Author: Han Han

Instructions:

For part 1:

data = pd.read\_csv(r”C:\Users\mrj\Desktop\TwoDimHard.csv”)

Here the CSV file directory is in disk C and under Users\mrj\Desktop folders. In the beginning, you also need to include a letter ‘r’ to import the CSV file. You might need to locate the TwoDimHard.csv file and copy the directory of this CSV file to here.

I create a method to calculate the K means algorithm. In the K means algorithm method, I first convert the data type from a data frame to a matrix and then calculate the true centroids based on the true cluster labels. I also create a Euclidean distance method to calculate the distance between each data point to the centroids. Then, I proceed with K means algorithm with visual graphs. In addition, I also calculate the SSE for the data with only one cluster to calculate the SSB of more clusters.

For part 2:

Since the data file is online, you don’t have to change the directory of this file. You can just run the code from the top to the bottom.

First of all, I proceed with K means for this wine dataset. In order to have the optimal K value for this wine dataset, I run the K means with the K value from 1 to 15. Then I have the optimal K value 6. Then I run the K means algorithm and the internal and external measures.

Secondly, I proceed with DBSCAN algorithm for this wine dataset. I run many tests to have better parameter values. Then I calculate the centroids for further analysis, Therefore, I can calculate the SSE, SSB, and purity as the validity measures.

Finally, I proceed with Agglomerative Hierarchical algorithm. Based on the dendrogram, I find the optimal numbers of clusters. Then I use the internal and external measure to verify the results.