CS:5101 Machine Learning

Term 3 (Dec 2020 - Feb 2021)

Programming Assignment - 7

Clustering Methods
K-Means, Gaussian Mixture Model & Hierarchical clustering

Due Date-4/1/2021

Follow the instructions given below carefully:

1. You are allowed to use all inbuilt libraries in today's assignment.
2. You must submit your code in a python .ipynb notebook with naming format as follows: Firstname_Lastname_assignment7.ipynb
3. For each question, create a separate text block containing the question followed by a code block containing the solution.
4. Your code must be properly commented explaining each step clearly.
5. If any of the above instructions are not followed, penalty will be there for the same.
6. Your code and answers will be checked for plagiarism and if found plagiarised, zero marks will be provided for assignment 7.

Question 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.

- 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:
  1)Choose and report optimal no: of clusters/components for the given dataset and show how you chose the value
  2)Find best hyperparameters for each clustering technique
  3) Output the scatter plot for the given data coloring each data point based on clusters assigned (one per clustering method)
  4) For agglomearive clustering visualize the dendrogram for the given data

- evaluation scheme:
  -1 mark-Implementation of each clustering technique (code)
  -2 mark - Choosing optimal no:of clusters/components (explanation/visualization of selection) and suitable hyperparameters for each technique
  -1.5 mark - Visualization of data (scatter plots assigning different colors for each clusters obtained) and dendrograms
  -0.5 mark- Write your observation about suitable clustering technique for the given dataset