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

Student's Name: Tarun Singla **Mobile No: 8872526396**

Roll Number: b19198 **Branch:**EE

1 a.

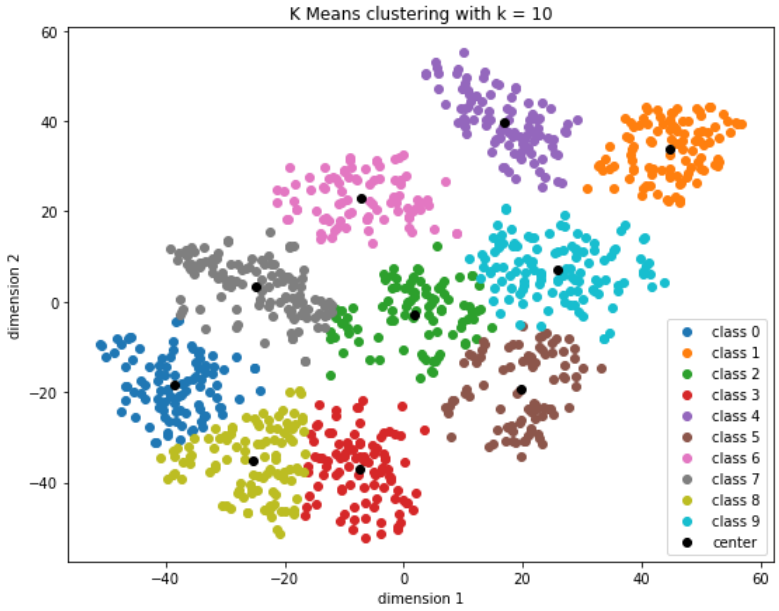


Figure 1 K-means (K=10) clustering on the mnist tsne training data

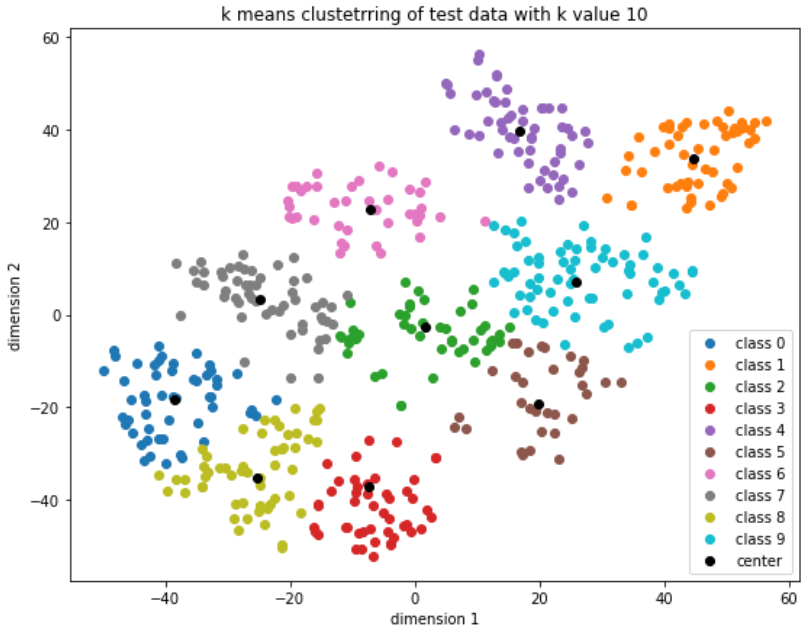
Inferences:

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. K-means algorithm assumes cluster boundaries to be circular in 2D. but the boundaries are not circular. There are not enough data points to form circle.
3. Boundary between class-0, class-3 and class-8 is linear in shape.
4. This is due to presence of outlier. K means clustering is sensitive to outliers

b.

The purity score after training examples are assigned to the clusters is .690

c.



\

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

Inferences:

1. Seeing both clusters we can see there is not much difference.

d.

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

Inferences:

1. Purity of train data is higher than test data. In test data we are predicted based on training data due to outliers its purity score is low. It also assumes circular data
2. It is sensitive to outliers.

2 a.

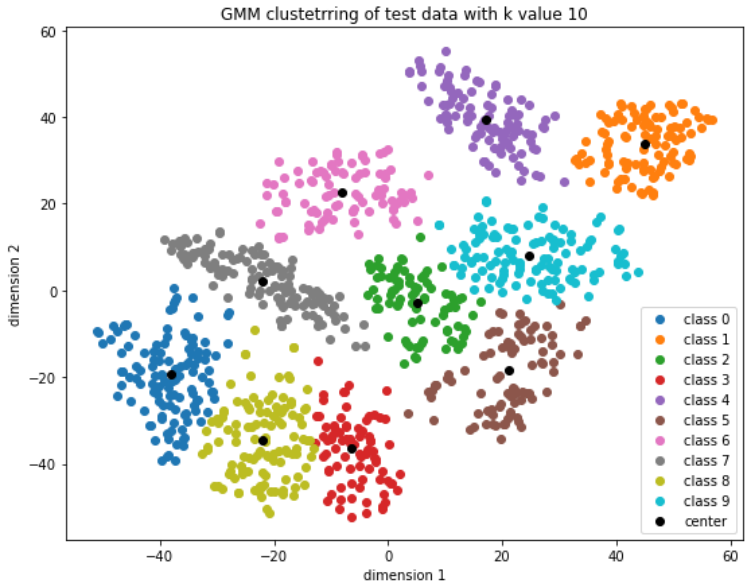


Figure 3 GMM clustering on the mnist tsne training data

Inferences:

1. In this we use mean and covariance to represent cluster and Expectation maximum is used to predict parameters.
2. No boundary is not elliptical. It assumes gaussian distribution and also due to presence of outliers.
3. No there is no observable difference between 1a and 2a. both type of clustering yield almost same clusters

b.

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

c.

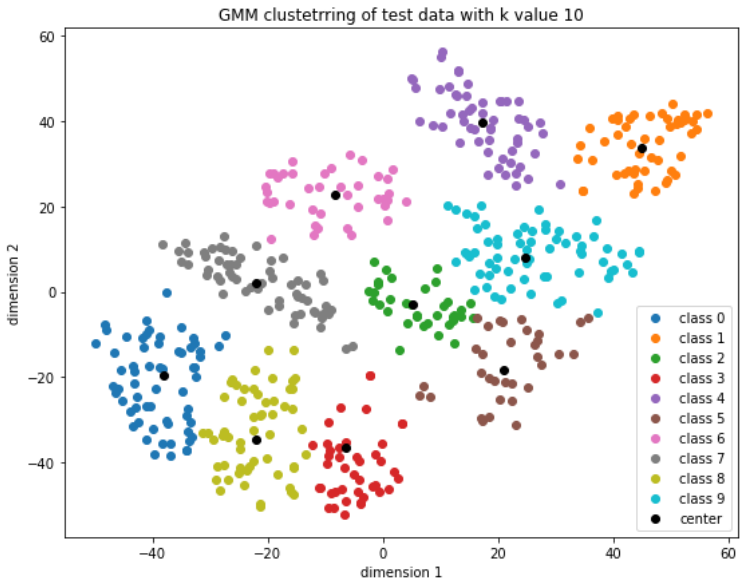


Figure 4 GMM clustering on the mnist tsne test data

Inferences:

1. There is not much difference between in 2a and 2b.

d.

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

Inferences:

1. Purity score of train data is higher because we are predicting based on training data. Due to presence of outlier higher data points are predicted wrong
2. It assumes gaussian distribution and it is computationally very expensive

3 a.

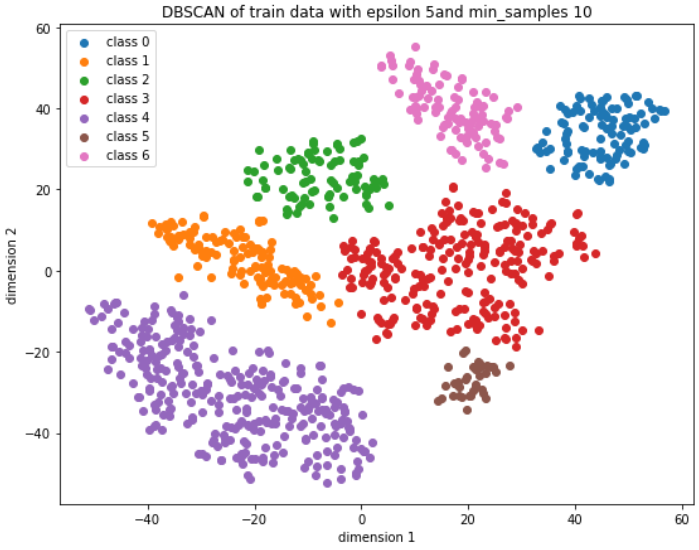


Figure 5 DBSCAN clustering on the mnist tsne training data

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 0.602

c.

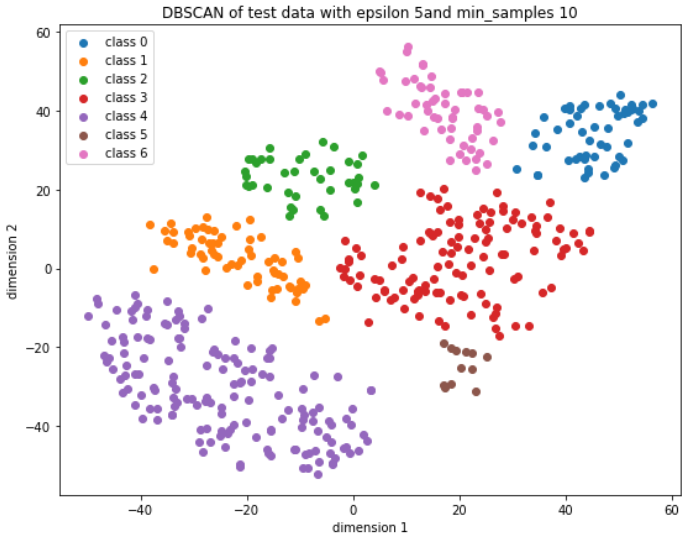


Figure 6 DBSCAN clustering on the mnist tsne test data

Inferences:

1. There is not much difference in the clustering of test and train data as we are using train data to predict test data.

d.

The purity score after test examples are assigned to the clusters is 0.586

Inferences:

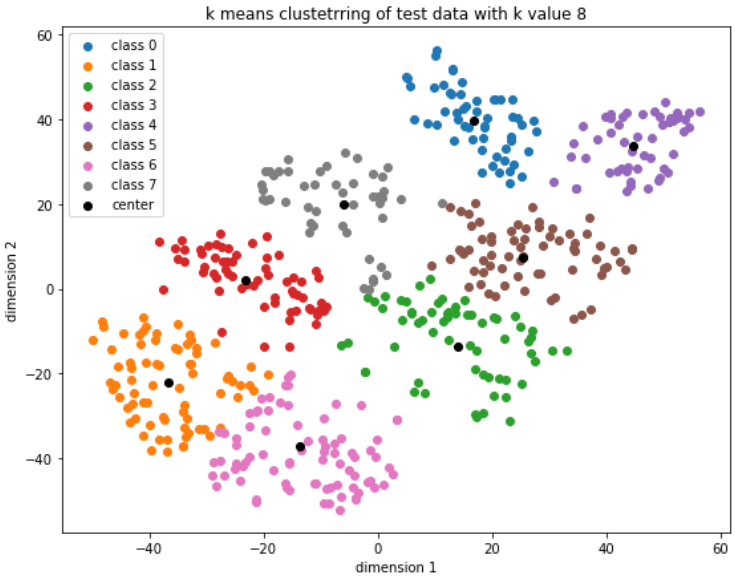
1. Purity score of train data is high then test data. As it randomly assigns test sample if number of training example of two or more clusters in eps radius are eps
2. This Clustering is not suitable if there are very much dense points.
3. It has low purity score as compared to GMM and K means this is because DBSCAN is not suitable much when density of data points is low.

Bonus Questions

1a For K Means Clustering and GMM for different values of k

Train Data: -

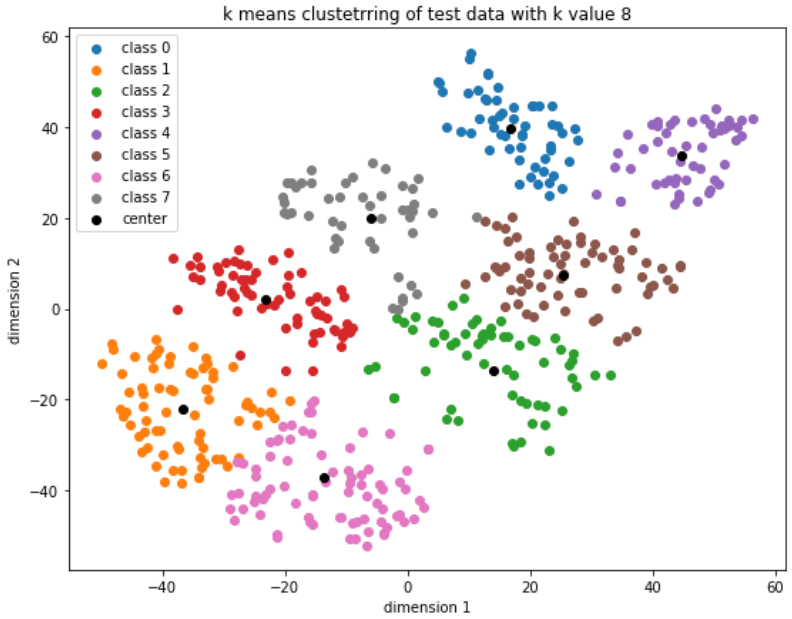
Best plot for train data is for k means for train data for K = 8



1. Purity score is maximum with k = 8.

2. Change in purity score is uneven with change in value of K. It first increased till 8 and then start decreasing.

Best plot for test data is for k means for train data for K = 8



1. Purity score is maximum with k = 8.
2. Change in purity score is uneven with change in value of K. It first increased till 8 and then start decreasing
3. There is also not much difference in clustering of test and train data.

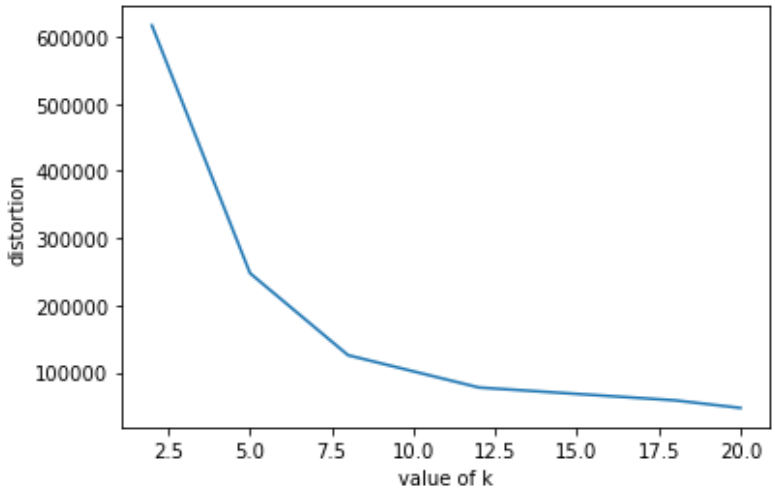
Purity score of K Means clustering of train data for different values of k

K value	Purity score
2	0.20
5	0.393
8	0.63
12	0.611
18	0.481
20	0.432

Purity score of K Means clustering of test data for different values of k

K value	Purity score
2	0.20

5	**0.398**
8	**0.624**
12	**0.612**
18	**0.46**
20	**0.416**

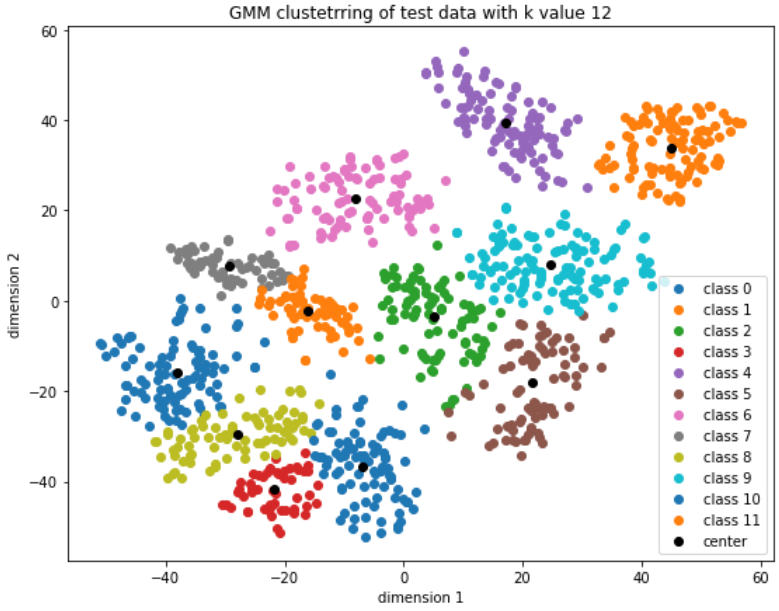


Distortion Measure for K value determination

1. Seeing the graph K = 8 is optimal value after 8 graphs is almost linear.
2. Distortion Measure decreases with increase value of K.

Best plot for test data is for GMM Clustering for train data for K = 12

Best plot for test data is for GMM Clustering for test data for K = 12



1. Value k = 12 has best purity score
2. Purity score first increase with increase in value of k then decrease.

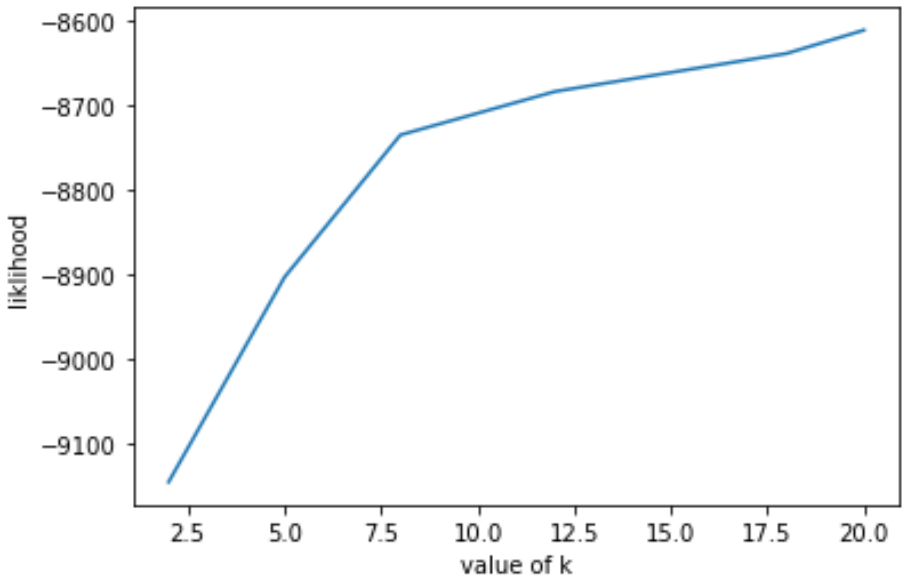
Purity score of GMM clustering of train data for different values of k

K value	Purity score
2	**0.20**

5	**0.46**
8	**0.629**
12	**0.66**
18	**0.508**
20	**0.455**

Purity score of GMM clustering of test data for different values of k

K value	Purity score
2	**0.20**
5	**0.448**
8	**0.628**
12	**0.646**
18	**0.51**
20	**0.46**



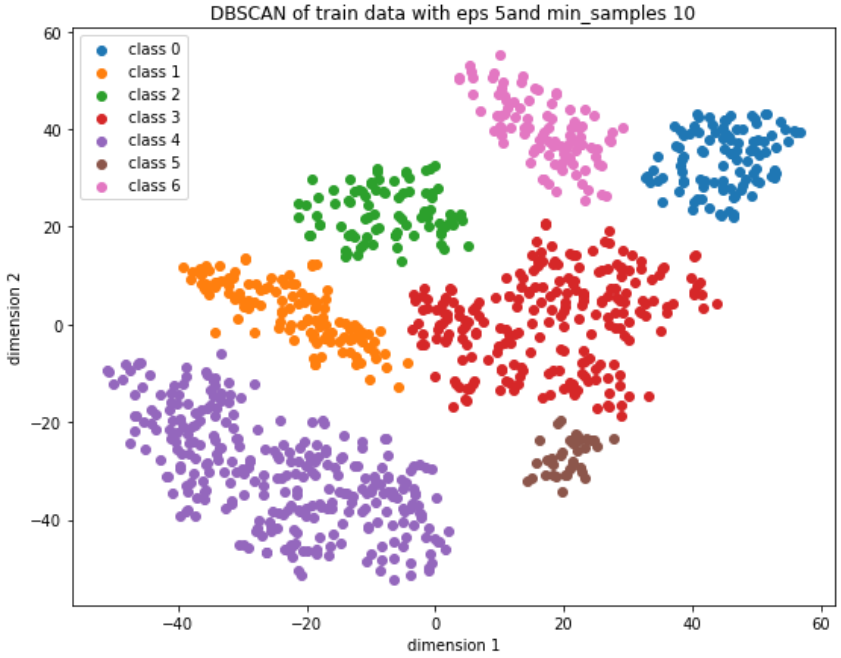
Log likelihood Measure

1. Total likelihood seems to be convergent after k=12.
2. Purity score is also optimum for k=12.
3. On increasing values ‘K’, purity score first increases then decreases due to overfitting

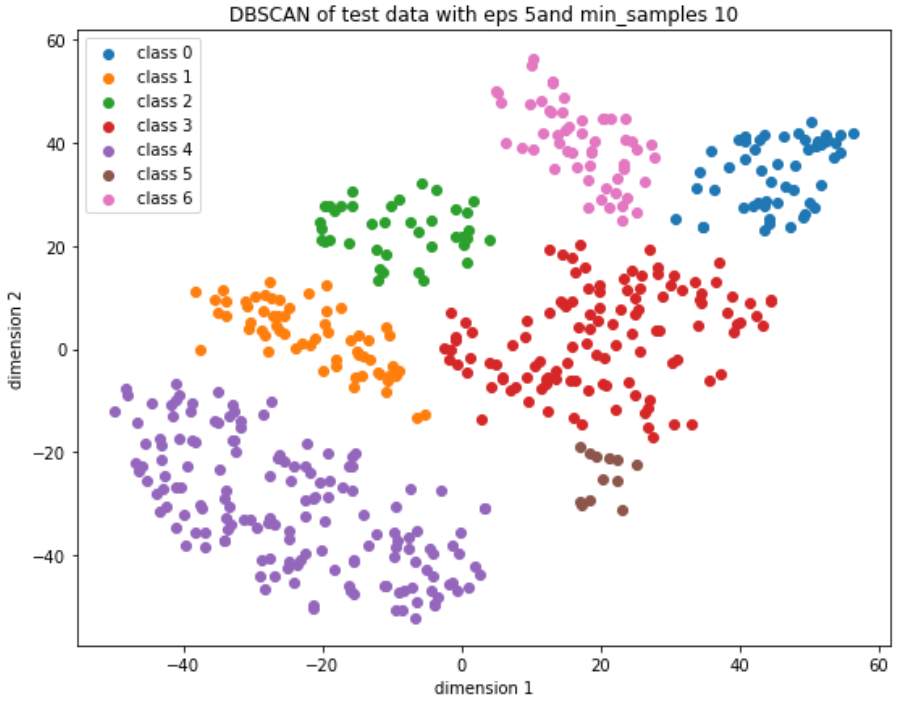
3..**1b**

DBSCAN for by varying Epsilon

Best plot for train data with varying epsilon = 5 and min samples = 10



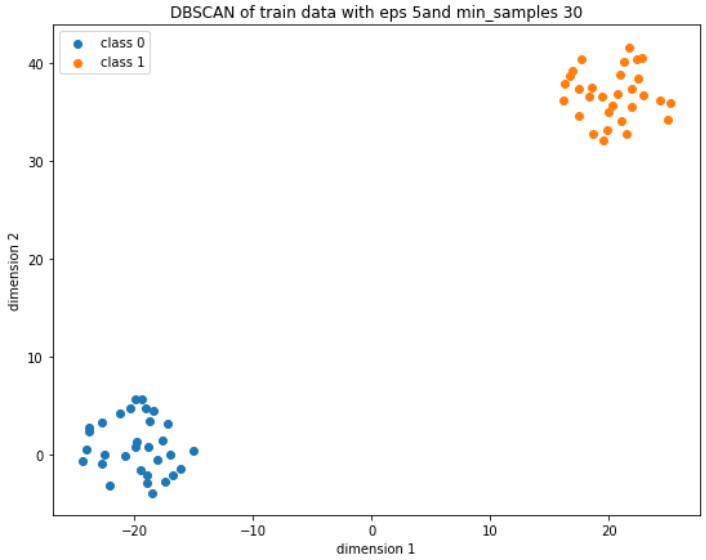
Best plot for test data with varying epsilon = 5 and min samples = 10



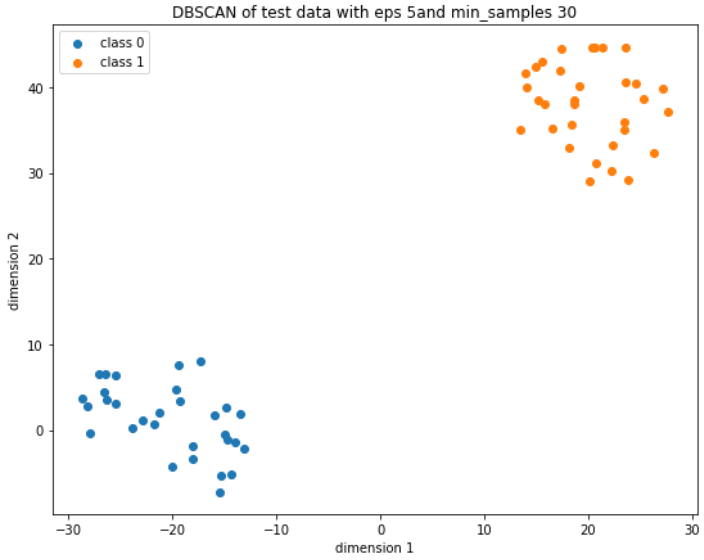
1. Min Samples = 10 and epsilon = 5 has best purity score
2. Epsilon and min Samples are experimentally determined
3. There with very high and low value of epsilon will result in low value of purity

DBSCAN for by varying Min Samples

Best plot for train data with varying min samples = 30 and epsilon = 5



Best plot for test data with varying min samples = 30 and epsilon = 5



1. Maximum Purity is the min samples = 30
2. Very high values of min sample with result in declaring large number of samples as outlier and this will result in high purity.

Purity score of train data different values of epsilon and min samples

Epsilon	**Min samples**	**Purity score**
5	**1**	**0.208**
5	**10**	**0.602**
5	**30**	**0.950**
10	**10**	**0.10**

Purity score of test data different values of epsilon and samples

Epsilon	**Min samples**	**Purity score**
5	**1**	**0.20**
5	**10**	**0.586**
5	**30**	**0.887**
10	**10**	**0.10**

Guidelines for Report (Delete this while you submit the report):

- **The plot/graph/figure/table should be centre justified with sequence number and caption.**
- **Inferences should be written as a numbered list.**
- **Use specific and technical terms to write inferences.**
- **Values observed/calculated should be rounded off to three decimal places.**
- **The quantities which have units should be written with units.**