P4 K-Means Clustering

Due before midnight on March 31st.

10% of your Final Grade

For project 4 you will implement a K-means clustering algorithm.

Your data file will be formatted with first line containing m and n, tab separated, where m is the number of lines of data and n is the number of features (for this assignment n will be 2 but assume we still put it into the file.)

Each line thereafter will contain two real values (feature x1 and feature x2), tab separated.

Example

4	2
6.3	6
6.7	5.8
5.7	4.1
5.6	3.9

1. Your program should prompt the user for the name of a data file.
2. Prompt the user for the name of a file containing two initial centroids.
3. Print out the coordinates of the two initial centroids.
4. Print out a plot of the data to the screen, including the two initial centroids
5. Run K-means (K=2) to cluster the data into two groups.
6. Print out a plot of the cluster data with each cluster color coded along with the final centroids.
7. Print out the coordinates of the final centroids.
8. Compute and print out the overall error (J function presented in class) for the entire data set.

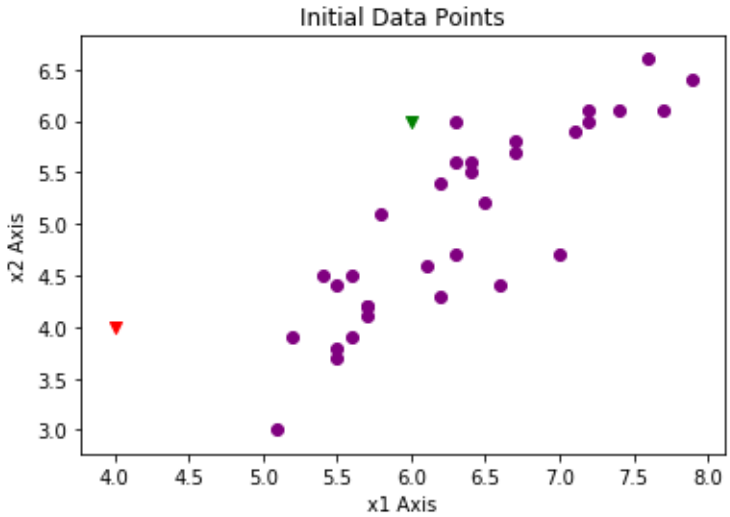
1. Your program should prompt the user for the name of a data file.
2. Prompt the user for the name of a file containing two initial centroids, formatted with the number of centroids on the first line and the coordinates of each centroid on the following lines, one centroid per line, tab separated.

Example Initial Centroid file

2
4 4
6 6

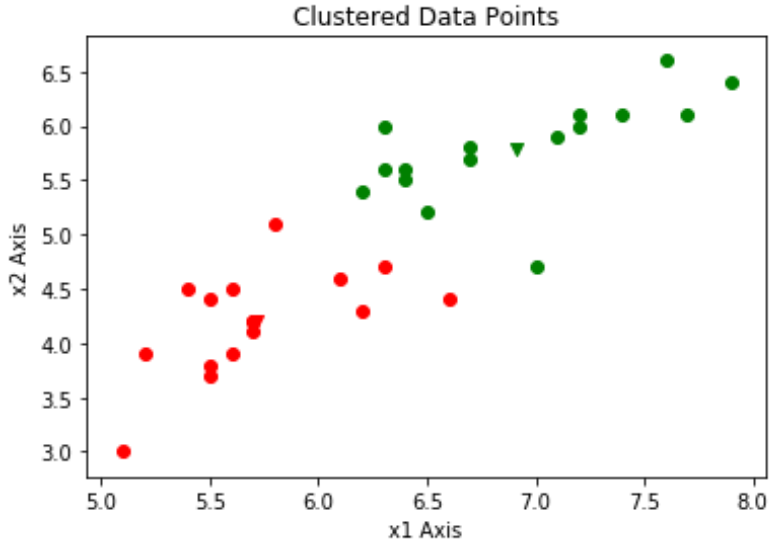
3. Print out the coordinates of the two initial centroids.

4. Print out a plot of the data to the screen, including the two initial centroids



5. Run K-means (K=2) to cluster the data into two groups.

6. Print out a plot of the cluster data with each cluster color coded along with the final centroids.



7. Print out the coordinates of the final centroids.

Final centroids are: [[5.71875 4.20625]
[6.9125 5.79375]]

8. Compute and print out the overall error (J function presented in class) for the entire data set.

Error is 0.43064453124999974

What to turn in: One **Zipped** Python file named yourlastname_yourfirstname_P4.py Make NO assumptions about other files being available. Your program should work with whatever data file with two features that we run it on in a Spyder environment.

I will put a practice input file name P4Data.txt and a practice Centroid file name P4Centroids.txt on Canvas. These may not be the files we use to test your program but will just be examples of the format for the two input files.