Project 4: K-Means Clustering

Due before midnight on Nov. 28th 2021

For project 4 you will implement a K-means clustering algorithm. Your data file will be formatted with the 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 x_1 and feature x_2), tab separated.

Example

4	2
6.3	6
6.7	5.8
5.7	4.1
5.6	3.9

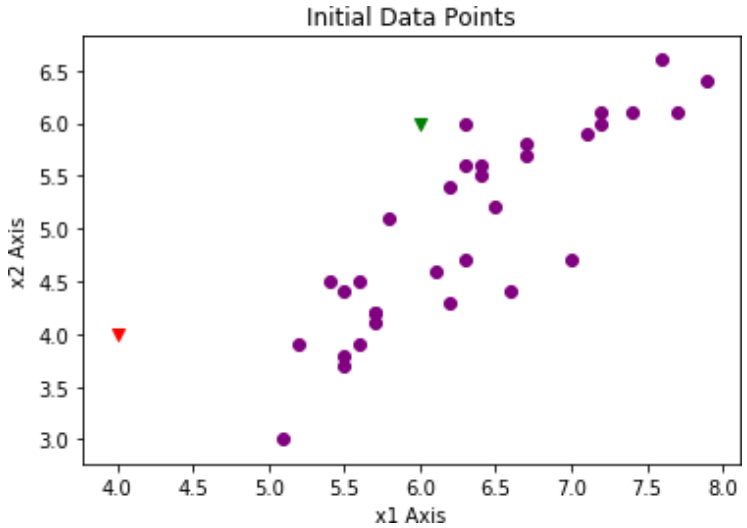
What to do:

1. Prompt the user for the name of a data file formatted as described above.
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

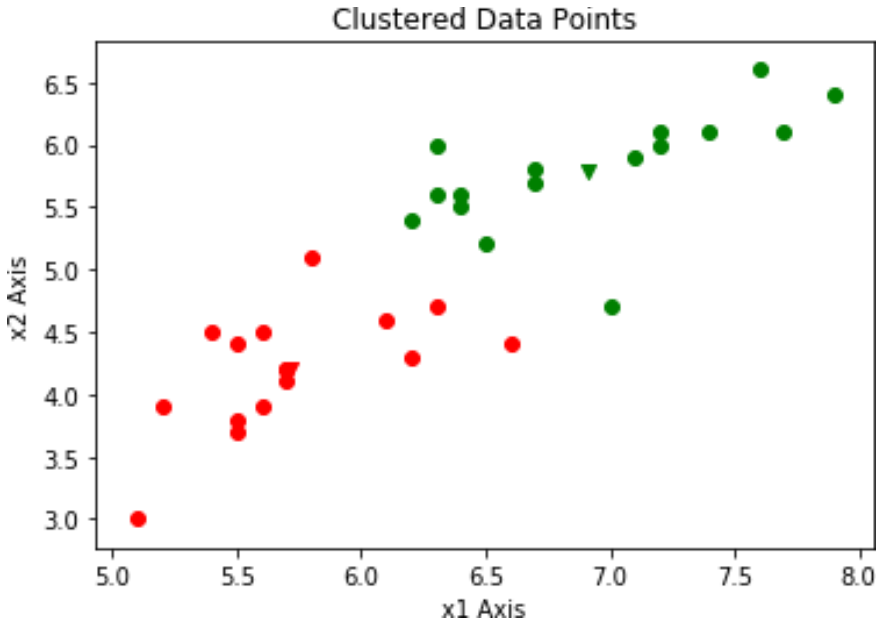
2	
4	4
6	6

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 (color coded).



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 the video) for the entire data set.

   Error is 0.43064453124999974

What to turn in:

One **Zipped file** containing:

- Python file named yourlastname_yourfirstname_P4.py

  Make NO assumptions about other files being available. Your program should work with any data file with two featuresthat we run it on in a Spyder environment and with a file of any two initial Centroids.

  I will put a practice input file name P4Data.txt and a practice Centroid file name P4Centroids.txt on Canvas. These **maynot be the files** we use to test your program but will just be examples of the format for the two input files.

  Note: All output should be to the screen, not to files.

- A report named yourlastname_yourfirstname_P4.pdf briefly introduces
  - The input and output of your code
  - The requested plots
  - Screenshot of your command window
  - A copy of your code