CS 3753 & 5163 Data Science Summer 2020
Homework 6 (60 points)

Submission:

1. submit a single python script (abc123_hw#.ipynb or abc123_hw#.py) through blackboard. All the results are outputted from your Python code.
2. You should have the instruction of running your code at the beginning of your code. It should run successfully either in the basic Python3 environment or in Jupyter Notebook.
3. If your code cannot run, we assume your code can run, then we will check whether your code is correct logically. If so, half points will be deducted. Otherwise, more points will be deducted if your code is wrong or there is no code.
4. Do not compress your files and make sure all your files are in the same folder. The compressed files will get a warning at the first time and will lose 10% points later.
5. You can submit your homework 3 times before the deadline. The late submission will lose 15% of the total points in the assignment.

Questions

1. (30 points) K-means clustering.

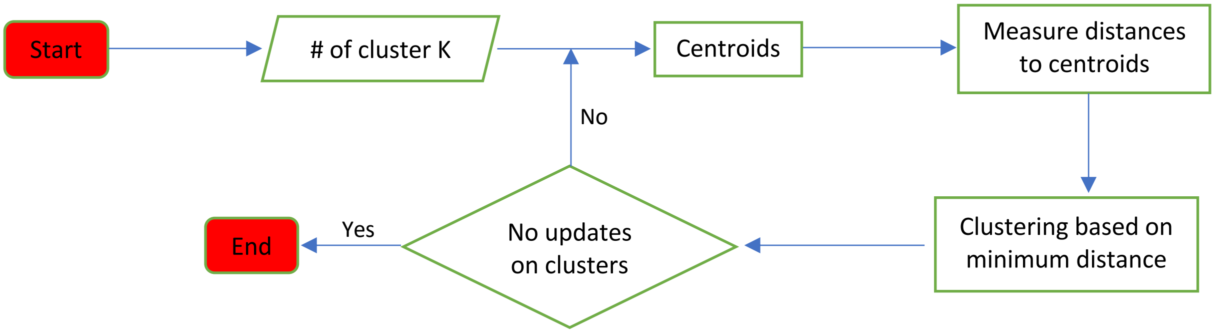
You do not need to import any libraries or modules about K-means clustering because you will implement it from scratch. The template of the code is provided, and you just need to write your code at specified locations with "your code is here".

Download the dataset 'k_means_clustering_data.csv' and save it into your working directory where we can find your source code about this homework. The dataset has two columns ('x' and 'y') and 42 records. They are 42 points in a 2D plane. Your goal is to group them into K clusters using K-means clustering algorithm.

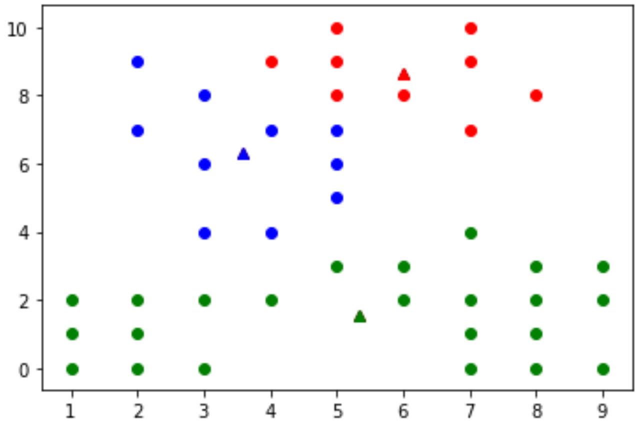
The basic step of k-means clustering is simple. Initially, we determine number of cluster K and select K centroid or center of these clusters from the dataset randomly.

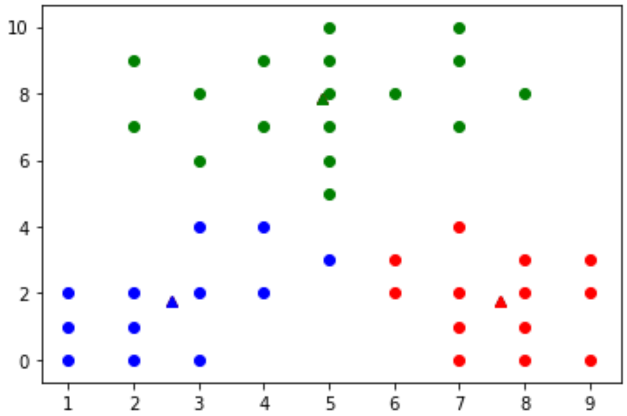
Then the K-means algorithm will iterate at the following steps until convergence.

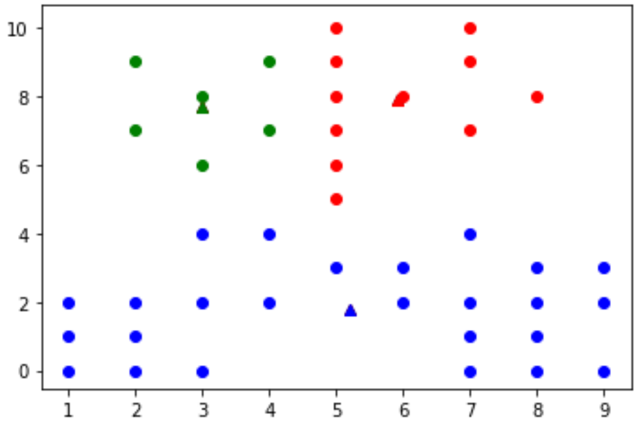
a. Update each centroid coordinate based on the data points in the cluster
b. Measure the distance of each point in the dataset to the K centroids
c. Group the point based on minimum distance


Start
# of cluster K
Centroids
Measure distances to centroids
Clustering based on minimum distance
No updates on clusters
No
Yes
End

Generally, the seeds or centroids are selected randomly. Here, we have three sets of selected initial centroids. They are described in the code. So, we can verify the correctness of your algorithm by the resulted figures. The results of the clusters are shown in the following three figures.







2. (30 points) Hidden Markov Model.

Please review the lecture of hidden Markov model and answer the following question. You will have the two same transition and emission tables on page 22.

Suppose the day you were locked in the room was Rainy. The caretaker did not bring in an umbrella on day 2 but on day 3. We assume the prior probability of the caretaker bringing an umbrella is 0.6. What is the probability that it is rainy on day 3?

Please write a Python code to output your major intermediate steps and final results.