Haorui Chen
116165549

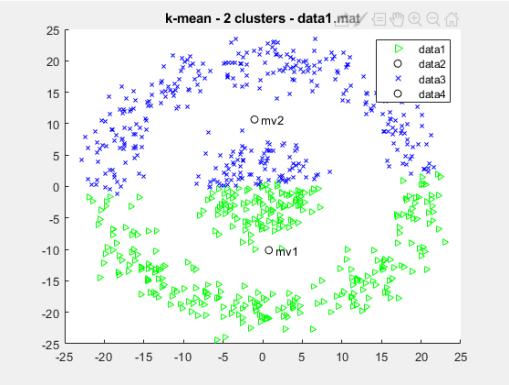
ENEE439M/ENTS669D

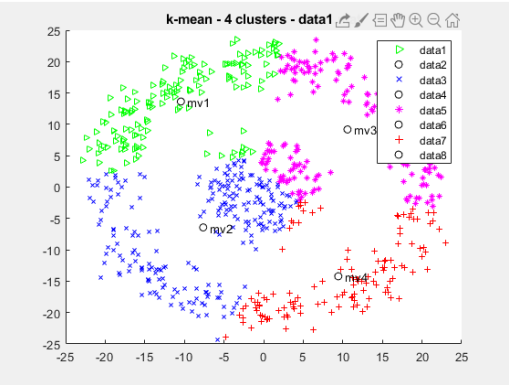
Homework 5 **Due 05/01/2019**

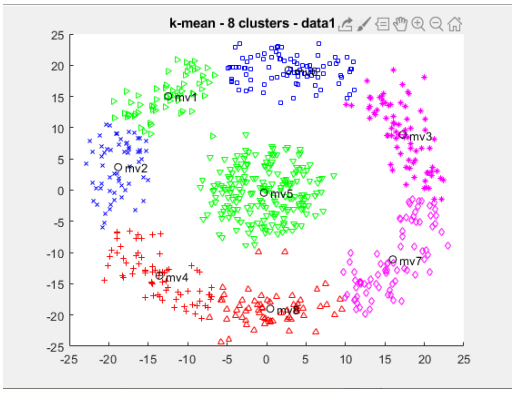
For these problems, you are give two synthetic datasets: `data1.mat` containing 700 samples of 2-dimensional features, and `data2.mat` containing 1400 samples of 2-dimensional features.

Problem 1. Implement the K-means algorithm over MATLAB, and cluster the samples in both datasets using $C = 2, 3, \cdots, 8$ clusters. Plot the members of each cluster using different symbols, e.g., 'x', 'o', etc., for $C = 2, 4$ and 8 only.

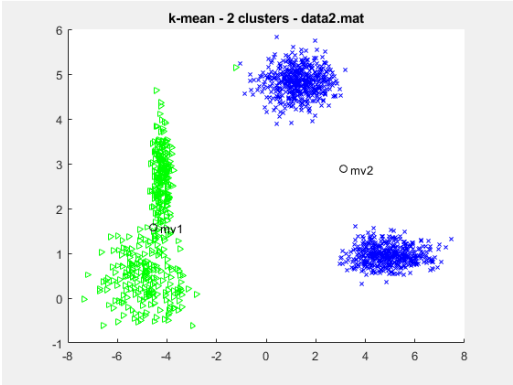
>>For data1.mat, in clusters = 2, 4, 8.

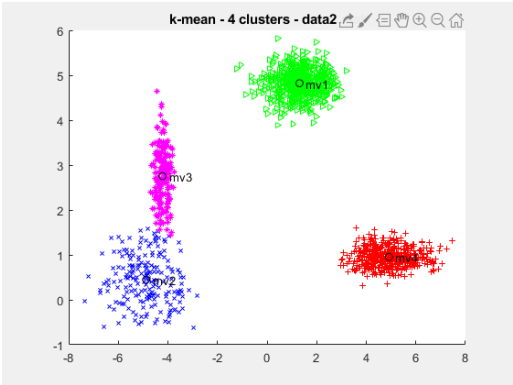


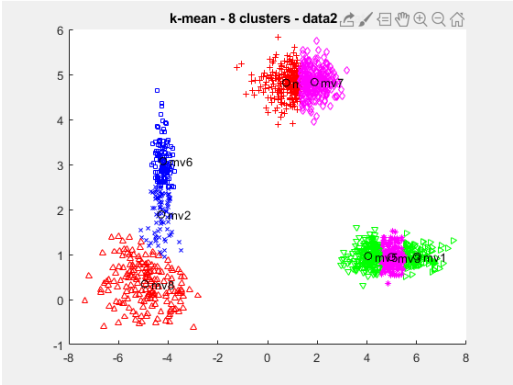




>>For data2.mat, in clusters = 2, 4, 8.







Haorui Chen
116165549

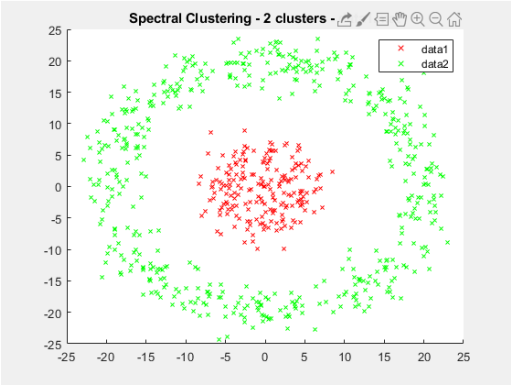
Problem 2. Implement the spectral clustering algorithm over MATLAB for dividing the data into 2 clusters. Use the similarity measure given by:

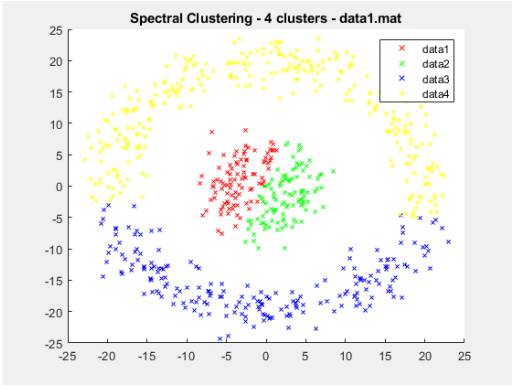
$$w_{ij} = \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{10}\right).$$

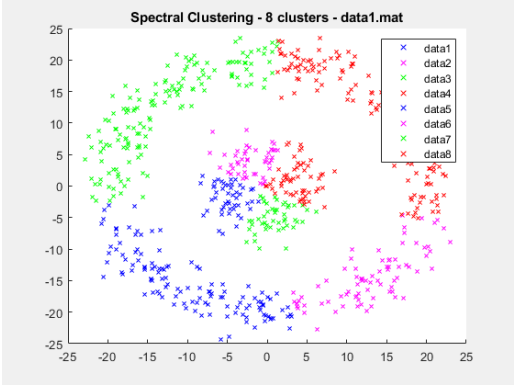
Recall that for spectral clustering you need to first find the eigenvector $\mathbf{z}_1$ *corresponding to the second-smallest eigenvalue of* $\mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{W})\mathbf{D}^{-1/2}$. *Then, the indices of the positive components of* $\mathbf{y}_1 = \mathbf{D}^{-1/2}\mathbf{z}_1$ *correspond to the first cluster, and the rest correspond to the other cluster.*

For each data set, first cluster the data into 2 clusters ($C = 2$), and then cluster each of the previous clusters into 2 clusters ($C = 4$), so on and so forth until you get $C = 8$ clusters. Plot the members of each cluster using different symbols, e.g., 'x', 'o', etc.

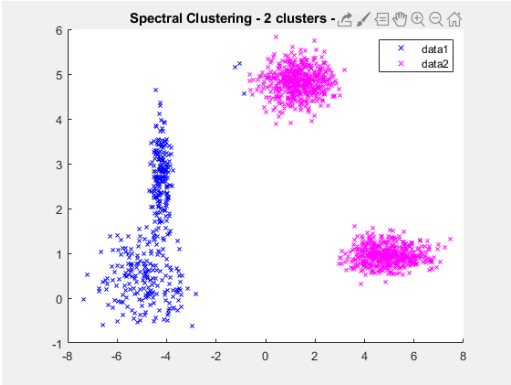
>>For data1.mat, in clusters = 2, 4, 8.

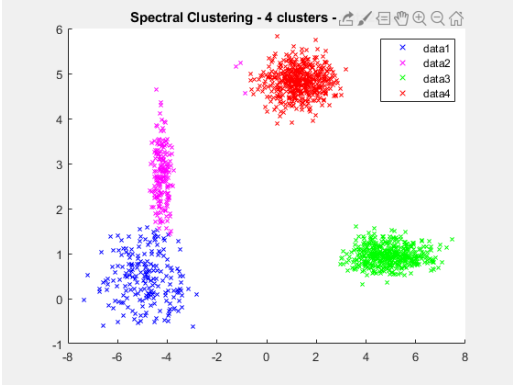


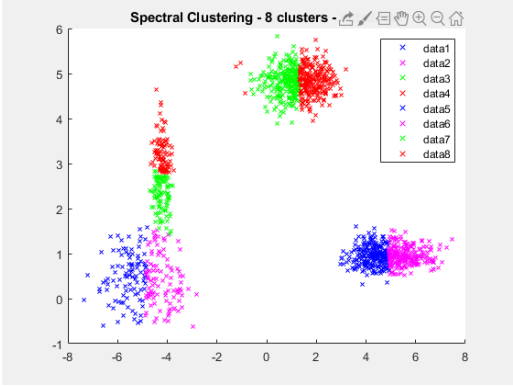




>>For data2.mat, in clusters = 2, 4, 8.







Haorui Chen
116165549

Problem 3. Using the cost functions J_e for the sum-of-squared-error and determinant criteria, compare the clustering quality of K-means (Problem 1) and spectral clustering (Problem 2) over each dataset, for $C = 2, 4$ and 8, and determine which algorithm works best for a given C. How do these results compare with the visual (qualitative) assessment of the two clustering outputs?

Compare with the visual figure output and cost function return value. Spectral Clustering works better than k-means for data. Result of J_e shows that spectral algorithm works better than k-means, especially when encounter such dataset from data1.mat.

>>Problem3.mat output

```
Clear Memory & Command Window
>>Calculate 2 clusters for data1.mat

>>Je(k-mean) = 1.318658e+05, Je(spectral) = 1.195756e+05

------------------------------------------------------
>>Calculate 4 clusters for data1.mat

>>Je(k-mean) = 6.477889e+04, Je(spectral) = 4.371157e+04

------------------------------------------------------
>>Calculate 8 clusters for data1.mat

>>Je(k-mean) = 2.124898e+04, Je(spectral) = 1.695908e+04

------------------------------------------------------

Clear Memory & Command Window
>>Calculate 2 clusters for data2.mat

>>Je(k-mean) = 1.072556e+04, Je(spectral) = 4.264196e+03

------------------------------------------------------
>>Calculate 4 clusters for data2.mat

>>Je(k-mean) = 1.096738e+03, Je(spectral) = 8.922758e+02

------------------------------------------------------
>>Calculate 8 clusters for data2.mat

>>Je(k-mean) = 7.508868e+02, Je(spectral) = 2.787056e+02

------------------------------------------------------
```