CS 412: Introduction to Data Mining **Spring 2020**

Homework 4

Handed Out: Apr 17, 2020 *12:00 pm* *Due: May* 01, 2020 *11:59 pm*

1 General Instructions

- The programming assignment will be hosted on HackerRank (`https://www.hackerrank.com/`) as a programming contest. To participate in this contest, please open a HackerRank account with your illinois.edu email netid.
- It is OK to discuss the problems with your classmates. However, it is NOT OK to work together or share code. You need to write your code independently and the TAs will not do the code debugging. Plagiarism is an academic violation to copy, to include text from other sources, including online sources, without proper citation. To get a better idea of what constitutes plagiarism, consult the CS Honor code (`http://cs.illinois.edu/academics/honor-code`) on academic integrity violations, including examples, and recommended penalties. There is a zero tolerance policy on academic integrity violations. Any student found to be violating this code will be subject to disciplinary action.
- Please use Piazza if you have questions about the homework.

2 Instructions

In this programming assignment, you are required to implement two clustering algorithms and apply them on multiple datasets. Participate in the programming contest hosted at HackerRank: `www.hackerrank.com/hw4-cs412-s20`.

- Please read the problem description carefully.
- The input will always be valid. We are mainly testing your understanding of data clustering, not your coding skills.
- Please pay special attention to the output format. We will be using the HackerRank based auto-grader, and it is extremely important that your generated output satisfies the requirement.
- We don't have specific constraints for this programming question. The only constrains are the standard environment constraints in HackerRank: `https://www.hackerrank.com/environment`.
- The grading will be based on if you pass the test cases. You are provided with one sample test case to debug your code. For the final grading, we will use additional test cases to test your code.

3 Programming Assignment

This assignment aims to familiarize you with the mechanism of two widely-used clustering methods: k-means and *AGNES* (a *single-link* based agglomerative nesting) algorithms.

k-means is an iterative clustering algorithm that aims to cluster points such that every point is assigned to its nearest cluster. Please use Euclidean distance to compute the distance between any pair of points.

For *AGNES*, you are asked to employ the hierarchical clustering algorithm to cluster N data points into k clusters. In other words, you shall always compute the agglomerative clustering via single linkage, and then stop the clustering when number of clusters reaches k, where $k < N$.

For **k-means**, the input shall comprise of N points and k initial cluster centroids. The points and the clusters are 0-indexed separately. In other words, the first point has index 0 and the second point has index 1 and so on. Similarly, the clusters are indexed from 0 to $k - 1$.

For **AGNES**, the input shall comprise of N points. The points are 0-indexed. In the absence of external clusters in case of AGNES, we shall assign `cluster-id` to a cluster of points as the minimum index of the points in the cluster. For example, if a cluster is composed of points {2, 3, 8, 9}, the `cluster-id` of the above group of nodes will be 2.

[The above design choice in implementation purely aims to ease auto-grading on HackerRank.]

Same for the previous assignments, you are not supposed to use any package (i.e. *numpy*, *pandas* in python) in this assignment.

Since we are to use HackerRank for grading, we have to eliminate additional randomness and generate the deterministic results. We therefore enforce the following rule in this assignment:

- For the k-means algorithm, we assign the points to the smallest indexed cluster in case of ties (i.e., when the distances are same). In other words, you need to break ties by assigning a point to the cluster with the lowest index if there are several equidistant clusters.
- For *AGNES*, we resolve the ties when merging two clusters using their `cluster-ids`. For example, consider two separate cluster pairs $P_1 = (c_1, c_2)$ and $P_2 = (c_3, c_4)$ that have the the same link separation i.e. distance between c_1 and c_2 is same as that between c_3 and c_4. We shall choose the smaller pair to combine in this case. We define the smaller pair by the following rule:

if $min(c_1, c_2) < min(c_3, c_4)$, then P_1 is smaller
if $min(c_1, c_2) == min(c_3, c_4)$ and $max(c_1, c_2) < max(c_3, c_4)$, then P_1 is smaller
else P_2 is smaller.

For example, if $P_1 = (0, 2)$ and $P_2 = (1, 2)$, then P_1 is smaller . If $P_1 = (0, 2)$ and $P_2 = (0, 1)$, then P_2 is smaller.

For all test cases in this assignment, we guarantee that deterministic results can be generated as long as the aforementioned requirements are satisfied.

Input Format and Sample

We ensure that the labels for N input points and k cluster are named by **non-negative integer** following zero-based numbering.

The first line of input will be N and k (space in between). This will be followed by N input points where the points co-ordinates are space separated. The data type of the input points is floating number. The k initial cluster points for k-means method shall follow the input points. For AGNES, the last k lines will not be there.

We provide the following toy input example alongside visualization in Figure 1.

```
10 2
8.98320053625 -2.08946304844
2.61615632899 9.46426282022
1.60822068547 8.29785986996
8.64957587261 -0.882595891607
1.01364234605 10.0300852081
1.49172651098 8.68816850944
7.95531802235 -1.96381815529
0.527763520075 9.22731148332
6.91660822453 -3.2344537134
6.48286208351 -0.605353440895
3.35228193353 6.27493570626
6.76656276363 6.54028732984
```

In this example the goal of the clustering task is to find the groups among 10 data points as illustrated in Figure 1. The k value for this example is 2, and we provide two random selected initial points at the end of the input.

Output Format and Sample

The output is the clustering results on the provided data made by k-means and $AGNES$. For each method, print the `cluster-id` corresponding to each point on separate lines.

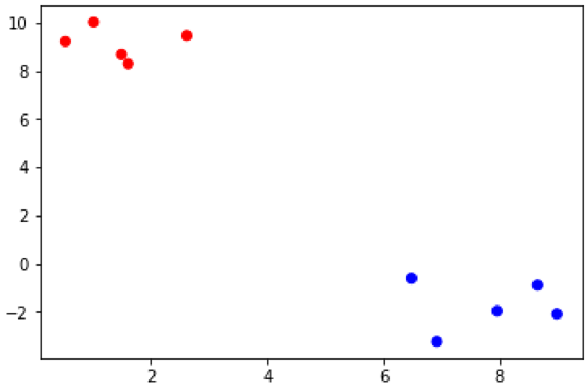

Figure 1: Visualization of the input of the toy example.

As an example, the outputs of the toy example are as follows. For K-Means, the sample output is:

```
1
0
0
1
0
0
1
0
1
1
```

For AGNES, the sample output is:

```
0
1
1
0
1
1
0
1
0
0
```

4 Grading Rubric

There are 5 test cases in total for this programming assignment. The total credit for this assignment is 70pt, and the extra credit is 30pt.

- input00 (sample): 10pt
- input01: 20pt
- input02: 20pt
- input03: 20pt
- input04 (extra): 30pt