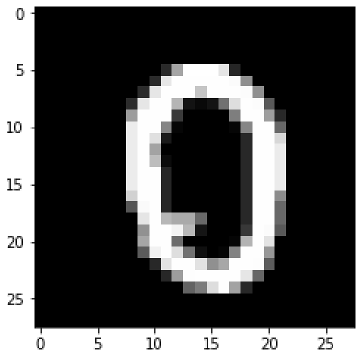
CS57300: Assignment 4

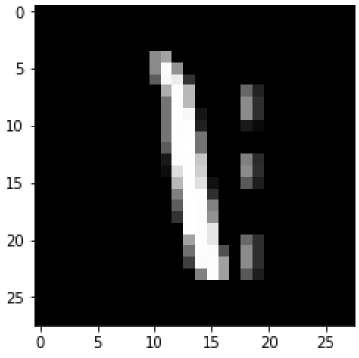
Pavani Guttula
pguttula@purdue.edu
late days used -1

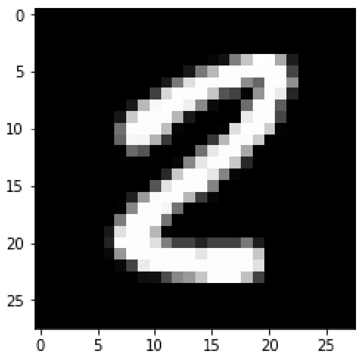
April 20, 2019

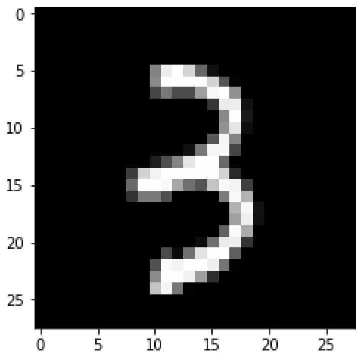
1 Exploration

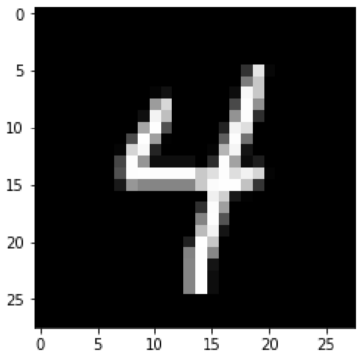
Below are the images of digits from 0-9.

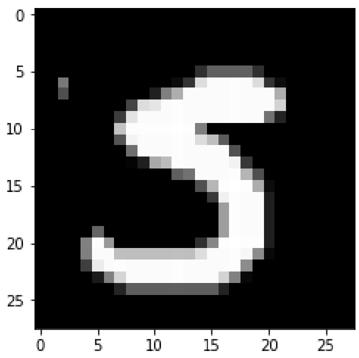


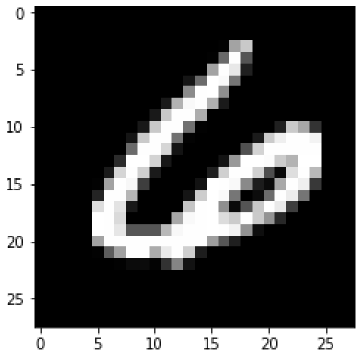


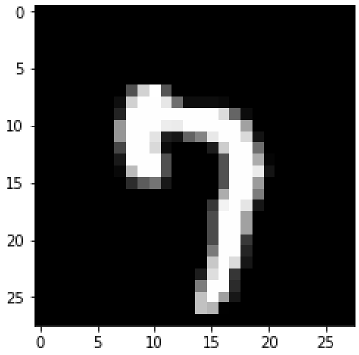

0
5
10
15
20
25
0
5
10
15
20
25

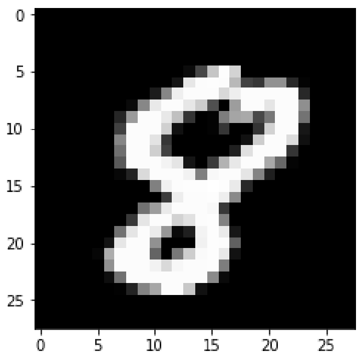

0
5
10
15
20
25
0
5
10
15
20
25

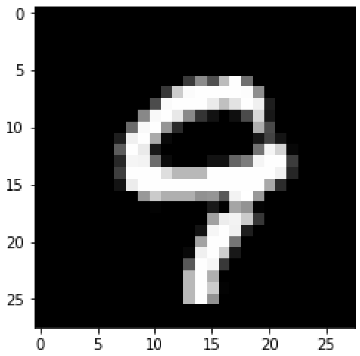

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25

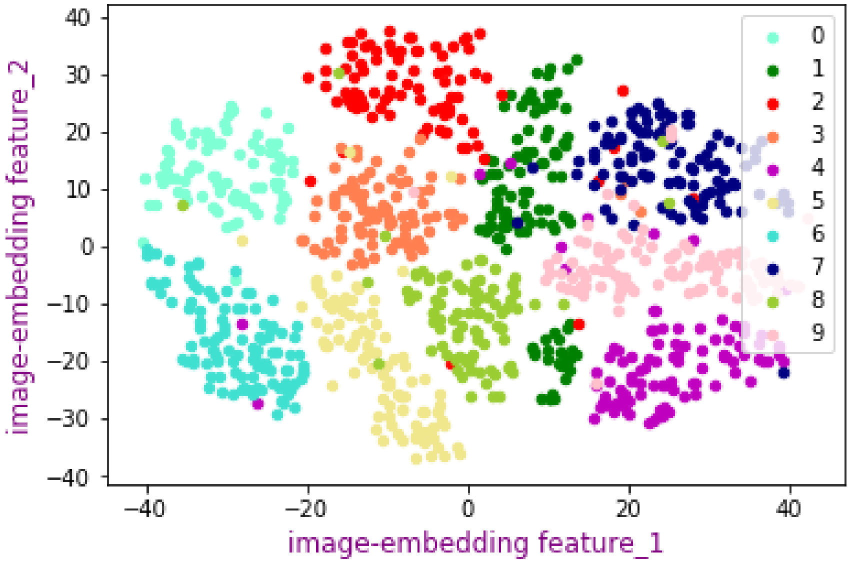

0
5
10
15
20
25
0
5
10
15
20
25


0
5
10
15
20
25
0
5
10
15
20
25





Below is the image showing clusters of data for all the class labels(0-9)



2 K-means Clustering

With a K value of 10 on the complete dataset digits-embedding.csv, below are the values for WC-SSD, SC and NMI.

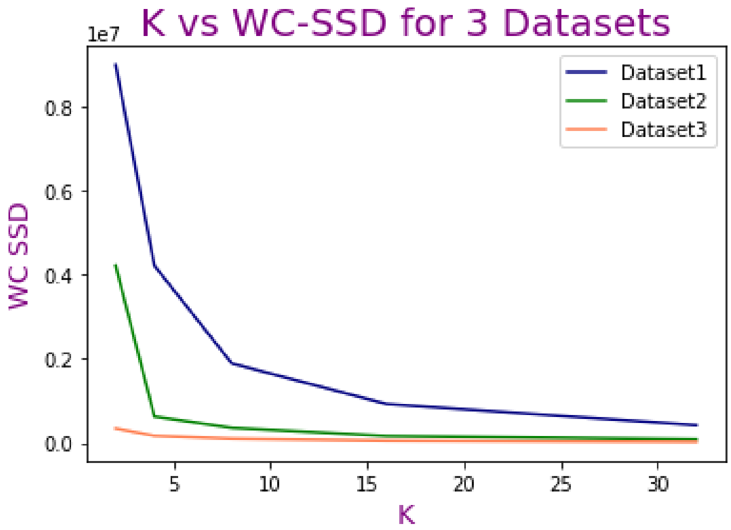
2.1 code

- WC-SSD: 1433531.469
- SC: 0.712 NMI: 0.356

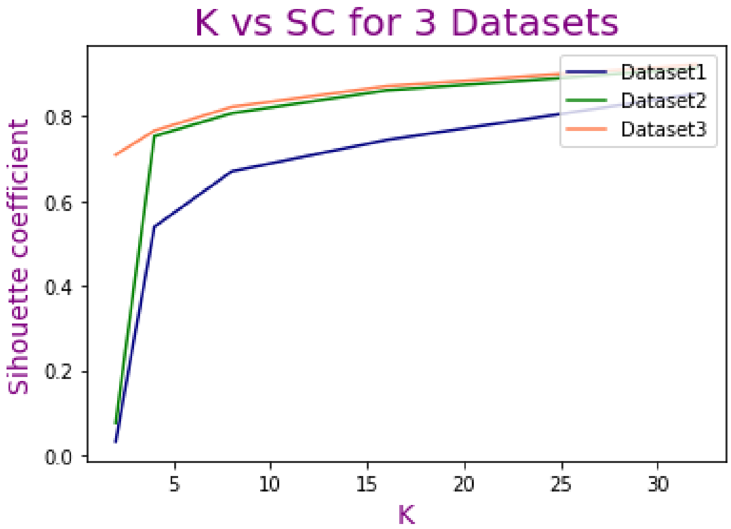
2.2 analysis

2.2.1

Below is the graph for WC-SSD values against different values of K plotted for different Datasets.



Below is the graph for SC values against different values of K plotted for different Datasets.



2.2.2

We decide the optimal K values based on elbow in the above plots in problem 2.2.1.
My K values would be 8 for Dataset 1, 4 for Datsaset 2 and 4 for Dataset 3.

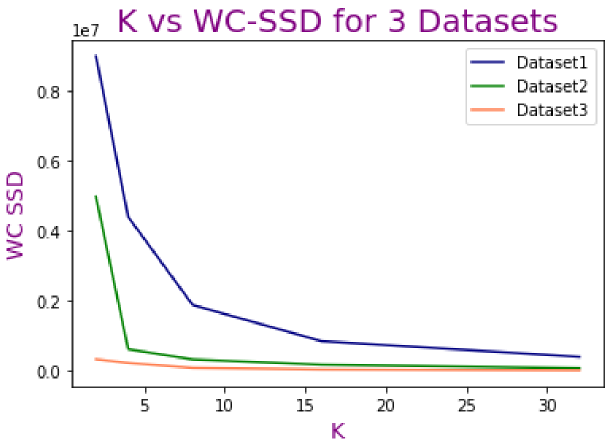
WC-SSD - Observations:
For Dataset1 and Dataset2, WC-SSD values decrease exponentially with an increase in the value of K till K= 16 and then reduction reduces making the graph constant. Clearly,with an increase in the number of clusters, the distance of the points to the nearest cluster reduces.

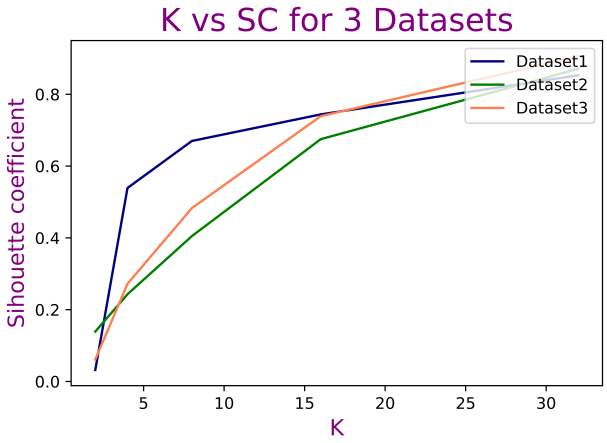
For Dataset 3, WC-SSD value does not change much as overall we have only 2 clusters in the dataset and making more and more clusters would not give us any optimization.

SC- Observations:
Again, for Dataset1 and Dataset2, we have an almost exponential increase in SC values till k=8 and then a gradual increase for the rest of the K values. This would be because with less number of clusters, we would not have much difference in distance between points in same cluster and different clusters.

For Dataset 3, we can see a gradual increase in SC values with an increase in K.

2.2.3





I ran k-means algorithm for different K sizes and random seeds ranging from 0 to 9. For all the three datasets, both WC-SSD and SC values remain almost the same for all the 10 seeds. Hence mean values for this problem are almost the same as the WC-SSD, SC values we get with seed = 0.
Standard deviation is zero for all the the 3 datasets obviously because of no changes in the WC-SSD and SC values with a change in seed.
WC-SSD values:

Dataset1 - 8: 1887621.7327022392, 16: 855453.19122656272, 2: 8983899.9995161984, 4: 4398977.8396483399, 32: 407945.05260677205
Dataset2 - 8: 335733.13198965101, 16: 184248.51627225231, 2: 4975048.3899274319, 4: 623865.31116823317, 32: 84269.728725130451
Dataset3- 8: 94277.207765708881, 16: 48931.980593146582, 2: 340372.41942807235, 4: 234388.57241930984, 32: 25965.939307835339

Same is the case with my SC values. They are almost similar to my SC values with random seed 0 with a nearly 0 standard deviation between different runs of seeds.
SC values:
Dataset1 - 8: 0.6740303499292396, 16: 0.77343105262960476, 2: 0.37361139635267121, 4: 0.53308193197312703, 32: 0.83906047640310122
Dataset2 - 8: 0.40535111814245672, 16: 0.67456627438925243, 2: 0.24333322000556451, 4: 0.13890974696835429, 32: 0.87075588140486637
Dataset3- 8: 0.48314883508933093, 16: 0.73893152637155646, 2: 0.061446159885829736, 4: 0.27191146937014843, 32: 0.90646480230767212

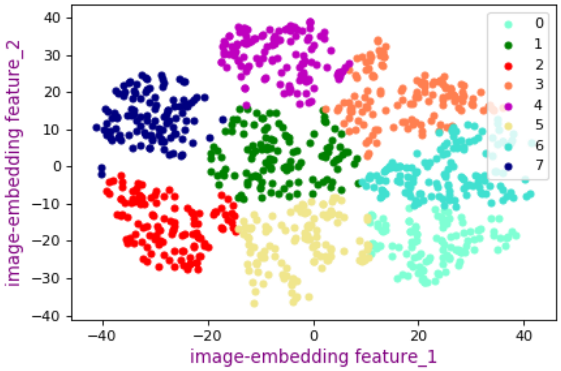
Hence, we can say that K-means sensitivity is null to initial starting conditions in this case. In other words, changing the initial random seed values make no difference to the k-means.

Hence, we can say that K-means sensitivity is null to initial starting conditions in this case. In other words, changing the initial random seed values make no difference to the k-means.

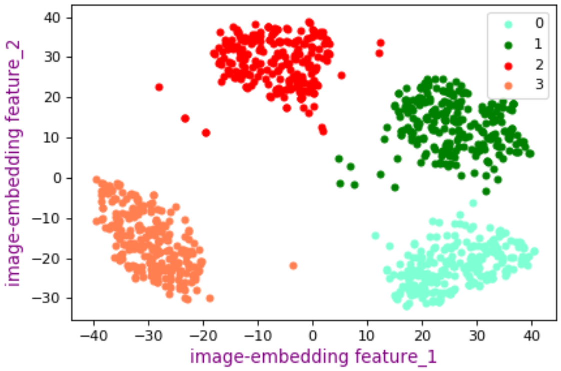
2.2.4

- NMI for Dataset 1 with K =8 : 0.347
- NMI for Dataset 2 with K =4 : 0.1796
- NMI for Dataset 3 with K =4 : 0.0

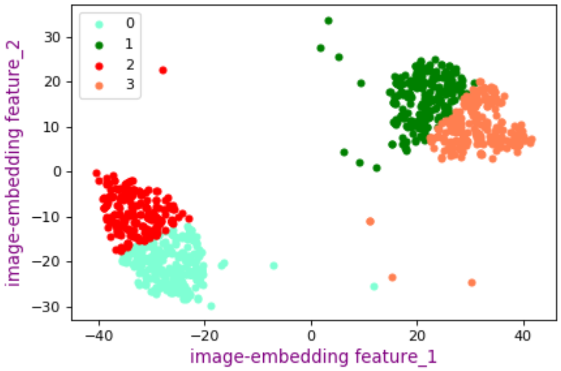
Below is the graph for visualization of data for Dataset1:



Below is the graph for visualization of data for Dataset2:



Below is the graph for visualization of data for Dataset3:

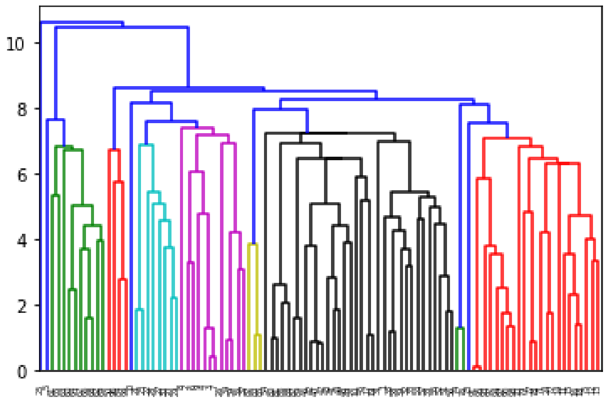


We can see that the NMI values align with the visualization of clusters. In the first dataset NMI is high since the mutual information between clustering is high. In the third dataset for example, two pairs of clusters are so closely aligned with each other that the mutual information between is the clusters is bill.

3 Hierarchical Clustering

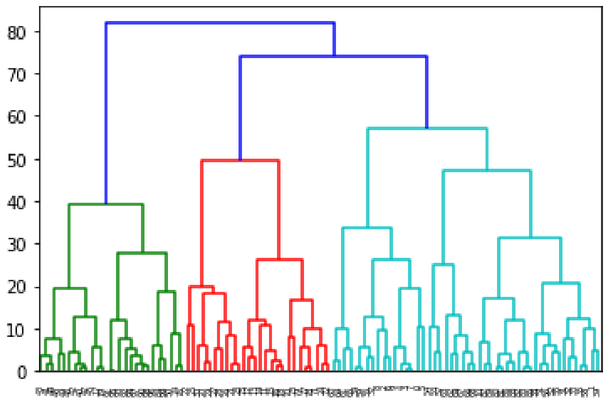
3.1

Dendrogram for Single Linkage

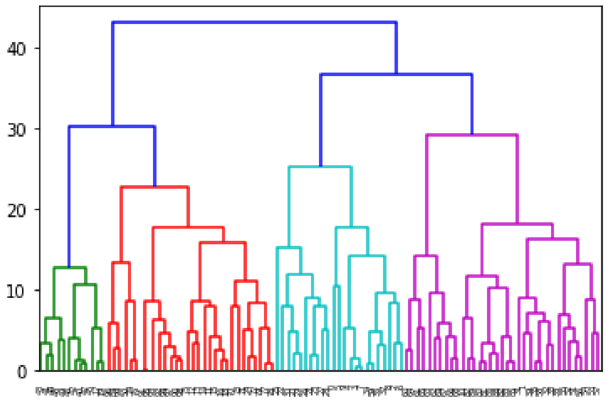


3.2

Dendrogram for Complete Linkage

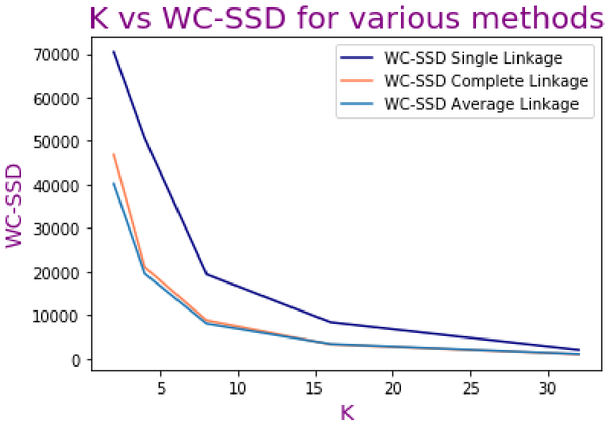


Dendrogram for Average Linkage

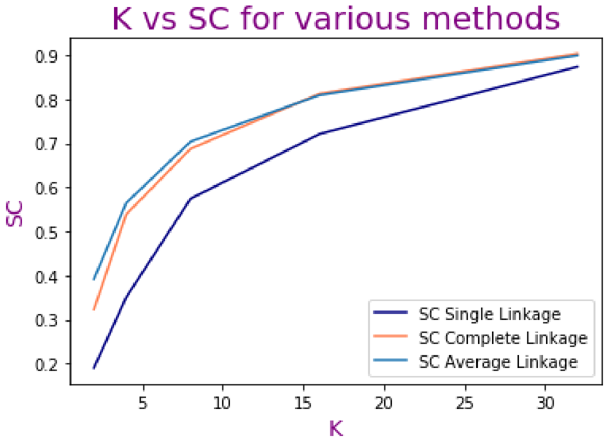


3.3

Below plot shows the WC-SSD values against K values for all three methods single linkage, complete linkage and Average linkage:



Below plot shows the SC values against K values for all three methods single linkage, complete linkage and Average linkage:



3.4

I would choose the below K values based on elbow of the plot.

- K for single linkage:8
- K for complete linkage :8
- K for Average linkage: 8

My K value is same as that of the K value I chose in part 2.

3.5

NMI values for the three methods:

- NMI for K = 8 for single method is 0.28
- NMI for K = 8 for complete method is 0.35
- NMI for K = 8 for average method is 0.37

  NMI for all the three methods is very close to the NMI of dataset 1 from k-means.