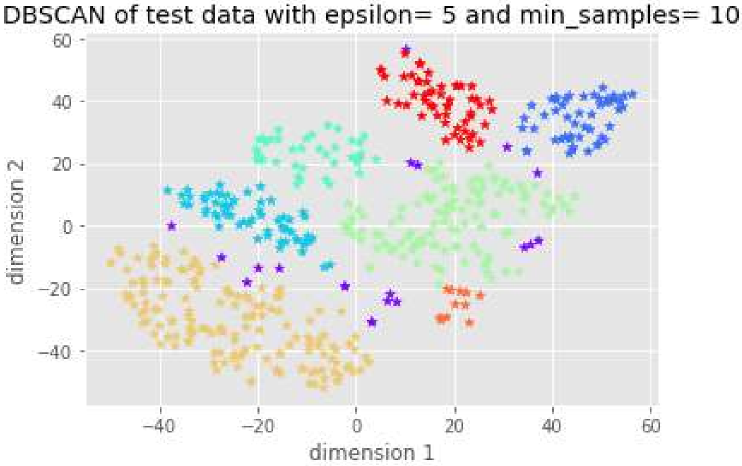


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. There is no difference in the distribution of the data as such. The only difference is that there is less number of data points.

d.

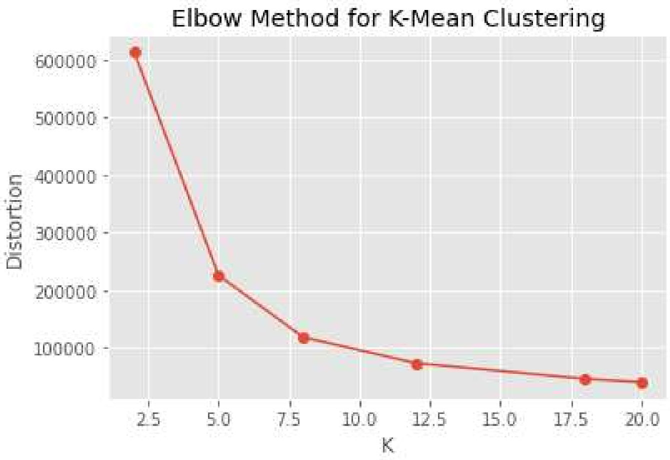
The purity score after test examples are assigned to the clusters is **0.584**

Inferences:

1. Train purity score is higher than test purity score. This is because the model is based on training examples but test data points are just assigned classes on the basis of this model.
2. This Clustering is not suitable if there is no region of low density.

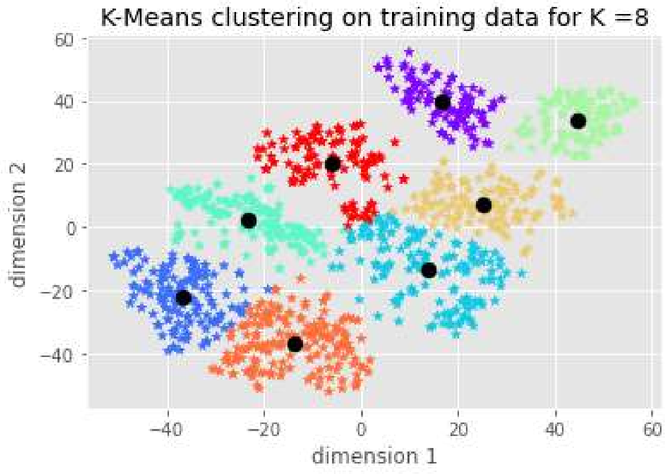
4 Bonus Question:

(A) (i)

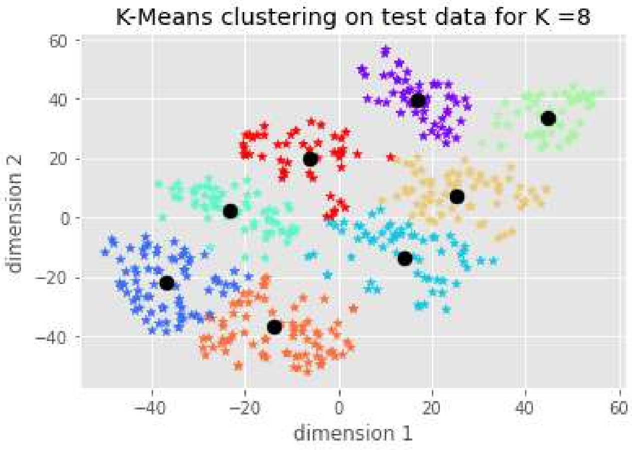


Inferences:

1. Above graph shows that **K=8** is optimal value of clusters for K-Means clustering.

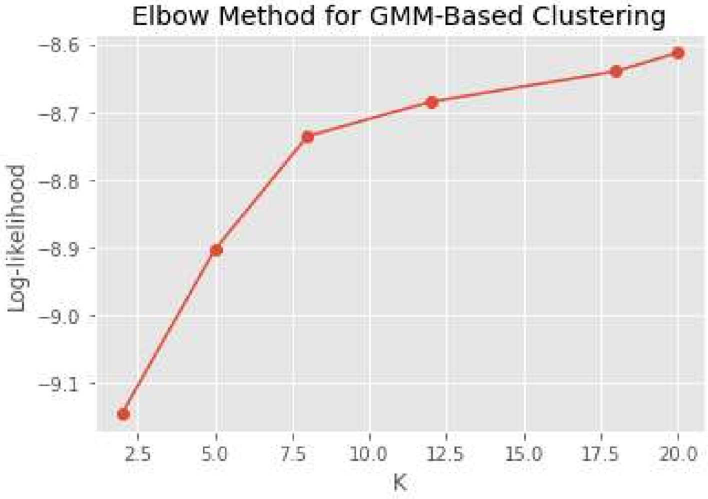


Purity score = **0.63**



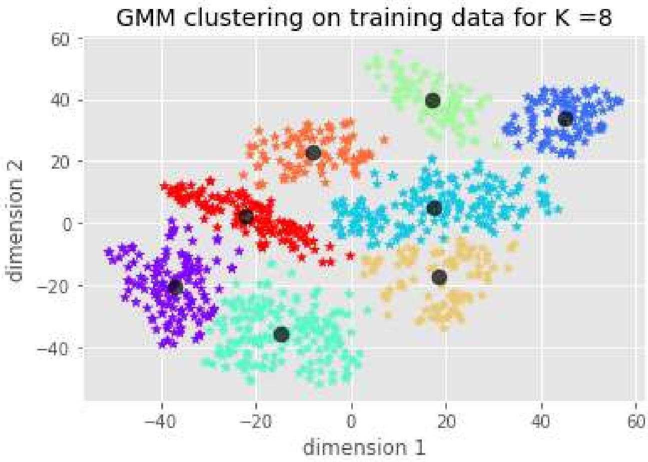
Purity score = **0.624**

(ii)

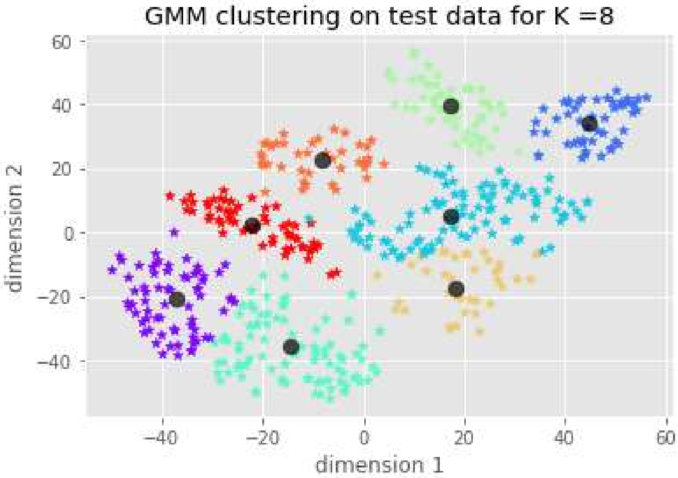


Inference:

1. Above graph shows that **K=8** is optimal value of clusters for GMM-Based clustering.

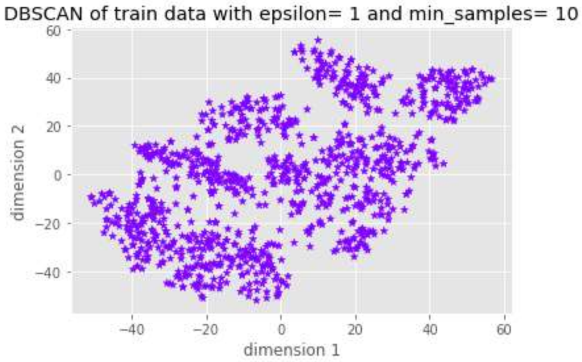


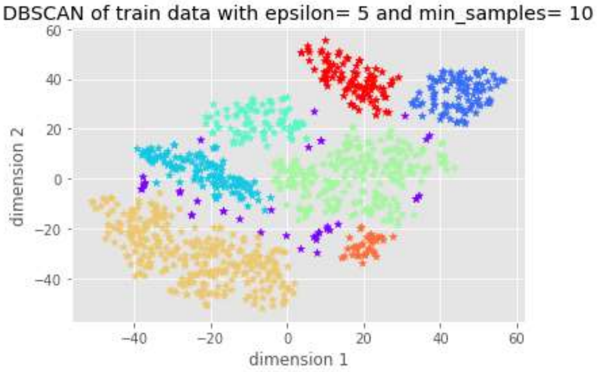
Purity score = **0.629**

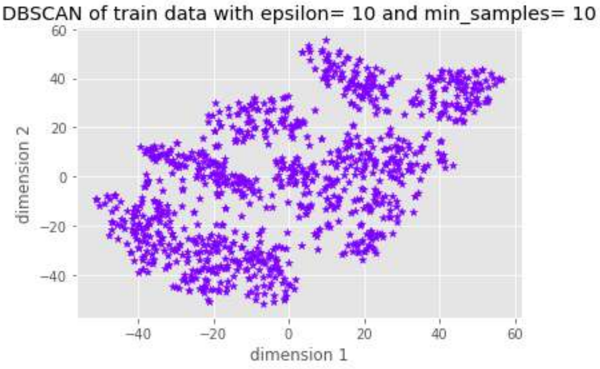


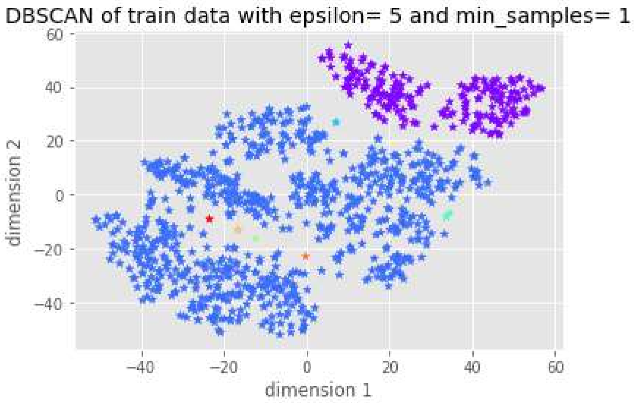
Purity score = **0.628**

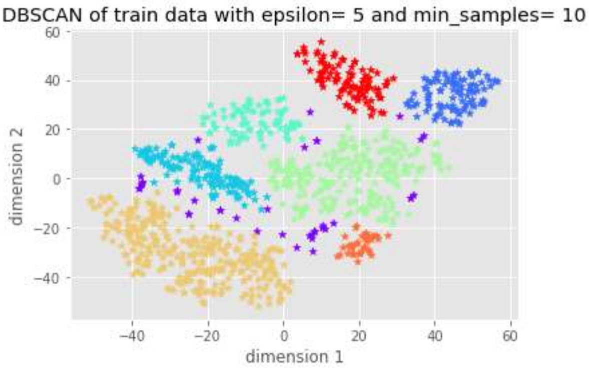
(B)

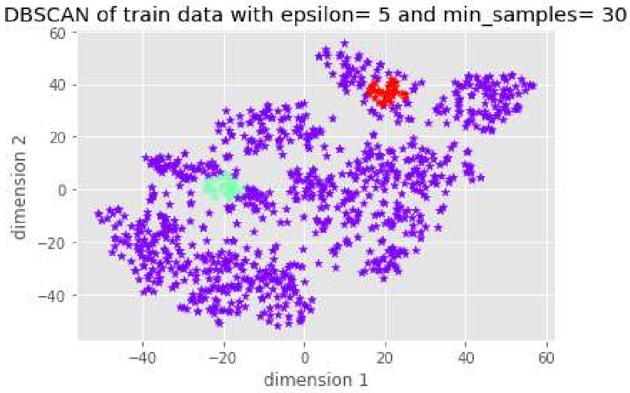

DBSCAN of train data with epsilon= 1 and min_samples= 10
60
40
20
0
−20
−40
dimension 2
−40
−20
0
20
40
60
dimension 1

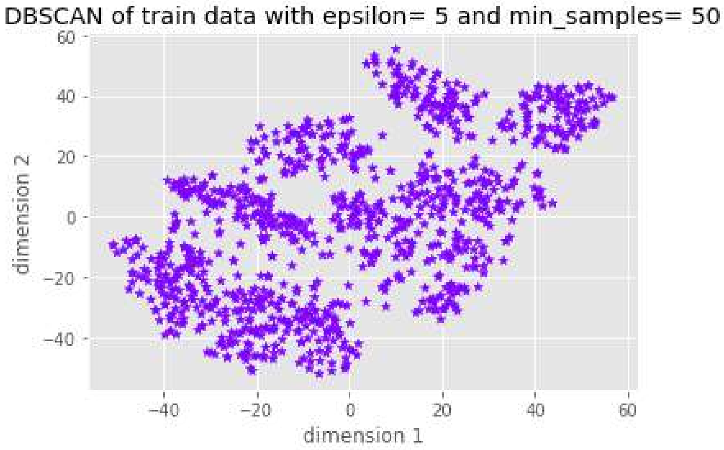

DBSCAN of train data with epsilon= 5 and min_samples= 10
60
40
20
0
−20
−40
dimension 2
−40
−20
0
20
40
60
dimension 1


DBSCAN of train data with epsilon= 10 and min_samples= 10
60
40
20
0
−20
−40
dimension 2
−40
−20
0
20
40
60
dimension 1


DBSCAN of train data with epsilon= 5 and min_samples= 1
dimension 1
dimension 2
60
40
20
0
−20
−40
−40
−20
0
20
40
60


DBSCAN of train data with epsilon= 5 and min_samples= 10
dimension 1
dimension 2
60
40
20
0
−20
−40
−40
−20
0
20
40
60


DBSCAN of train data with epsilon= 5 and min_samples= 30
dimension 1
dimension 2
60
40
20
0
−20
−40
−40
−20
0
20
40
60


DBSCAN of train data with epsilon= 5 and min_samples= 50
dimension 2
dimension 1
60
40
20
0
−20
−40
−40
−20
0
20
40
60