Indian Institute of Technology Mandi

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

1 a.

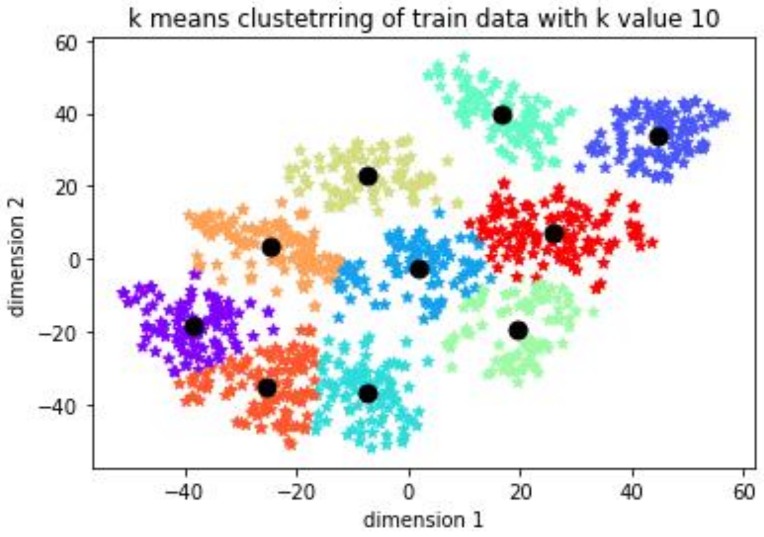


Figure 1 K-means (K=10) clustering on the mnist tsne training data

Inferences:

1. K-Means algorithm is easy to understand as well as implement. It can be used for large data sets. It guarantees convergence in almost all cases except some. As in this case the distribution of the cluster and their size comparing to each of them are approximately equal and the cluster center is very well placed making it a suitable K number for clustering.
2. Yes, the boundary in this case seems to be a circular boundary but not an exactly circular boundary, it is evident from the plot as from the center of the clusters(x) the points in the cluster show a distribution which seems to be forming approximately circular boundary.

b.

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

c.

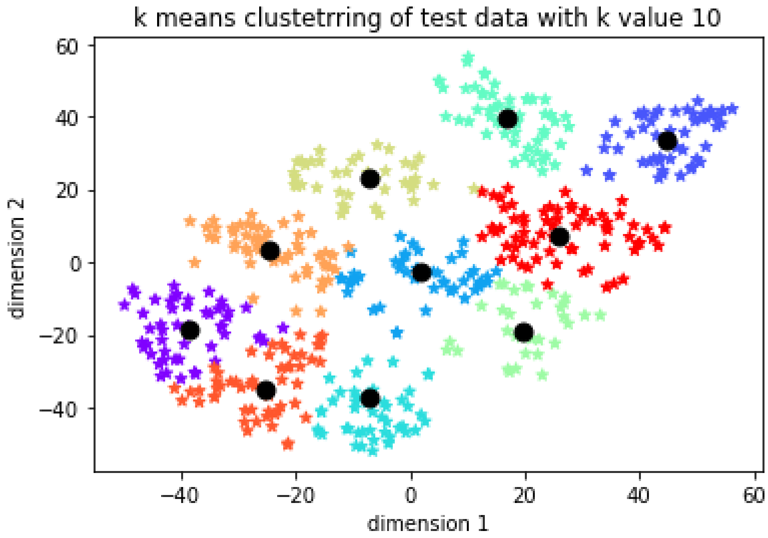


Figure 2 K-means (K=10) clustering on the mnist tsne test data

Inferences:

1. Inferring from both the plots we can say that the difference is not very identical according to the distribution of the data but the only difference here is the number of data points are quite less as compare to the train data set so, it is a little bit less dense.

d.

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

Inferences:

1. As compare between test and train purity score, the purity score of train data is higher than test data purity score. This is because the model is basically based on training examples or we can say that it is modeled or learned from training examples but test data points are just assigned classes on the basis of this model.
2. We must choose the k value manually. Outliers severely impact the performance of k-means algorithm. It depends on initial values so with each new run on the same data we do not get the same result, this is due to randomly choosing centers in the starting step.

2 a.

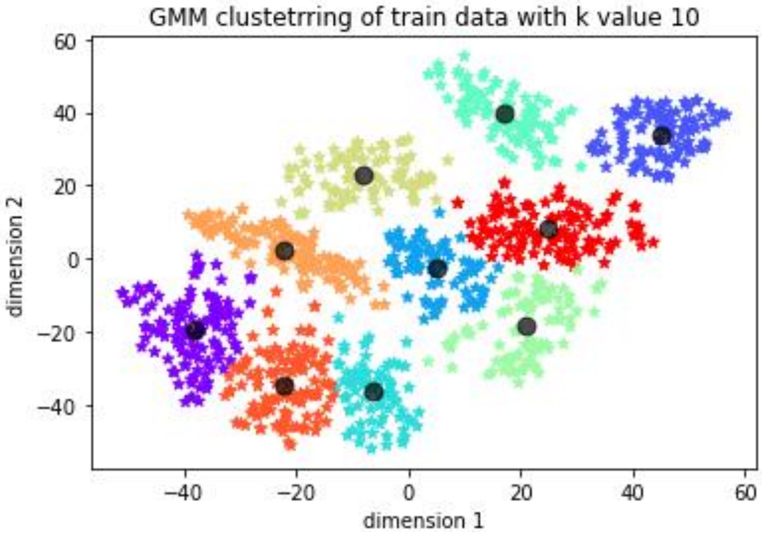


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. GMM is more flexible in terms of cluster covariance and hence is more efficient then K- Means in most cases as using GMM method we can say that for larger no of data pints the relationship it shows is basically forming a gaussian distribution which also results in increasing its purity score as compare to the kmeans purity score.
2. Using GMM algorithm clustering method we can only say that by the inference of the above plot the boundary seems a bit elliptical in shape but obviously it is not exactly elliptical shape but a bit it is similar

3. Inferring from plot 1.(a) and plot 2.(a) we can say that the clustering is more uniformly distributed among all the clustery as well as the boundaries in GMM is approximately elliptical in shape on the other hand Kmeans boundary form shape which is approximately circular.

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

c.

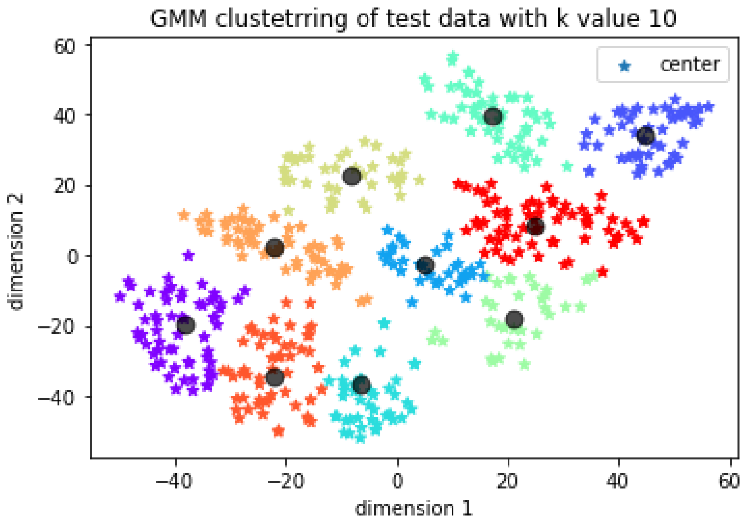


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. Inferring from both the plots we can say that the difference is not very identical according to the distribution of the data but the only difference here is the number of data points are quite less as compare to the train data set so, it is a little bit less dense

d.

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

Inferences:

1. As compare between test and train purity score, the purity score of train data is higher than test data purity score. This is because the model is basically based on training examples or we can say that it is modeled or learned from training examples but test data points are just assigned classes on the basis of this model.
2. It perform better in case of the data containing larger number of values as due to large number of data set the relation ship between them is something like gaussian distribution finally resulting in a better purity score as well as a better model.

3 a.

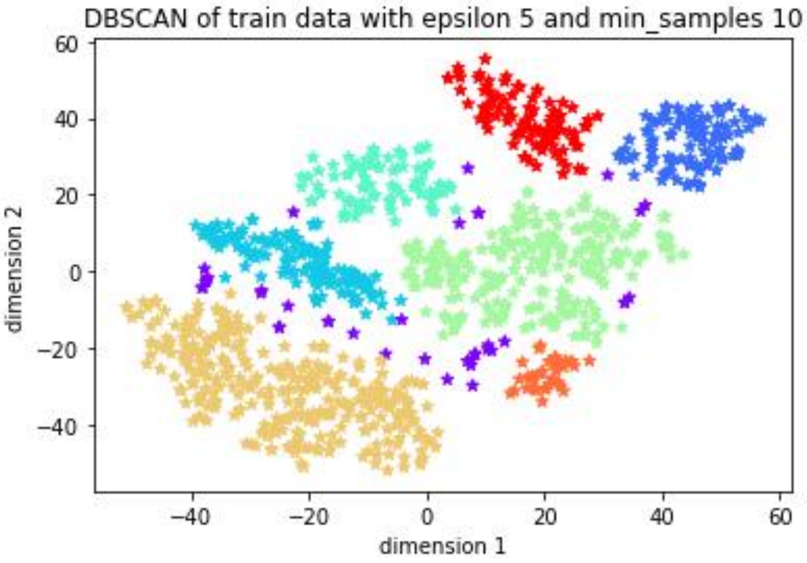


Figure 5 DBSCAN clustering on the mnist tsne training data

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 and also the clusters forms are not uniformly distributed among each cluster form in this case. Also, the main quality of this method is that it is not affected by outliers present in the given data.
2. From the above plot we can infer that the distribution using DBSCAN of each cluster form is not uniform or we can say that it is not equally distributed among all the cluster.
3. 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 are assigned to the clusters is 0.585

c.

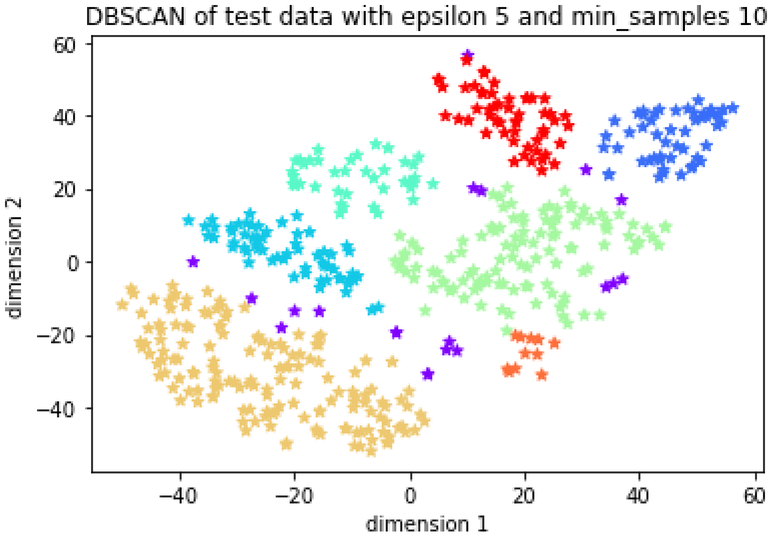


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. From the inference of the above both plots of DBSCAN we can say that there not much difference between both of them except the number of data used to form the plot in both cases but also they are differ a bit as it is not exactly equal.

d.

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

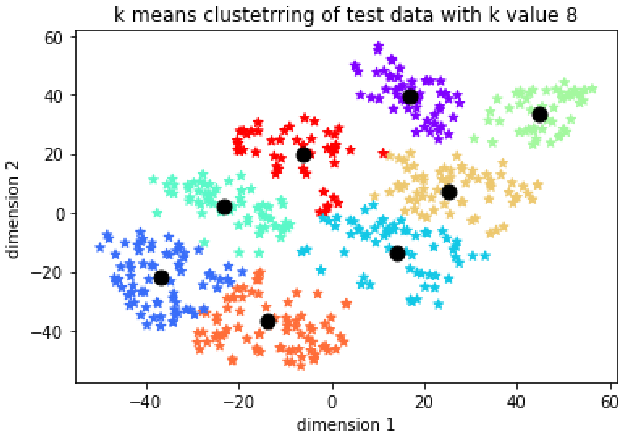
Inferences:

1. The purity score in this method of DBSCAN is not much differ from one another as the difference of purity score in both the case is quite negligible which is the best part of this method. But as the model is build using the train data so in this case also the purity score is higher for train data as compare to the test data.
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.

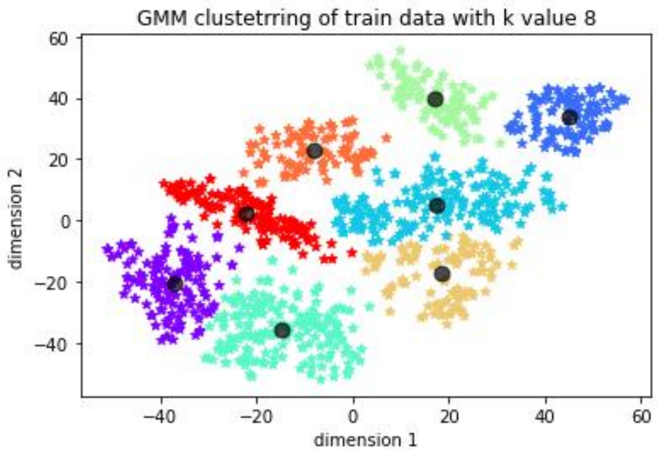
Bonus Questions ()

1) **Bonus ques1**

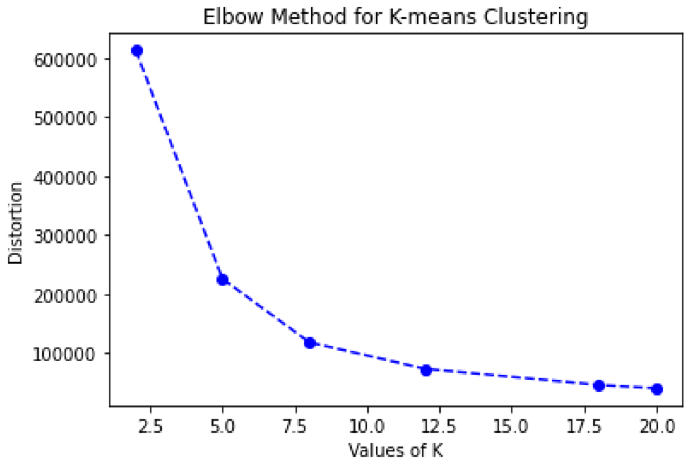
1. Accuracy for best purity score for K = 8 using Kmeans method: -



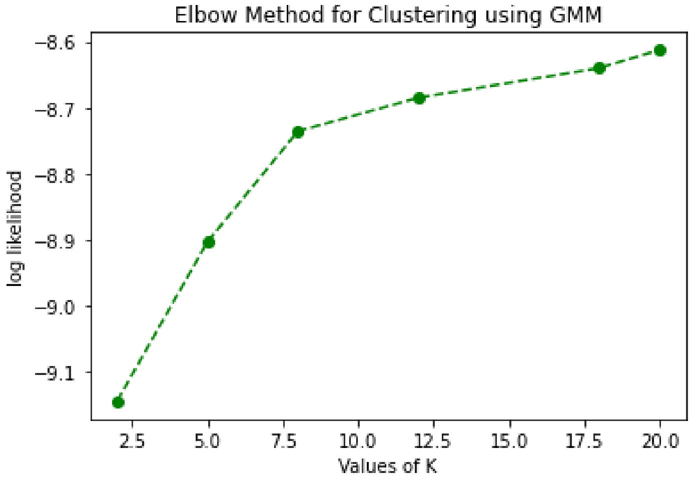
2. Accuracy for best purity score for K = 8 using GMM method: -



3. Plot of Kmeans using elbow method: -



4. Plot of GMM using elbow method: -



5. Table containing purity score for each value of K using Kmeans method for train data: -

K value	Purity Score
2	0.2
5	0.393
8	0.63
12	0.611
18	0.481
20	0.432

6. Table containing purity score for each value of K using Kmeans method for test data: -

K value	Purity Score
2	0.2
5	0.398
8	0.624
12	0.612
18	0.46
20	0.416

7. Table containing purity score for each value of K using GMM method for train data: -

K value	Purity Score
2	0.2
5	0.46
8	0.629
12	0.66
18	0.508
20	0.455

8. Table containing purity score for each value of K using GMM method for test data: -

K value	Purity Score
2	0.2
5	0.448
8	0.628
12	0.646
18	0.51
20	0.46

2) Bonus ques2

- Table containing purity score of DBSCAN method when min sample = 10 for train data: -

Epsilon value	Purity Score
1	**0.1**

5	**0.585**
10	**0.1**

- Table containing purity score of DBSCAN method when min sample = 10 for test data: -

Epsilon value	Purity Score
1	**0.1**
5	**0.584**
10	**0.1**

- Table containing purity score of DBSCAN method when epsilon = 5 for train data: -

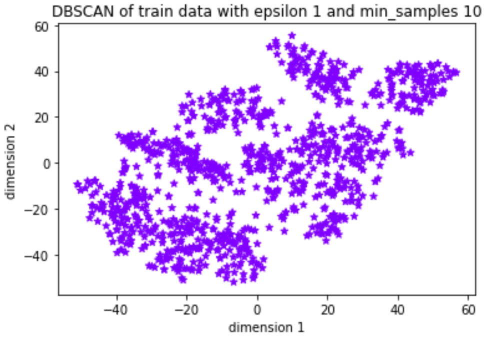
Min Sample value	Purity Score
1	**0.208**
10	**0.585**
30	**0.158**
50	**0.1**

- Table containing purity score of DBSCAN method when epsilon = 5 for test data: -

Min Sample value	Purity Score
1	**0.212**
10	**0.584**
30	**0.14**
50	**0.1**

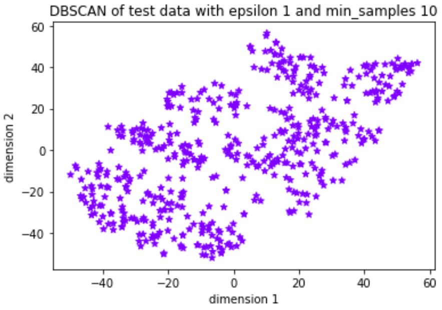
3) Plots for all the above values of epsilon and min value: -

i.

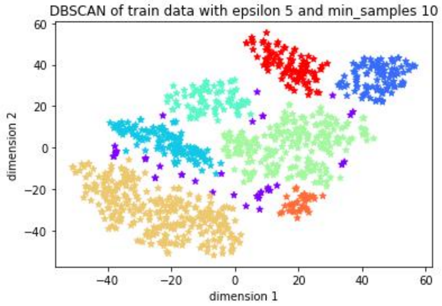


DBSCAN of train data with epsilon 1 and min_samples 10

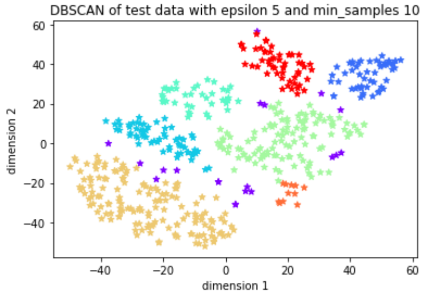
ii.



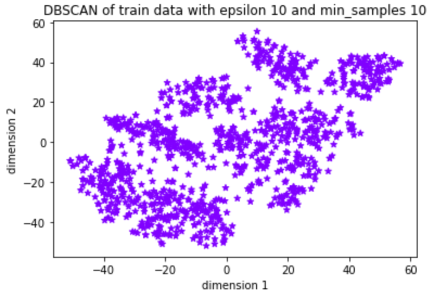
DBSCAN of test data with epsilon 1 and min_samples 10 iii.



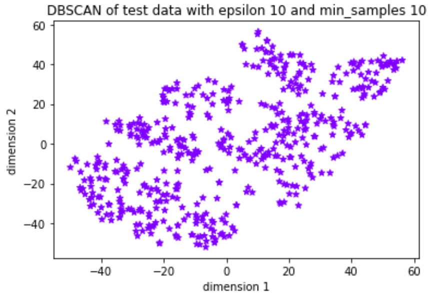
iv.



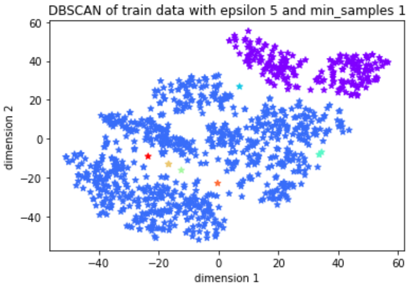
v.



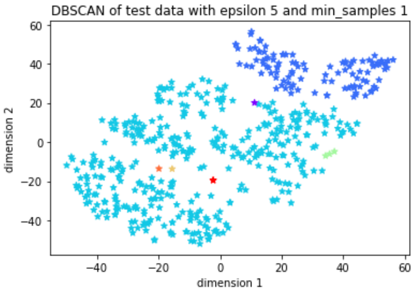
vi.



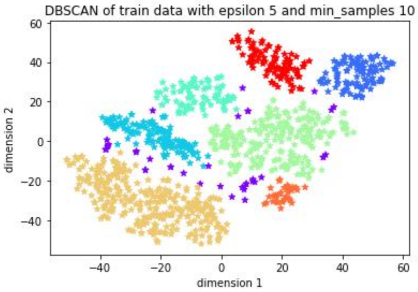
vii.


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

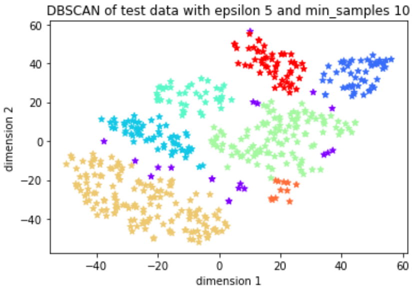
viii.


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

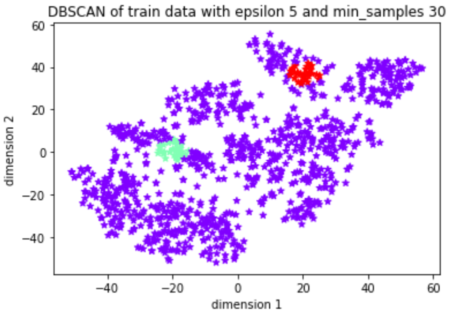
ix.


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

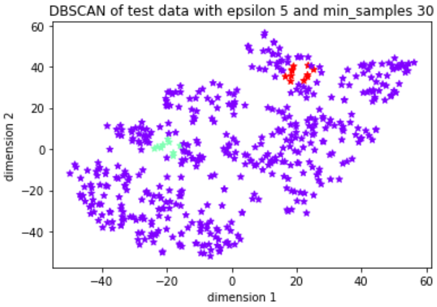
x.


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

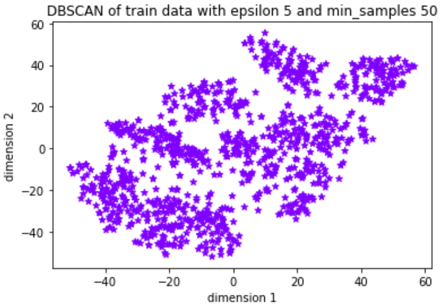
xi.



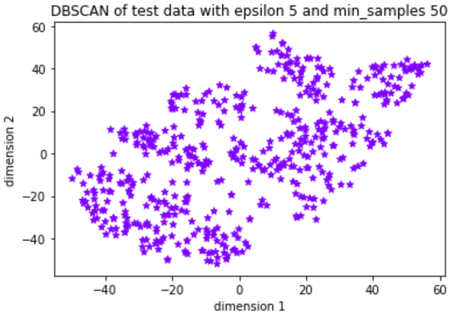
xii.



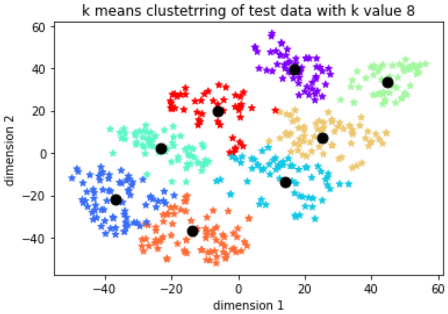
xiii.



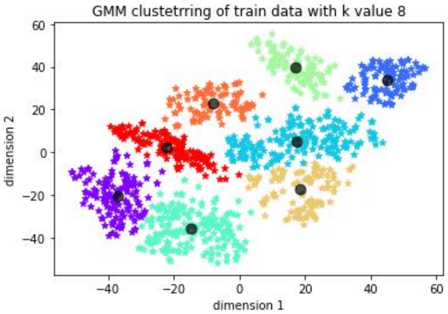
xiv.



xv.


k means clustetrring of test data with k value 8
dimension 2
dimension 1
60
40
20
0
−20
−40
−40
−20
0
20
40
60

xvi.


GMM clustetrring of train data with k value 8
dimension 2
dimension 1
60
40
20
0
−20
−40
−40
−20
0
20
40
60