Negar Nejatishahidin
G01207447
HW3 (Part 1) - Iris Clustering

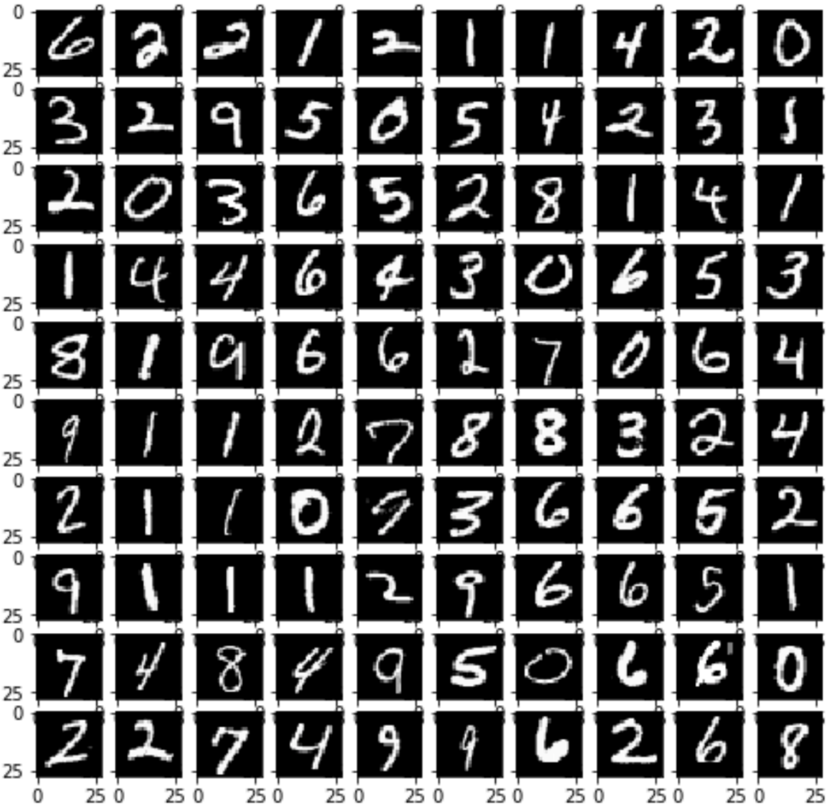
The Kmean is exactly the one which is described in the previous part of home work.

Digits clustering:

The objective in this assignment is to increase the accuracy of the Kmean clustering on the test dataset. It has 10740 records and 784 features.

In this part I first read the data as a panda data frame and then visualize the dataset to see what I am really try to cluster.

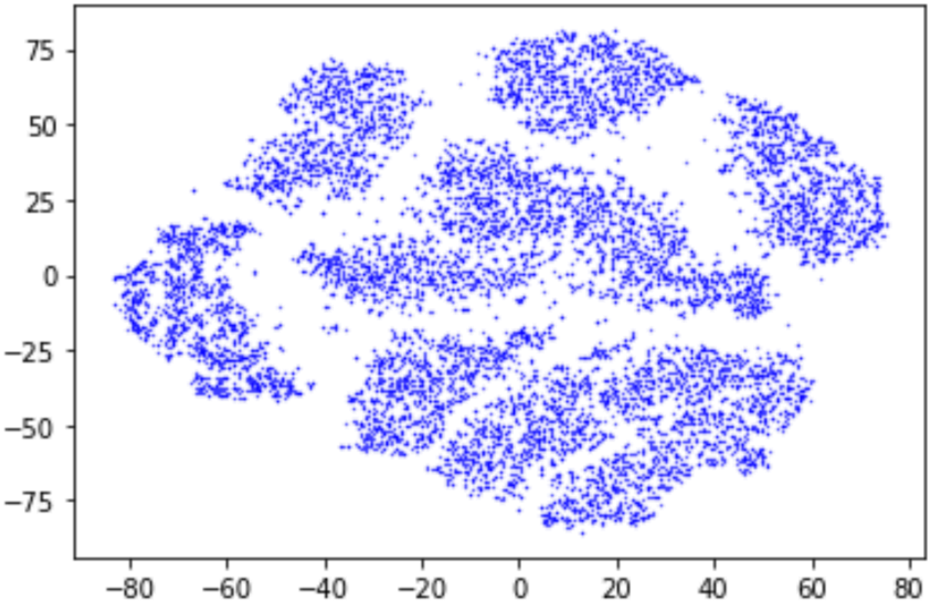
Hear is the visualization of data records, It is impossible to visualize the hole data because visualizing of 784 feature is impossible.



In this project I first just used the Kmans algorithm but the results was 48% so I understood the the preprocessing is necessary.

Preprocessing¶

For this part I first normalized the data. Then I used PCA and reduced the dimension to 200.
After that I normalized the data again and feed it to the T-SNE. The output has 2 dimension so it is possible to visualize it in 2D .



You can know see that 10 classes are separable know.

Kmean Implementation :

have used my Kmean clustering algorithem with K= 10 and 100 times iteration. Since the first centroid is random, I did this task for 100 times and finally use the one with the least error.

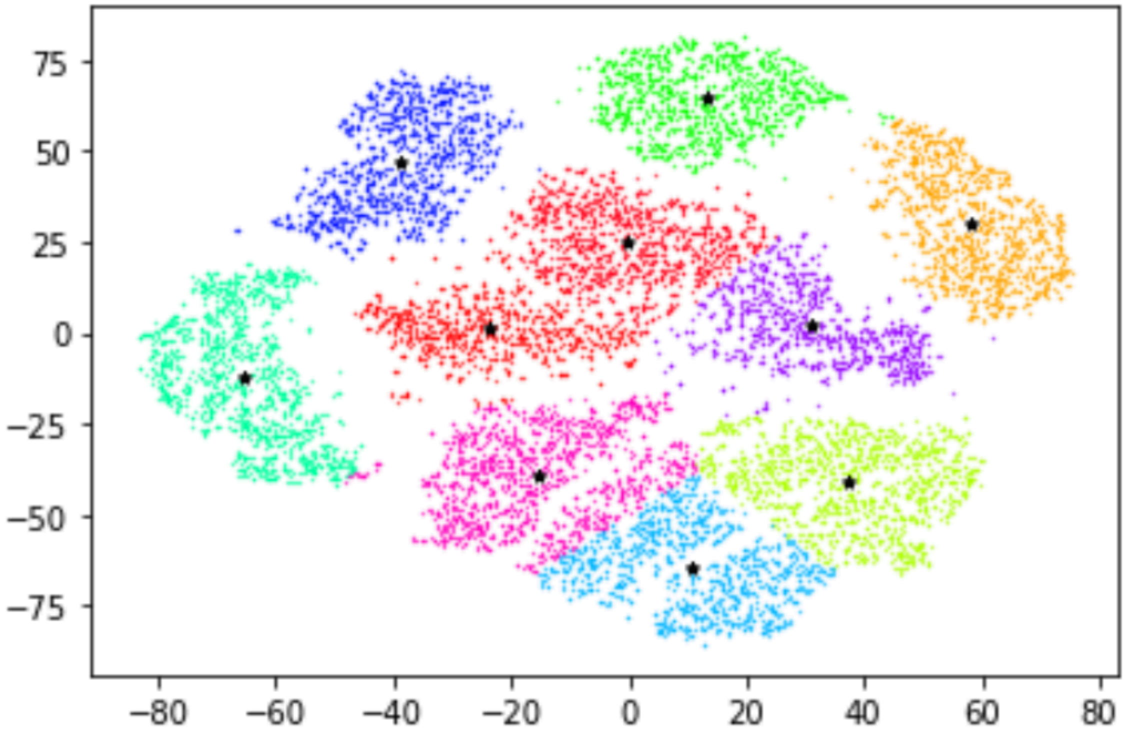
Run Time :

The whole KMeans algorithem for this dataset takes

accuracy:

The accuracy is 76%.

The result in 2D with centered is as follow :



It shows that the algorithm could not work perfectly in the bottom of the picture.