CPSC-6430 Machine Learning: Implementation & Evaluation

Project 4: K-Means Clustering

Dineshchandar Ravichandran
C19657741
Email: dravich@g.clemson.edu

Contents

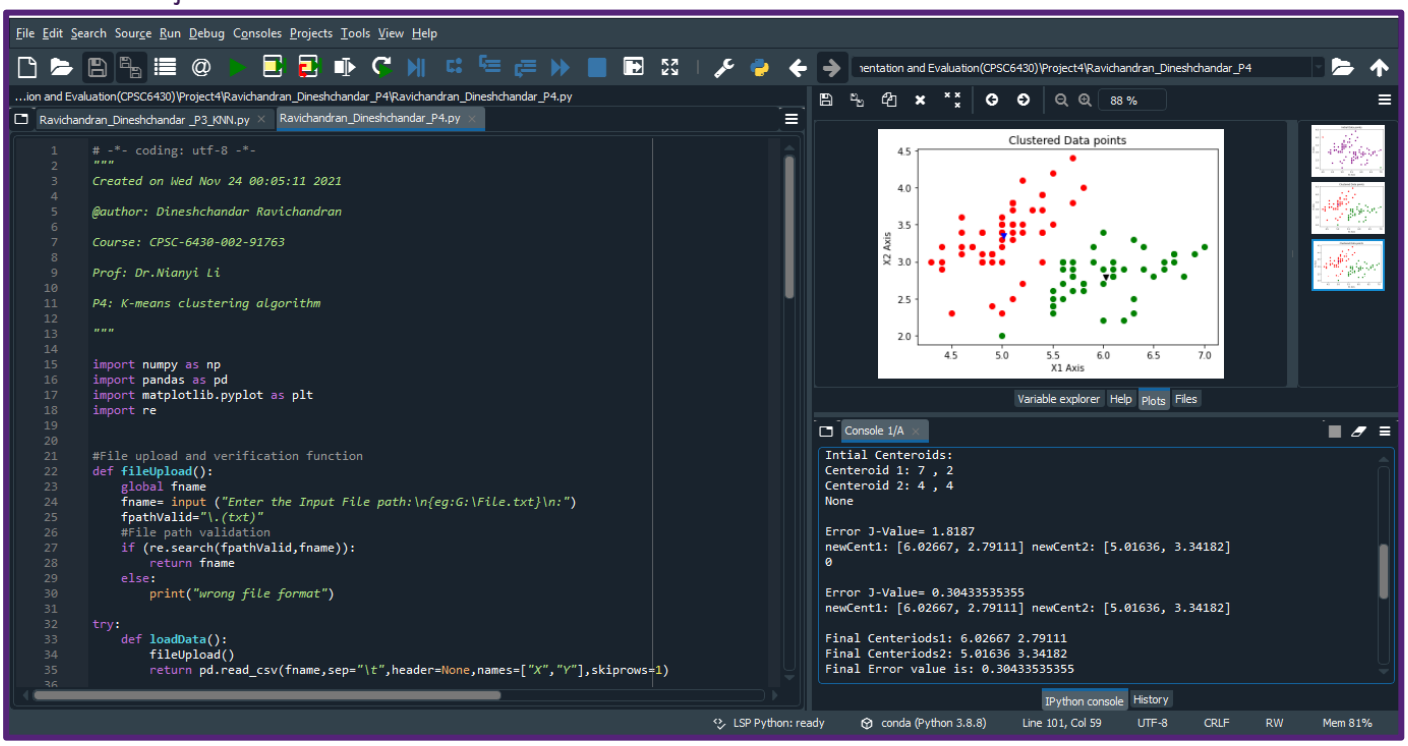
Introduction

Project 4, to implement a “K-means clustering algorithm” program to perform unsupervised classification on a set of points received from the data file.

1. Problem Statement

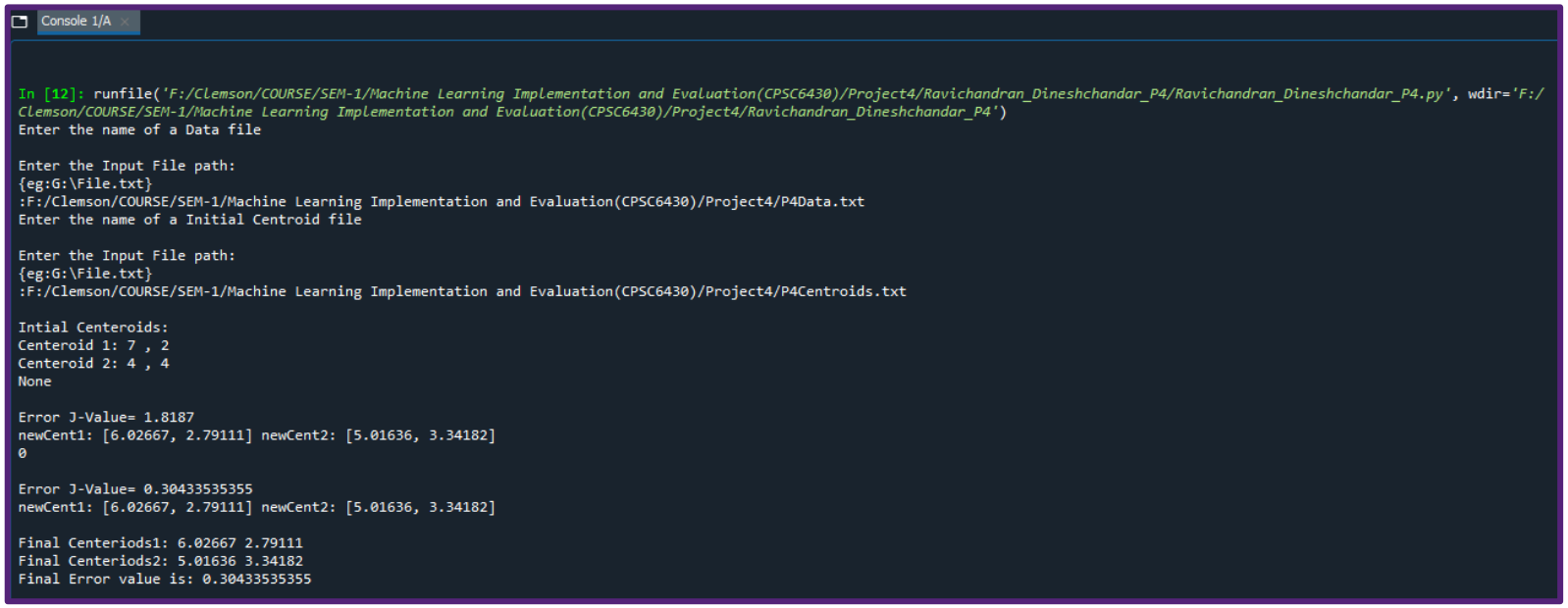
To implement a Python program to read the data from a “.txt” file containing the data set received from the user. And cluster them in 2 clusters based on the centroids.

2. Project Screenshot:



- The above screenshots represent the code in the SPYDER IDE, along with the response in the console, where the cost value for the final centroid is **0.30433535355**.
- The value of initial centroids as extracted from the user-provided file are:
  - Centroid 1:[7, 2]
  - Centroid 2:[4, 4]
- The value of final centroids are:
  - Final Centroid 1: [6.02667, 2.79111]
  - Final Centroid 2: [5.01636 3.34182]
- The above graph illustrates the data points are being classified into clusters:
  - Cluster1: With centroid [6.02667, 2.79111] illustrated in red dots.
  - Cluster2: With centroid [5.01636 3.34182] illustrated in green dots.

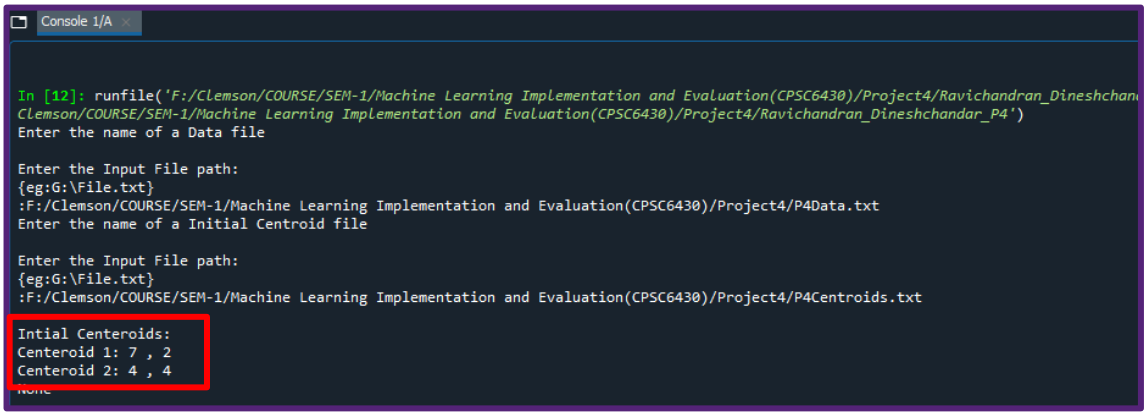
- Console Screen Shot for the same:



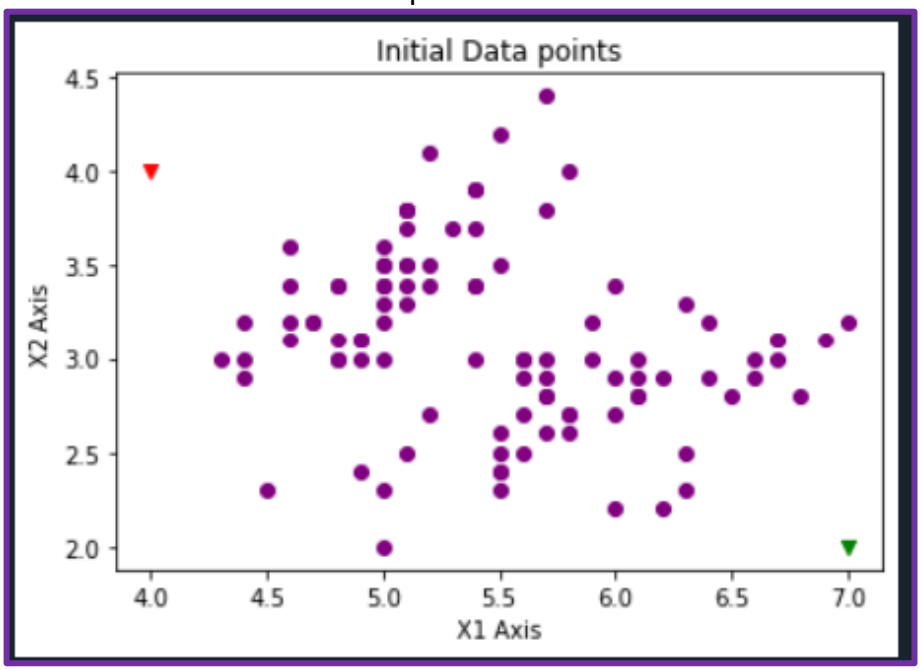
3. Project Input and Output

3.1. Input:

- User-provided Data file containing coordinates of 100 samples named “P4Data.txt”.
- User-provided Centroid file containing two centroids named “P4Centroids.txt”.
- Following is the data extracted from the user mentioned above centroid file, presented in the console(highlighted in red):



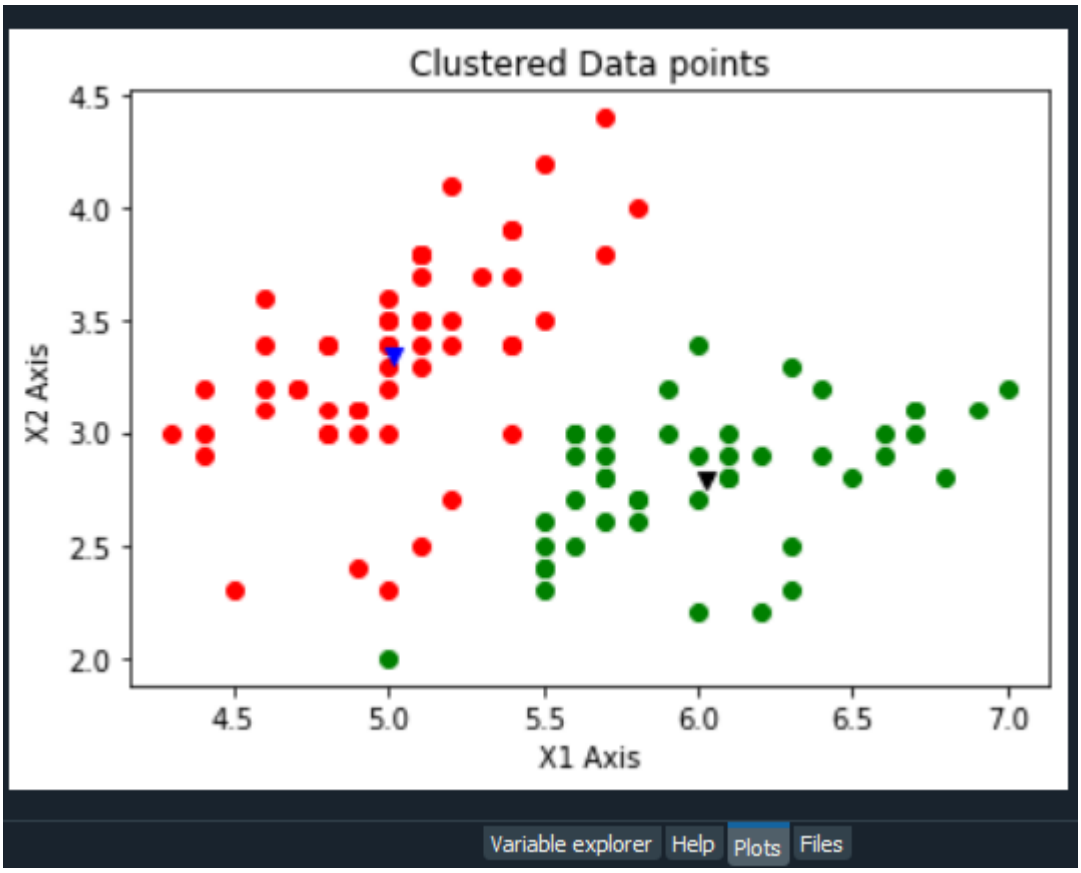
- The printout of the plot based on the data points and centroid points provided by the user before KMean actions are performed:



- Here the data points are presented in purple dots, and the centroid 1 [7,2] is represented as a Green triangle and centroid 2[4,4] as a red triangle.

3.2 Output:

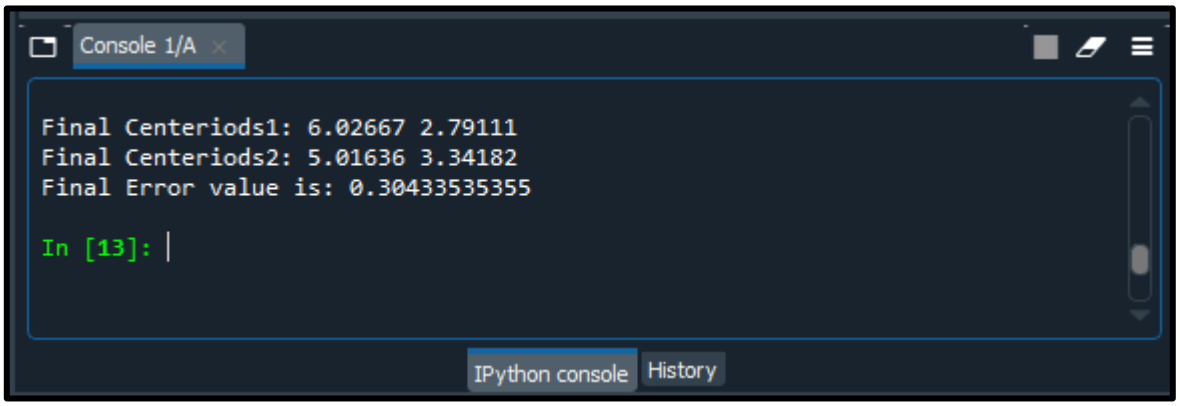
3.2.1 Cluster data plot along with the final centroid points.



- All the data clustered with respect to centroid **[5.01636 3.34182]** (marked in the blue triangle) is represented as red dots.
- All the data clustered with respect to centroid **[6.02667, 2.79111]** (marked in the black triangle) is represented as green dots.

3.2.2 Coordinates for final centroid.

- The values for the final centroid are **[6.02667, 2.79111]** and **[5.01636 3.34182]**, as illustrated below:



3.2.3 Overall error.

- The overall error = 0.30433535355, as highlighted in red in the below screenshot.