CSCI 4734 – Machine Learning
Spring 2020 – Homework 3
K-means clustering
Due date: February 29, 2020

1. *K*-means clustering

In this this exercise, you will implement the K-means algorithm. You will experiment with an example 2D dataset that will help you gain an intuition of how the *K*-means algorithm works. The *K*-means algorithm is a method to automatically cluster similar data examples together. Concretely, you are given a training set $\{x^{(1)}, \dots, x^{(m)}\}$ (where $x^{(i)} \in \mathbb{R}^n$), and want to group the data into a few cohesive "clusters". The intuition behind *K*-means is an iterative procedure that starts by guessing the initial centroids, and then refines this guess by repeatedly assigning examples to their closest centroids and then recomputing the centroids based on the assignments.

The inner-loop of the algorithm repeatedly carries out two steps: (i) Assigning each training example $x^{(i)}$to its closest centroid, and (ii) Recomputing the mean of each centroid using the points assigned to it. The *K*-means algorithm will always converge to some final set of means for the centroids. Note that the converged solution may not always be ideal and depends on the initial setting of the centroids. Therefore, in practice the *K*-means algorithm is usually run a few times with different random initializations. One way to choose between these different solutions from different random initializations is to choose the one with the lowest cost function value (distortion). At the end, you may try to visualize your data to see whether your code works correctly (Figure 1).

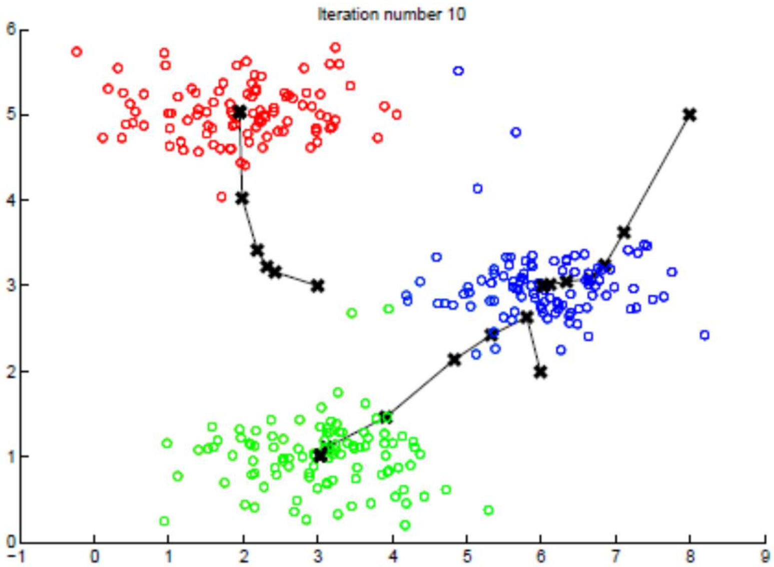

Figure 1: The expected output.

After training the data, define indexes of classes for the following pairs:
0,3 5,2

2	4,4
1,1	5,2
5,2	2,9
5,3	3,3
6,1	2,8

Your homework is divided into:

10% - loading data

10% - visualization

10% - Calculate cost function

30% - K-means algorithm from scratch

10% - testing

20% - Solve same problem by K-means algorithm using python library and test given pairs

10% - calculate silhouette_score

Please follow the instruction below when you submit your homework assignment to the Blackboard System.

- Submit homework solution in 2 files: 1) python file (.py or ipynb) 2) word file containing your codes and achieved results.
- Write your full name and number of homework assignment in the FILE NAME.
- Don't submit files in Zip file format.

Note. Please be informed that your assignment on coding will be checked through Safe Assign in Blackboard. In case of two same (similar) assignment submissions, both students will be scored as a zero. Furthermore, any student whose solution may arise a question or, will be asked for some explanations as well.