Lab - 6
CSL2010: Introduction To Machine Learning
AY 2022-23
K-Means Clustering, Agglomerative Clustering Algorithms

(Due: 21 Sep 2022, 11:59 PM)

General Instructions

1. You need to upload a zip ***<Lab6_Your_Roll_No>.zip***, which contains one file for the task in ***<Lab6_Your_Roll_No>.py*** format and the report for the entire assignment in ***<Lab6_Your_Roll_No>.pdf*** format.
2. Provide your colab file link in the report. **Make sure that your file is accessible**.
3. Submit a single report mentioning your observations for all the tasks.
4. Report/Cite any resources you have used while attempting the assignment.
5. Attempt Q1 (a) and (b) during the lab.
6. Q2 is for your practice and will not be graded.

Question 1: K-Means clustering is an unsupervised learning algorithm which groups the unlabeled dataset into different clusters. [20 marks + 10 marks bonus]

You have been given the Wine Quality Dataset (Dataset Description) and the details about the dataset is given in the link. Ignore the category of samples, pre-process the data if required, and perform the following tasks:

a) Visualize the distribution of data points by picking different pairs of attributes, and by looking at the scatter plot, estimate what value of \`k' (i.e., number of clusters) might be best suited for k-means clustering and why? (There is no need to use any method to find the optimal value of 'k'.) [10 marks]

b) Perform k-means clustering on this data (can use sklearn library) using the value of \`k' which you have chosen above. Visualize by showing the clusters along with the centroids. [10 marks]

c) BONUS: Implement k-means clustering algorithm from scratch and perform all the operations previously performed in *part b*. [10 Marks]

Question 2 [NON-GRADED]: Hierarchical clustering is another unsupervised learning algorithm that is used to group together the unlabeled data points having similar characteristics.
Dataset: Brain Cancer Dataset.

A. Normalize/Scale the data so that the scale of each variable will be the same.

B. Visualize the dataset & find out the number of communities available.

C. Visualize the communities from part A.

D. Apply Agglomerative hierarchical clustering (using sklearn).

E. Apply K-means (sklearn) and make a comparison between these two approaches & justify your results.