Programming Assignment: Hierarchical Clustering

Problem Description

In this programming assignment, we'll be implementing the agglomerative hierarchical clustering algorithm. We will implement three different cluster similarity measures: **Single link**, **Complete link**, and **Average link**.

We will be using geographical data for this assignment. Each of the data points is a 2D vector, with longitude and latitude as its dimensions.

Here is a brief review of agglomerative clustering. During the clustering process, we iteratively aggregate the most similar two clusters until there are K clusters left. For initialization, each data point forms its own cluster.

The similarity of two clusters C_i, C_j is determined by a distance measure. We use the following three measures in this assignment:

Single link:

$$D(C_i, C_j) = \min\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

Complete link:

$$D(C_i, C_j) = \max\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

Average link:

$$D(C_i, C_j) = \mathrm{mean}\{d(\mathbf{v}_p, \mathbf{v}_q) \mid \mathbf{v}_p \in C_i, \mathbf{v}_q \in C_j\}$$

The smaller the distance is, the more similar the two clusters are.

In the equations, $d(\cdot, \cdot)$ is a distance measure between two data points. In this assignment, for simplicity, we use the **Euclidean distance**, defined by:

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{\sum_i (p_i - q_i)^2}$$

where p_i, q_i are dimensions of $\mathbf{p}, \mathbf{q}$.

Your task: complete the missing functions in the given programming template files.

Programming Template

Find in `template/`

You are allowed to use *one of three* programming languages for this assignment: Python (3.10), C++ (14), or Java (JDK version 17). You are provided with the template code for each language containing descriptions for the expected input and output for each function that you would have to write.

Sample Input and Output

Find in `sample_test_cases/`

To aid your understanding, we have included one test case for each of Single Link, Complete Link, and Average Link clustering within this problem statement. The input format is as follows:

The first line of the input contains three space-separated integers N, K, M:

1. N: The number of data points (lines) following the first line.

2. K: The number of output clusters.

3. M: The cluster similarity measure. $M = 0$ for single link, $M = 1$ for complete link, $M = 2$ for average link.

Starting from the second line, each line will have exactly two floating point numbers, representing the longitude and latitude of a location. Each line corresponds to a 2D data point which will be fed into the clustering algorithm.

Thus, in the input file, there would be $N + 1$ lines in total. The objective would be to use the similarity measure specified by M to cluster the N data points into K clusters.

Note here that the test case format is provided for you to understand the test case files that will be used. For whichever language you use, you would NOT have to read in test cases from STDIN. That will be handled by the grader. You simply need to complete the required functions in the template code file.

Test Scale

For all public and hidden test cases, $N \leq 100, K \leq 40$.

Allowed Libraries

For each language, only standard built-in libraries would be usable. For example, if using Python, you would NOT be able to use libraries like NumPy or Scikit-Learn.

Submission

You need to submit ONE completed template code file to Gradescope. `submission.py` or `submission.cpp` or `Submission.java`. Note the capitalization for the `.java` file.

On Gradescope, your code will be tested on a list of public test cases, for which, you will be able to view both the input and the expected output. At the end of the submission deadline, you will be able to see the results produced by your code on a list of hidden test cases.