ENGR 421 / DASC 521: Introduction to Machine Learning
Homework 08: Spectral Clustering

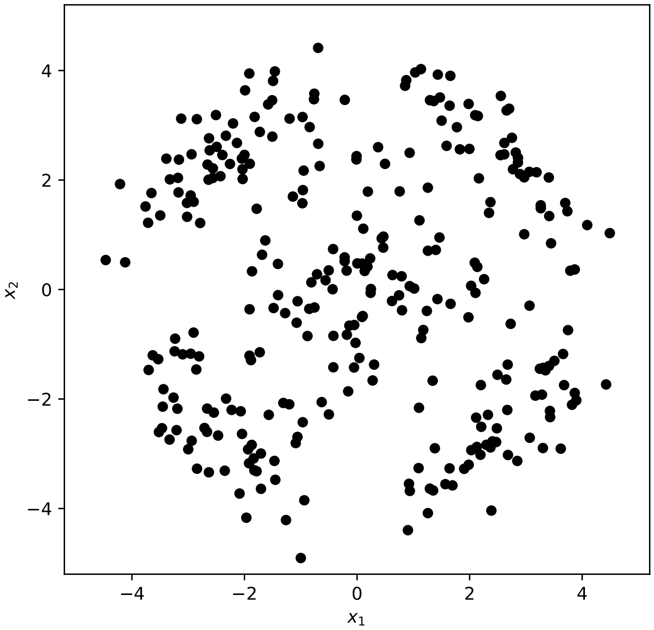
Deadline: January 3, 2022, 11:59 PM

In this homework, you will implement a spectral clustering algorithm in Python. Here are the steps you need to follow:

1. You are given a two-dimensional data set in the file named `hw08_data_set.csv`, which contains 300 data points generated randomly from five bivariate Gaussian densities with the following parameters.

$$\mu_1 = \begin{bmatrix} +2.5 \\ +2.5 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, \quad N_1 = 50$$
$$\mu_2 = \begin{bmatrix} -2.5 \\ +2.5 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, \quad N_2 = 50$$
$$\mu_3 = \begin{bmatrix} -2.5 \\ -2.5 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, \quad N_3 = 50$$
$$\mu_4 = \begin{bmatrix} +2.5 \\ -2.5 \end{bmatrix}, \quad \Sigma_4 = \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, \quad N_4 = 50$$
$$\mu_5 = \begin{bmatrix} +0.0 \\ +0.0 \end{bmatrix}, \quad \Sigma_5 = \begin{bmatrix} +1.6 & +0.0 \\ +0.0 & +1.6 \end{bmatrix}, \quad N_5 = 100$$

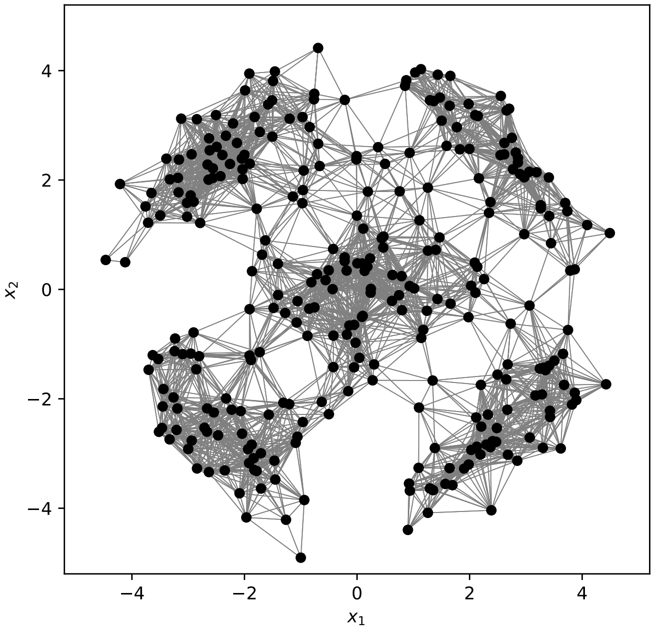
The given data points are shown in the following figure.



2. You should first calculate the Euclidean distances between the pairs of data points. The data point pairs with distance less than or equal to $\delta = 1.25$ are considered as connected. Construct the matrix **B** as follows:

$$b_{ij} = \begin{cases} 1, & \left\| \boldsymbol{x}_i - \boldsymbol{x}_j \right\|_2 < \delta \\ 0, & \text{otherwise.} \end{cases}$$
$$b_{ii} = 0$$

You should also visualize this connectivity matrix by drawing a line between two data points if they are connected. Your figure should be similar to the following figure.



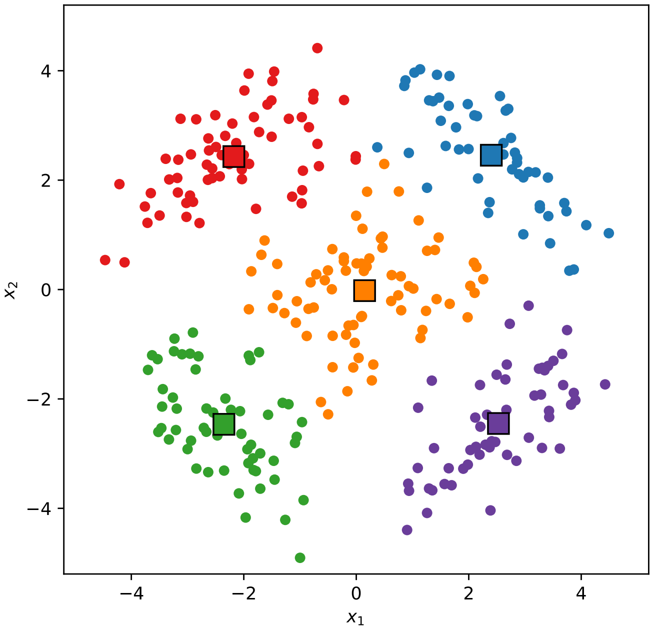
3. You should then calculate **D** and **L** matrices as described in the lecture notes. You should normalize the Laplacian matrix using the following formula:

   $$\mathbf{L}_{symmetric} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{B}\mathbf{D}^{-1/2}$$

4. Find the eigenvectors of the normalized Laplacian matrix and pick $R = 5$ eigenvectors that corresponds to R smallest eigenvectors (eigenvectors that corresponds to 2nd smallest, 3rd smallest, 4th smallest, 5th smallest and 6th smallest eigenvalues since the smallest eigenvalue is 0). Using these eigenvectors construct the matrix **Z** as described in the lecture notes. Please note that the eigenvalues might not be returned in a decreasing or increasing order from the eig function.

5. Run *k*-means clustering algorithm on **Z** matrix to find $K = 5$ clusters. When initializing your algorithm, use the following rows of **Z** matrix for initial centroids: 29, 143, 204, 271, and 277.

6. Draw the clustering result obtained by your spectral clustering algorithm by coloring each cluster with a different color. Your figure should be similar to the following figure.



What to submit: You need to submit your source code in a single file (.py file) and a short report explaining your approach (.doc, .docx, or .pdf file).

How to submit: Submit the two files (source code and short report) you created to Blackboard. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.