C S 488/508 Introduction to Data Mining
Homework 6: Clustering

Objective

In this **individual** homework, you will do exercises to write program to test different clustering algorithms.

Requirements

This assignment uses a data set called Travel Review Ratings Data Set from the UCI machine learning repository. It contains Google reviews on attractions from 24 categories across Europe. Google user rating ranges from 1 to 5 and the average user rating per category is calculated. Its csv file can be downloaded from Canvas (google_review_ratings.csv). For this dataset, we want to do clustering for understanding how users give reviews to different attractions across Europe. Refer to the data set description at `https://archive.ics.uci.edu/ml/datasets/Tarvel+Review+Ratings` for more details about its attribute information and instances. Please DO NOT use the csv file from the UCI as it is slightly different from the one on Canvas.

Q1. (10 points) (Data Preprocessing)
   (1) Remove all rows that contain missing values. Save the new data to `data.csv` and submit it. Use the new data (`data.csv`) generated in this question to perform the following tasks.

Q2. (25 points) (k-means) We want to train a k-means model.
   (1) (15 points) We will use the Silhouette coefficient to select the best number of clusters (k) from [1,10]. For each k, run k-means **two** times and compute the average Silhouette coefficient across two running times and clusters. Plot the average Silhouette coefficients for different k. Submit the plot. What is the best k?
   (2) (10 points) Train a k-means model with the best k above. Report the centroids of clusters.

Q3. (15 points) (Agglomerative Clustering) Cluster the data using Agglomerative Clustering. One thing to note is that for each k, running Agglomerative Clustering one time is good enough. Set the number of clusters you get to be the best k you found in Q2.(1).

Q4. (15 points) (Gaussian Mixture Model) Cluster the data using Gaussian Mixture Model. Set the number of clusters to be the best k you found in Q2.(1).

Q5. (15 points) (Spectral Clustering) Cluster the data using Spectral clustering. Set the number of clusters to be the best k you found in Q2.(1).

Q6. (10 points) (Comparison) Draw a plot that includes the average Silhouette coefficients for all the methods above. Submit the plot. Which method is better in terms of the average Silhouette coefficient?

Q7. (10 points) (Comparison, **CS 508 only**) Calculate WSS and BSS values for the clusters you found from all the methods above. Submit the values. Which method is better in terms of WSS and BSS?

Submission instructions

A zipped file `hw-lastname.zip` consisting of all the code and the PDF files containing discussions and figures.

Grading criteria

(1) CS 508 students need to answer all the questions.

(2) CS 488 students do not need to answer questions marked with **(CS 508 only)** although you have the freedom to work on them. Your scores will be scaled to 100. If CS 488 students answer the questions marked with **(CS 508 only)**, you will not have any points deducted if your answers are wrong; you will not get any extra points either if your answers are correct.

(3) The score allocation has been put beside the questions.

(4) Please make sure that you test your code **thoroughly**.

(5) FIVE points will be deducted if files are not submitted in the required format.