CS5101 Machine Learning Lab (Aug - Dec 2021)
Assignment 8

(Clustering)

Rules:

1. The assignment should be submitted via moodle in .ipynb format.
2. Single notebook should be submitted including all solutions
3. **Plagiarism will be checked and Zero marks will be given if found**.
4. Your code should be neat and well commented
5. Maximum marks for this assignment are 5.

Questions:

Q1 - You are provided with a dataset with each data point having two features namely weight and height. Perform all three clustering techniques: K-Means, GMM and Agglomerative clustering on this dataset. You should optimize hyperparameters available for all the clustering techniques wherever possible. (3 marks)

- Your code should input the entire data from the given csv file and perform all the three above mentioned clustering techniques.
- Report following outputs in the python notebook itself with proper headings mentioning clustering technique used:
    - Choose and report optimal no: of clusters/components for the given dataset and show how you chose the value
    - Find best hyperparameters for each clustering technique
    - Output the scatter plot for the given data coloring each data point based on clusters assigned (one per clustering method)
    - For agglomerative clustering visualize the dendrogram for the given data •

- evaluation scheme: -
    - 1 mark - Implementation of each clustering technique (code)
    - 1 mark - Choosing optimal no:of clusters/components (explanation/visualization of selection) and suitable hyperparameters for each technique
    - 1 mark - Visualization of data (scatter plots assigning different colors for each clusters obtained) and dendrograms

Q2 - Outlier Detection (2 marks)

- Apply DBSCAN on the dataset given in data.csv to group the data into clusters and also predict the outliers (noise points). Plot in 2D the clusters obtained along with the noise points for a clear visualisation.
- Add a column ”Outliers” in data.csv which will contain value -1 for noise points and value 0 otherwise and submit this csv file also.