Homework 4

Note: As a solution to each part, upload your python codes, discussion narrative, and image files in Canvas.

The attached “prnn_viruses.csv” file contains a dataset of 61 viruses with rod-shaped particles affecting various crops (tobacco, tomato, cucumber, and others) described by Fauquet et al. (1988) and analyzed by Eslava-Gomez (1989). There are 18 measurements on each virus, representing the number of amino acid residues per molecule of coat protein.

The whole dataset is in order Hordeviruses (3), Tobraviruses (6), Tobamoviruses (39) and Furoviruses (13). These were added as the last (target) attribute.

Goal: Use the scikit-learn methods to perform the clustering of the 61 viruses and determine which clustering method is best to use for recognizing the true classification (target) for these viruses.

A. Read the data. For analysis, drop the column corresponding to ‘virus_type’. Set the column ‘virus_type’ as the target.

B. Print information about the data contained in the dataset, such as the header, description or summary. Print the mean for each column.

C. Graph the dataset using a pair plot representation.

D. Plot the dataset in the (‘col_1’,’col_2’) plane using colors based on the 4 types of viruses.

E. Use the **Kmeans method** to cluster the viruses data. Start by determining the optimal number of clusters using the elbow method, the average silhouette score, the Calinski-Harabasz score, and the DBI score. Plot the within cluster sum of squares (WCSS) versus the number of clusters (check up to 15 clusters), as well as the other scores versus the number of clusters. How many clusters are optimal based on this analysis? Is this what you were expecting?

F. Apply the Kmeans method using the number of clusters from E. Save the cluster labels for the 61 points. How well does Kmeans perform in labeling the viruses?

G. Plot the dataset in the (‘col_1’,’col_2’) plane using colors based on the cluster labels. Include the cluster centers. How does this compare with the plot from D? How distinct are the clusters in this plane?

H. Apply the **Gaussian Mixture Model (GMM) Clustering algorithm** to the viruses dataset using the number of clusters learned from part E. Save the probabilities for each of the 61 observations to belong to each of the clusters. How good is the GMM clustering in recognizing the classes for these observations?

I. Use the **Agglomerative Hierarchical clustering method** to cluster the viruses data. Start by determining the dendrogram which allows us to know the clusters that we want our data to be split to. Plot the dendrogram for the *complete* linkage. Use the various scores listed in part E to determine the optimal number of clusters. How many clusters are optimal based on this analysis? Is this what you were expecting?

J. Apply the Agglomerative Hierarchical method with *complete linkage* and Euclidian distance using the number of clusters from I. Save the cluster labels for the 61 points. How well does this method perform in labeling the viruses?

K. Plot the dataset in the (‘col_1’,’col_2’) plane using colors based on the cluster labels from J. How does this plot compare with the plot from D? How distinct are the clusters in this plane?

L. Repeat I to K for the **Agglomerative Hierarchical clustering method** with average *linkage.*

M. Which of the 2 types of the **Agglomerative Hierarchical clustering methods** performs best on the viruses dataset? How does this compare with the performance of the **Kmeans** and **GMM** methods? Which clustering method performs best overall on the viruses dataset?