CS 584 - Theory and Application of Data Mining Spring 2019
HW 3 – Part 2

Name : Akshaya Damodaran
G No. : G01129364

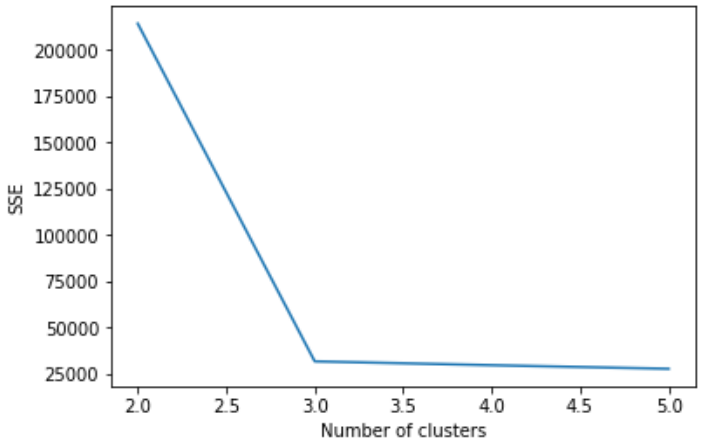
April 4, 2019

Q.1

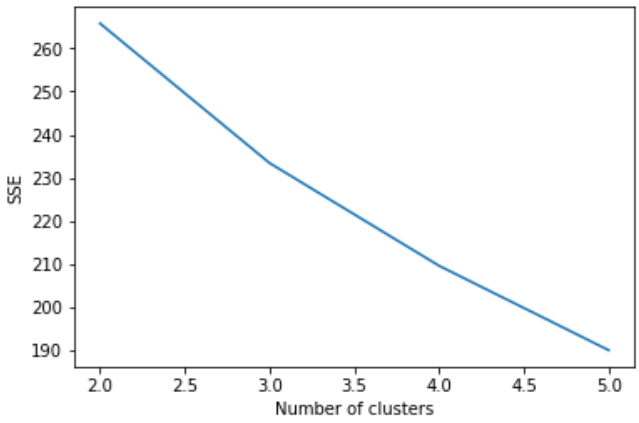
Given the two datasets, read as 'dataset1.txt' and 'dataset2.txt' in the code, clustering was done on each using K-Means (for k in [2,5]), EM and DBSCAN methods.

Q.2

By plotting the SSE for each value of k in the range 2 to 5, it was found that the elbow appears for k=3 and beyond k=3, the SSE curve remains almost constant. Therefore, k=3 is the best value of k for dataset1.



However, for dataset2, we can see that for k in the range 2 to 5, the SSE curve does not have a sharp elbow. We can therefore infer that dataset2 might have a very large number of clusters i.e. it does not exhibit proper clustering structure.



Q.3

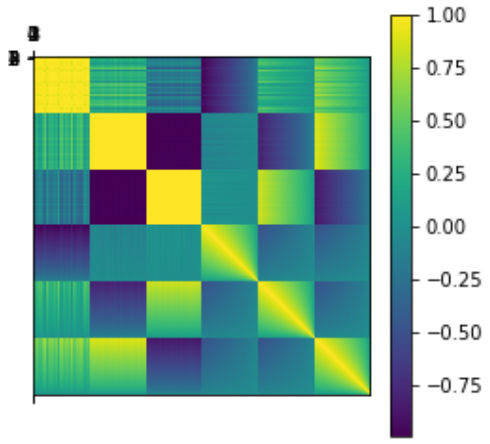
Clustering validation was performed by finding correlation of two matrices –

- The proximity (distance) matrix
- The incidence matrix
    - One row and one column for each data point
    - An entry is 1 if the associated pair of points belong to the same cluster
    - An entry is 0 if the associated pair of points belongs to different clusters

These two matrices are created for each dataset and clustering method and the correlation between these two matrices is computed.

K = 3 for dataset1.txt

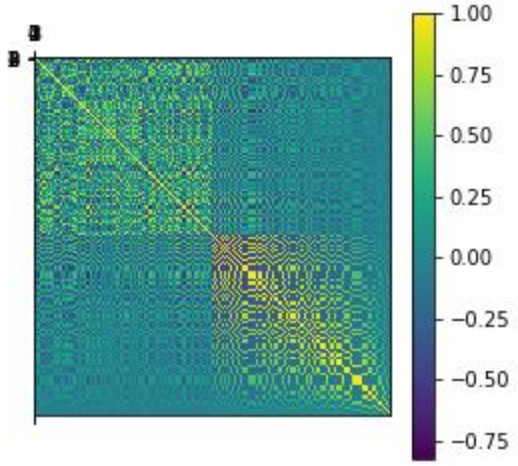
Clustering dataset1.txt using K Means
Plot of correlation:



Silhouette Score = 0.7876425008781422

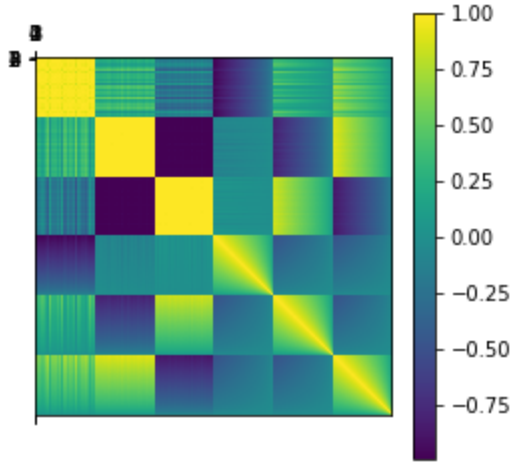
K = 5 for dataset2.txt

Clustering dataset2.txt using K Means
Plot of correlation:



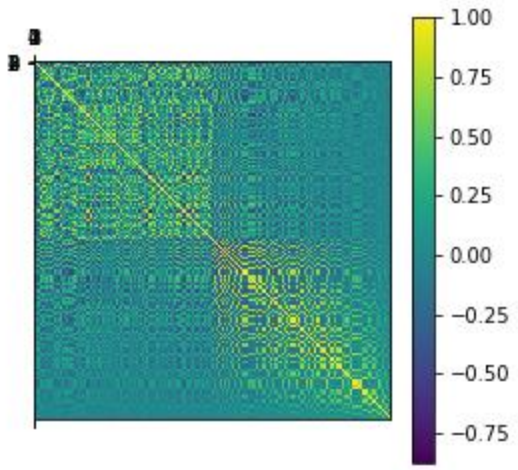
Silhouette Score = 0.1479054990959254

Clustering dataset1.txt using EM
Plot of correlation:



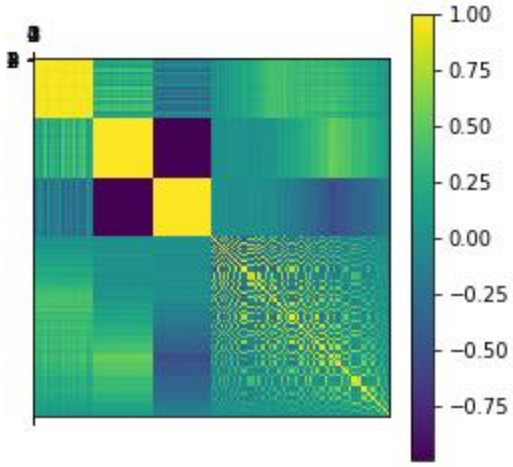
Silhouette Score = 0.3695063756086118

Clustering dataset2.txt using EM
Plot of correlation:



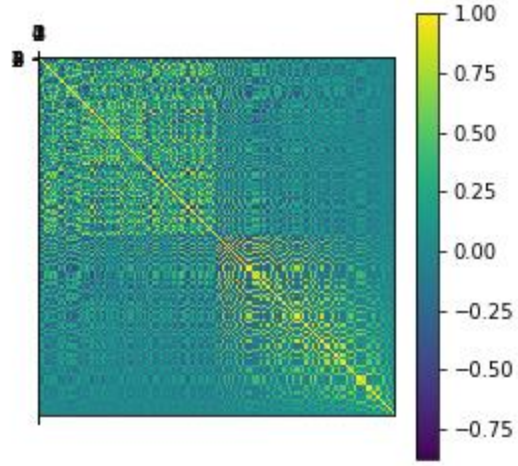
Silhouette Score = 0.1248272569143979

Clustering dataset1.txt using DBSCAN
Estimated number of clusters: 3
Plot of correlation:



Silhouette Score = 0.6382490337456681

Clustering dataset2.txt using DBSCAN
Estimated number of clusters: 14
Plot of correlation:



Silhouette Score = -0.34234868082524816

Q.4

High correlation indicates that points that belong to the same cluster are close to each other and this can be seen on the heat map plotted for the correlation between the proximity and incidence matrices.

For each of the clustering methods, we can clearly see from the heat maps that dataset1 shows a better clustering structure that dataset2.

The Silhouette Coefficient is calculated using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample. The Silhouette Coefficient for a sample is (b - a) / max(a, b). The

best value is 1 and the worst value is -1. Values near 0 indicate overlapping clusters. Negative values generally indicate that a sample has been assigned to the wrong cluster, as a different cluster is more similar.
Dataset1 has a high Silhouette Score in all the methods of clustering and therefore, Dataset1 exhibits a good clustering structure unlike dataset2.