Lab - 8

CSL2010: Introduction To Machine Learning
AY 2021-22

General Instructions

1. Prepare separate Python code files for each task :
   Task 1 : Name the file as ***<Your_Roll_No>_Task1.py***.
   Task 2: You need to upload a zip ***<Your_Roll_No>.zip***, which contains two files for the task in ***<Your_Roll_No>_Task2.py*** format and the report for the entire assignment in ***<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, mention your observations for all the tasks.[Include plots]
4. Report any resources you have used while attempting the assignment.

Any submission received in another format or after the deadline will not be evaluated.

Task 1 (Due: 11:59 PM, 06 Oct 2021)

K Means **[4 marks]**

i. Download the dataset from the given link.

ii. Perform K Means clustering using the scikit learn library, try at least three different values of **n_clusters**, and use **random_state = 2021.**

iii. Print the cluster_centers_ obtained.

iv. Plot a scatter plot after performing the K means clustering between any two attributes/features of the dataset.

Task 2 (Due: 5:30 PM, 13 Oct' 2021)

Hierarchical Clustering and K Means [w/o inbuilt function] [16 marks]

I. i. Plot a dendrogram on the dataset to see how clusters are being formed. [Using scipy library]

ii. Perform agglomerative clustering using the scikit learn library,use different values of n_clusters and linkage.

iii. Plot a scatter plot after performing the agglomerative clustering to see how the data is clustered.

II. Implement K-Means clustering from scratch without using the inbuilt library.

i. Choose a value k as the number of clusters.

ii. Initialize the centroid randomly for the data points.

iii. Compute the euclidean distance from the centroid and assign the cluster based on the minimum distance.

iv. Calculate the mean of the data points assigned to each cluster and assign a new centroid for the clusters.

v. Repeat iii & iv till convergence is achieved.

vi. Visualize to see how the data is clustered.