COMP4432 Machine Learning - Assignment 5

Learning Objective: This assignment will offer the opportunity to identify the optimal clustering of a more challenging data set than those shown in class. For each clustering algorithm considered in this assignment, the respective optimal hyperparameter settings will be identified with appropriate methods. The clusterings and their evaluation metrics will be compared to identify the best clustering of the data.

Part 1: Data Exploration

a) Load the provided data (*Assignment5_Data.csv*) into a Pandas DataFrame and examine the first few rows

b) Scale the data (fit and transform) using a StandardScaler

c) Create the scatterplot with surrounding density plots shown in the live session
Tip: Review the code shared after session 8 for reference

d) In a Markdown cell, document the following:

   i) Estimate the number of clusters from the scatter and density plots

   ii) Also, note any potential issues observed in the data. For example, for the data observed, are the cluster densities consistent? Discuss any regions of ambiguity or overlap.

Part 2: K-Means Clustering

a) Estimate the optimum number of clusters present in the dataset using the following methods:

   i) Elbow Method

   ii) Silhouette Score

   iii) Davies-Bouldin Score
   Hint: Try using a single loop over potential *k* values. See the notebook shared from Week 8.

b) Show a graph of each of the methods in Part 2a versus the number of clusters *k.*

c) In a Markdown cell, document the optimum number of clusters found from each method, and discuss the number you selected.

d) Fit a K-Means clustering algorithm with the optimal *k* value selected in Part 2c, and store the trained clusterer into an appropriately named variable. This represents the best version of a K-Means clusterer for the data

Part 3: Gaussian Mixture Model

a) Estimate the optimal number of clusters from AIC/BIC calculations included within the GaussianMixture method.

b) Show a graph of AIC and BIC from Part 3a versus the number of clusters.

c) In a Markdown cell, document the optimum number of clusters

d) Fit a Gaussian Mixture Model with the optimal k value selected in Part 3c, and store the trained mixture model into an appropriately named variable. This represents the best version of a Gaussian Mixture Model for the data

Part 4: DBSCAN

a) Knowing the number of dimensions in the dataset, select the appropriate values for *min_samples*
b) With this value of *min_samples*, identify the optimal value for the epsilon hyperparameter (the distance criteria)
   Hint: Review the discussion and code shared from Week 9
c) Fit a DBSCAN clusterer with these hyperparameters, and store the trained mixture model into an appropriately named variable.
d) Create a plot of data set color coded by the cluster labels form Part 5c.
e) The outliers (labels_ = -1) are problematic to evaluation.
   i) Copy the scaled data set into a new data set named *dbscan_df*
   ii) Add a new column to *dbscan_df* for the labels predicted from the best DBSCAN clusterer in Part 5c
   iii) Remove all rows with these labels = -1
   iv) Show a graph of *dbscan_df* color coded by cluster labels.
   v) In a Markdown cell, compare the plots with and without outliers and discuss whether you should adjust the distance hyperparameter epsilon.

Part 5: Model Evaluation

a) You have three optimized clustering methods: K-Means, Gaussian Mixture Model, and DBSCAN. For each clustering method, construct a plot of the data color coded to the cluster label predictions.
   Hint: plt.subplots(nrows= 1, ncols= 3) and Seaborn scatterplot are really helpful.
b) In a Markdown cell, answer the following:
   i) Visually inspect the plots in Part 5a. In your opinion, which method produced the best clustering of the data and why?
   ii) Discuss the similarities and differences in the clusters identified by the different methods.
   iii) Review the discussion in Part 1 regarding the estimation for number of clusters, regions of overlap, and densities. Discuss how each clustering method performed with these topics.
c) Gather or calculate the Silhouette Score and Davies-Bouldin Score for each of the three clusterings.
   Tip: Be cautious with DBSCAN calculations due to the data points labeled as outliers. Use the *dbscan_df* data set from Part 4e for these calculations
   In a Markdown cell, answer the following:
   i) Which method performed best per metric?
   ii) Do you agree with the metric values assessment?
   iii) One clusterer optimized the Silhouette Score and another clusterer optimized the Davies-Bouldin Score. Why the disagreement between scoring metrics?