[Project Code: PWHC-AC]

Portugal Weather using Complete Linkage Agglomerative (Bottom-Up) Clustering Technique

Project Duration: 24-Mar-2024 – 13-Apr-2024
Submission Information: (via) CSE-Moodle

Objective:

Based on the different features related to weather forecasting in Portugal between 2022 and 2023, 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 **Complete Linkage Agglomerative (Bottom-Up) 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 Details:

Dataset Filename: **`weather.csv`**
Dataset Description:

- Name: The name refers to the name of the country
- Datetime: Datetime denotes the date and time at which the weather conditions were observed or predicted.
- Tempmax: Tempmax represents the maximum temperature recorded or forecasted for a specific day in Portugal.
- Tempmin: Tempmin signifies the minimum temperature observed or predicted for aa specific day in Portugal.
- Temp: Temp denotes the current temperature in Portugal at the time of observation or prediction.
- Feelslikemax: Feelslikemax indicates the maximum "feels like" temperature experienced or expected in Portugal, which takes into account factors such as humidity and wind.
- Feelslikemin: Feelslikemin represents the minimum "feels like" temperature experienced or expected in Portugal.
- Feelslike: Feelslike refers to the current "feels like" temperature in Portugal.
- Dew: Dew represents the temperature at which air becomes saturated with water vapor, leading to the formation of dew.

- Humidity: Humidity indicates the amount of moisture present in the air, expressed as a percentage.
- Precip: Precip signifies the amount of rain precipitation that fell in Portugal.
- Precipprob: Precipprob represents the probability or likelihood of precipitation occurring in Portugal.
- Precipcover: Precipcover denotes the extent or coverage of an area that will be affected by precipitation in Portugal.
- Preciptype: Preciptype specifies the type of precipitation that is occurring or expected to occur in Portugal (rain, snow, sleet, etc.).
- Snow: Snow represents the amount of snowfall in Portugal, measured in centimeters.
- Snowdepth: Snowdepth indicates the depth of accumulated snow on the ground in Portugal.
- Windgust: Windgust signifies the maximum speed of wind gusts recorded in Portugal.
- Windspeed: Windspeed denotes the current speed of the wind in Portugal.
- Winddir: Winddir represents the direction from which the wind is blowing in Portugal.
- Sealevelpressure: Sealevelpressure refers to the atmospheric pressure measured at sea level in Portugal.
- Cloudcover: Cloudcover indicates the extent to which the sky is covered by clouds in Portugal, expressed as a percentage.
- Visibility: Visibility measures the horizontal distance at which objects are clearly visible in Portugal.
- Solarradiation: Solarradiation represents the amount of solar energy or radiation received at the Earth's surface in Portugal.
- Solarenergy: Solarenergy denotes the amount of solar energy that is available for use or absorbed by the Earth in Portugal.
- UVindex: UVindex represents the level of ultraviolet (UV) radiation from the sun in Portugal, which indicates the risk of sunburn or skin damage.
- Severerisk: Severerisk signifies the level of risk or potential severity of adverse weather conditions in Portugal.
- Sunrise: Sunrise denotes the time at which the sun rises above the horizon in Portugal.
- Sunset: Sunset represents the time at which the sun sets below the horizon in Portugal.
- Moonphase: Moonphase refers to the current phase or appearance of the moon in Portugal (the proportion of the moon that we saw on a specific day), expressed as a percentage.
- Conditions: Conditions describe the overall weather conditions prevailing in Portugal, which include cloud cover.
- Description: Description provides additional details or information about the weather conditions in Portugal, such as the state of a specific day.
- Icon: Icon represents a visual representation or symbol used to depict the weather conditions in Portugal.
- Stations: Stations denote the specific locations or weather stations from which the data or observations are recorded in Portugal.

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

1. ZIPPED Code Distribution in CSE-Moodle (it should also contain a README file mentioning the instructions to run your program).
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. Please report the approximate time taken by your program to run all the steps in a reasonable PC configuration.
3. ZIPPED experimental outcomes containing the following:
    a. 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)

    b. A file **agglomerative.txt** that contains final cluster information from step 4 in the same format as kmeans.txt.

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 name your file as <RollNo_ProjectCode.extension>
   (e.g., 23CS10000_PWHC-AC.pdf or 23CS30000_PWHC-AC.zip).
4. The submitted program file should have the following header comments:
   # Roll Number: Name of the student
   # Project Code
   # Project Title

You should not use any code available on the Web. Submissions found to be plagiarized 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 TA:
Kajori Ghosh (Email: kajorighosh4@gmail.com)