Lab 6 - K-Means & Hierarchical Clustering

Lab 6 on K-Means and Hierarchical Clustering for DS3010 - Machine Learning

OVERVIEW & PURPOSE

In this lab, you will experiment with Clustering Algorithms of Machine Learning

Instructions

1. Please submit the assignment through Moodle in .ipynb format (python notebook)
2. The submission should contain a single notebook containing all the solutions, including the requested documentation, observations, and findings.
3. The naming convention for the notebook is <firstname>_<lastname>_<rollnumber>.ipynb
4. You must adequately comment on the code to improve its readability.
5. The lab is worth 5 points
6. This graded lab is due on October 20th at 6 pm

Lab

1. K-Means Clustering (2 Marks)

a. Read the provided 'BMI_Data.csv' file using Pandas Dataframe.
b. Plot the scatter plot of the data points.
c. Optimize the hyperparameter n_clusters(K) value using the elbow method. Take the K value in the range(1, 10). Print the objective cost value for each K. Find the optimal K from the plot of objective cost (sum of distances of points from the cluster centroid) as a function of number of clusters. Observe the knee point in the plot and estimate the optimal number of

clusters.

d. Train the dataset with the optimal K value using K-Means Clustering Algorithm. Also, use the init, n_init, max_iter, tol and random_state parameters.
e. Plot each clusters with their data points. Also plot the centroid of each cluster on the same plot. Display each clusters and their centroids with different colors and different markers. Plot axes should be labelled. Also add the legend.

3. Hierarchical Agglomerative Clustering (3 Marks)

a. Read the provided 'preprocessed_customer_segmentation.csv' file using pandas dataframe. Consider the random sample of 100 rows of created dataframe as the new dataset.
b. Implement the Hierarchical Agglomerative Clustering from scratch. Create a function find_HAC which will take the numpy array of generated data in **part a**, number of data points N and type of linkage as their parameters. The function will return clusters in each iteration like this

```
Iteration 1: [[0], [1], [2], [3, 5], [4]]
Iteration 2: [[0], [1, 4], [2], [3, 5]]
Iteration 3: [[0, 3, 5], [1, 4], [2]]
Iteration 4: [[0, 3, 5, 1, 4], [2]]
Iteration 5: [[0, 3, 5, 1, 4, 2]]
```

c. Plot the dendogram for the BMI_Data.csv dataset using the single linkage using sklearn library.
d. Write all the informations provided by the dendogram.