[Mini Project 3] **Project Code : VC4**

Virus Clustering using Complete Linkage Hierarchical Clustering Technique

Project Duration: 08-Mar-2021 ~~ 22-Mar-2021
Submission Information: (via) CSE-Moodle

Objective:
Consider the data about features of different kinds of viruses found in the bodies of different people. These features include size of the virus, number of nucleic acid molecules and transmission rate. You have to cluster these viruses into groups.
Your task is to cluster the dataset into an optimal number of clusters. *In particular, you shall be doing the following:*

1. **K-means clustering**
   Write a program to perform k-means clustering on the given dataset. Consider k=3 clusters. Consider cosine similarity as the distance measure. Randomly initialize k cluster means as k distinct data points. Iterate for 20 iterations. After the iterations are over, save the clustering information in a file. This file may be used in step 4 if the value of k is the optimal number of clusters.

2. **Evaluation of the clustering algorithm**
   Evaluate the result of your clustering algorithm using the Silhouette coefficient metric and print the value of s.

$$s = \frac{b - a}{max(a, b)}$$

   The Silhouette Coefficient is defined for each sample and is composed of two scores:
   a: The mean distance between a sample and all other points in the same cluster.
   b: The mean distance between a sample and all other points in the next nearest cluster.

   The Silhouette Coefficient for a set of samples is given as the mean of the Silhouette Coefficient for each sample. The score is bounded between -1 for incorrect clustering and +1 for highly dense clustering. Scores around zero indicate overlapping clusters. Larger value of Silhouette Coefficient denotes that clusters are denser and well-separated in adherence to the idea of clustering algorithms.

3. **Find optimal value of k**
   Repeat steps 1 and 2 for k = 4, 5 and 6 as well. Report the value of k for which you get the highest value of the Silhouette Coefficient. This will be the optimal number of clusters. You will be using this number in the next step.

4. **Hierarchical Clustering**
   Implement a **bottom-up hierarchical clustering algorithm** considering the same notion of similarity as in step 1. Find k clusters (optimal number of clusters from step 3) using **complete linkage strategy.**

   Now you have k clusters from the k-means algorithm and k clusters from hierarchical clustering on the same dataset. Or in other words, the dataset is divided into k sets of data points as a result of the k-means algorithm (case A). Similar is the case for the hierarchical clustering algorithm (case B). You need to compute the Jaccard similarity between corresponding sets of both the cases. Consider the following example to understand the process clearly.

Let's say k=4 and our dataset consists of numbers from 0 to 99. case A divides the dataset into 4 sets. For simplicity, let's say that the groups are 0-24, 25-49, 50-74 and 75-99. Now, since the second algorithm is also a clustering algorithm, the dataset should be divided into more or less similar groups with slight deviations. But, we can assume that most of the numbers from 0-24 will be in the same group. So, if we consider the Jaccard Similarity of the group 0-24 from case A with all the groups of case B, one group will show high similarity while the other three will be quite dissimilar. **This task requires you to first map each set of case A to a distinct set of case B (one-to-one and onto mapping) considering the Jaccard similarity as shown in the aforementioned example. After the mapping, print the Jaccard Similarity scores for all the k mappings.**

Note: *The program can be written in C / C++ / Java / Python programming language from scratch. No machine learning /data science /statistics package / library should be used for model creation.*

Dataset Description:
The dataset consists of the distinguishing characteristics of few of the viruses that exist on earth, and includes the following attributes:

1. Size [Float, Nanometers]
2. Number of strands [Integer]
3. Number of protein layers [Integer]
4. Sam ratio [Float]
5. Latency [Integer, Hours]

Read as Attribute [Datatype, Measurement Unit]

Submission Details: (to be submitted under the specified entry in CSE-Moodle)

1. ZIPPED Code Distribution in CSE-Moodle
2. A brief report that contains the optimal number of clusters and your analysis of the results about similarity coefficients in step 2 and step 4.
3. A file **kmeans.txt** that contains your final cluster information considering the optimal number of clusters that you have found out in step 3. **The format should be as follows**:
   Each line will represent a different cluster, and will contain a sorted comma separated list of the indices of the data points in that cluster. Sort the clusters by the minimum index of the data points present in that cluster.
   Eg: if suppose you obtain clusters [1,3,5], [2], [4,0], then the file should contain:
   0,4
   1,3,5
   2
   Here the numbers represent the index of the corresponding documents in the dataset (excluding the header)
4. A file **agglomerative.txt** that contains final cluster information from step 4 in the same format as kmeans.txt.
5. A README file containing the instructions to run your program. Please report the approximate time taken by your program to run all the steps in a reasonable PC configuration.
6. You are advised to write all the programs in a single file following a modular approach and ensure that the main function of your program runs all the steps in sequence as asked in the assignment.

Submission Guidelines:

1. You may use one of the following languages: C / C++ / Java / Python.
2. Your program should be standalone and should not use any special purpose library. **Numpy or Pandas may be used.** And, you can use libraries for other purposes, such as generation and formatting of data.
3. You should submit the program file and README file and not the output/input file.
4. You should name your file as <RollNo_ProjectCode.extension> (e.g., 19CS10000_VC4.pdf or 19CS30000_VC4.zip).

5. The submitted program file should have the following header comments:
   # Roll Number: Name of the student
   # Project Number
   # Project Title
6. The zip should contain one file for the program (in any of the prescribed languages), kmeans.txt, agglomerative.txt, README file, and the report in pdf format.

You should not use any code available on the Web. Submissions found to be plagiarised or having used ML libraries (except for parts where specifically allowed) will be awarded zero marks.

For any questions about the assignment, contact the following TAs:
Rupak Kumar Thakur (rupakthakur97@gmail.com)
Saurav Roy (saurav.roy.edu@gmail.com)
Shubham Gautam (shubhamg0510@gmail.com)
Vindhyansh Mall (m.vindhyansh@gmail.com)