Lab5-Clustering

Total Marks: **10** + 2 (individual assessment)= 12 marks

In this part you will use Python to create the clusters in the heart disease data set (the link and explanation is included here)

Heart Disease Dataset: Here, is the link for heart disease dataset of patients.
http://archive.ics.uci.edu/ml/datasets/Heart+Disease
After going to this link you will find two folders: One: Data Folder and two: Dataset description. Data folder that has the dataset. It is better to use processed cleveland data. In the dataset description folder, you will find the description about the columns' names referring to the14 column of the dataset as the following: The last one attribute (number 14) is the result

1. #3 (age)
2. #4 (sex)
3. #9 (cp)
4. #10 (trestbps) 5. #12 (chol)

6. #16 (fbs)
7. #19 (restecg)
8. #32 (thalach)
9. #38 (exang)
10. #40 (oldpeak)
....
13. #51 (thal)
--
14. #58 (num) --------------------------->result

Implement the following using python notebook and submit your source code with results. Before you start, refer to the clustering lecture notes to understand how to create k-means and agglomerative clusters using a python library, how to find the right value of "k" for the k-means, and details about silhouette score.

1. Using the Python "sklearn" library, create and visualize the k-means clusters (with k=5) for the given heart disease dataset. For visualization, draw the scatter plot using the age and cholesterol features on each group of clusters. (2 marks)

2. Apply k-means clusters on the heart disease dataset with varying numbers of clusters from 1 to 10 and compute their corresponding Sum of squared Error (SSE) value. Plot the graph using Python "matplotlib" library and estimate the right "k" value.

   Hint:
   - The "elbow" in the plot of SSE vs the number of clusters can help to estimate "k" value

- Use python kmeans.inertia_ method to compute SSE
- Use KneeLocator from kneed python library to find the elbow value

(2 marks)

3. Create and visualize the k-means clustering with the “k” value obtained in Q2. The clustering algorithm ultimately groups similar patients by matching its features. Thus, for the visualization, draw the scatter plot using the age and cholesterol features on each cluster group and label them as “Group-A”, “Group-B” etc., (2 marks)

4. Plot a dendrogram using Python scipy.cluster.hierarchy method. (1 marks)

5. Create the agglomerative clustering with the number of clusters is equal to the “k” value obtained in Question-2. Visualize the clusters similar to Question-2. (2 marks)

6. Compute silhouette score for both K-means and agglomerative clustering and tell us which clustering is better for the given dataset. (1 marks)