# DSME6650: Data Mining for Managers
# Assignment 1

***Due Date: 23:59 Jan.19, 2020***

Submission Instruction:

For this assignment, please submit electronic version only. For the programming questions, you need to submit BOTH your Python codes and your results. Submit codes as zipped tar file and the output of your program in plain-text file. For other questions, answer them in a word document. You should place the relevant files in their separate directory (preferable A and B for each part).

After that, please compress all your files as one zip named using your student id, e.g., 11xxxxxxxx.zip, and submit to Blackboard.

## Part A. Clustering

*The following problem is based on lecture 5.*

1. Practice the K-means algorithm using Manhattan distance.

   Suppose there are 15 data points:

- A1(1.0, 1.0), A2(1.0, 2.0), A3(2.0, 0.0), A4(2.0, 3.0), A5(5.0, 10.0),
- A6(5.0, 12.0), A7(6.0, 2.0), A8(6.0, 4.0), A9(6.0, 10.0), A10(7.0, 2.0),
- A11(7.0, 5.0), A12(9.0, 9.0), A13(10.0, 7.0), A14(10.0, 9.0), A15(11.0, 8.0).

Assume we cluster these 15 data points into 4 clusters. The initial seeds are A1, A5, A7, A12.

(1) Please draw all data points and the initial seeds in a figure with grid line in the range of x: [0, 12] and y: [-2, 14].

(2) After running the K-means algorithm for 1 epoch only, answer the following questions one by one:

   a. The new clusters (i.e., the examples belonging to each cluster)

   b. The centers of the new clusters

   c. How many more iterations are needed to converge? Please write down the final centers and plot the results.

(3) If we use A7, A8, A10, A11 as initial seeds, how many more iterations are needed to converge? Please write down the final centers and plot the results.

(4) If we apply K-mean++ to the data and choose A1 as the first seed, which point will have the highest probability to be chosen as the second seed?

For these problems, write down your answers for all questions in the word file. For question (2), (3) and (4), you need to write a program to simulate the process of K-means and put your codes in directory A which will be zipped into the final submitted file. Please note that you should write a **README.txt** file (in the directory A), which indicates how to run the program to get the results you provided (hyper-parameters of the model should be given if needed). Without an appropriate README.txt will be deducted certain scores.

## Part B. Graph Representation

*The following problem is based on lecture 6.*

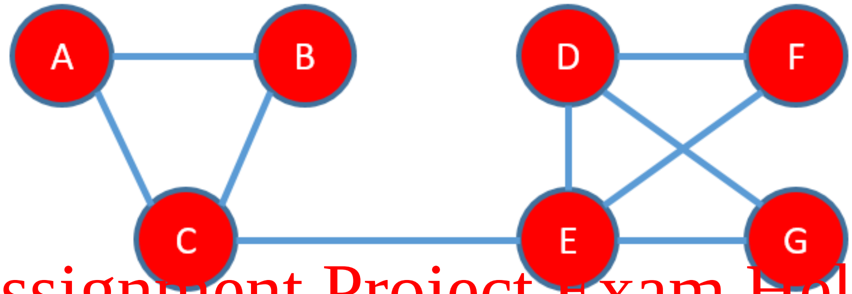

Figure 1: Figure for Graph Analysis.

Given a small graph shown in Fig. 1, please answer the following questions:

1. Compute the adjacency matrix, the degree matrix, the unnormalized Laplacian matrix, and the normalized Laplacian matrix.
2. You are required to implement a Spectral Clustering method to cluster the clusters of the given graph (k=2) on both unnormalized and normalized Laplacian matrix.

For question 1, write down your answers in a word file. For question 2, you need to write a program to simulate the process of spectral clustering and put your codes in directory B which will be zipped into the final submitted file. Please note that you should write a **README.txt** file (in the directory B), which indicates how to run the program to get the results you provided (hyper-parameters of the model should be given if needed). Without an appropriate README.txt will be deducted certain scores.