The University of Newcastle
School of Electrical Engineering and Computer Science

COMP2230 Introduction to Algorithmics

Assignment

Total mark: 100
Due 30 October 2016
11:59pm, in the Blackboard

You have been contracted by emergency services to develop an algorithm they can use to place temporary emergency stations in the case of natural disaster. They would provide you with a map of hotspots affected by a disaster and a maximum number of stations they are able to support at that time.

In the first instance, they want to ensure that if two hotspots belong to different emergency stations, they are as far from each other as possible. This is important as natural disasters such as fire can easily spread and in general it is not possible to fully synchronise the efforts of all neighboring emergency stations.

In the second instance, it is also important that hotspots belonging to the same emergency station are as close together as possible so that emergency services can move from one hotspot to another. Therefore your program should suggest the number of emergency stations to maximize the ratio of distances between hotspots belonging to different emergency stations and those belonging to the same station.

DATA MINING BACKGROUND

1. Your task is to implement an efficient algorithm for solving a clustering problem.
2. The clustering problem is defined as follows. Given a set of items, $a_1, a_2, \ldots, a_n$, and a set of distances between the items, $d_{i,j}=d(a_i, a_j), i,j \in [1,n],$ find the optimal clustering of the items.
3. Distance between two items is a measures of dissimilarity between the two items. If the items are points in a two-dimensional plane, distance can be defined as the Euclidean distance.
4. Inter-clustering distance (InterCD) is the minimum distance between any two items belonging to different clusters. Formally, $InterCD = \min \{ d(a_i, a_j) \mid i,j \in [1,n] \text{ AND } i \neq j \text{ AND } a_i \text{ and } a_j \text{ belong to different clusters}\}$.
5. Similarly, intra-clustering distance is (IntraCD) is the maximum distance between any two items belonging to the same cluster. Formally, $IntraCD = \max \{ d(a_i, a_j) \mid i,j \in [1,n] \text{ AND } i \neq j \text{ AND } a_i \text{ and } a_j \text{ belong to the same cluster}\}$.
6. There are different ways to define optimal clustering. We will call clustering optimal if it maximizes the inter-clustering distance for the given number *k* of clusters.

7. The centroid of the cluster is the “average” point in the cluster. The x-coordinate of the centroid is the average of the x-coordinates of all the points in the cluster; similarly, the y-coordinate of the centroid is the average of the y-coordinates of all the points in the cluster.

TASKS

Task 1 – Individual and pairs assignment: Use Kruskal’s algorithm for finding a minimum spanning tree to find an optimal clustering (clustering that maximizes the inter-clustering distance) of the given set of *n* points in a two-dimensional plane, for the given number *k* of clusters.

Task 2 – Pairs assignment only: Among the optimal clusterings obtained by Kruskal’s algorithm for different values of k, $2 \leq k \leq n\text{-}1$, find the one that maximizes the ratio $\frac{InterCD}{IntraCD}$ between the inter-clustering distance and intra-clustering distance.

INPUT and OUTPUT

1. Your program should first read the input file *input.txt*.

2. Then you should greet the user and prompt them for the number of clusters.

3. If you are working in pairs, the user should also have an option to have the number of clusters automatically computed to maximize inter-clustering distance to intra-clustering distance ratio.

4. The following is an example of the input file *input.txt*.

```
1    1    1
2    2    2
3    3    5
4    7    8
5    8    7
```

The first number on each line is the hotspot ID number. The second number is the x-coordinate and the third is the y-coordinate of the hotspot.

5. The following is an example of what your program should output for the input file *input.txt* given above. User input is presented in brackets – if user enters 5, it is written as <5>. Please observe that your program should be robust and display error messages if the user chooses an invalid input. You need to strictly follow this example.

```
*********************INDIVIDUAL   ASSIGNMENT*********************************
Hello and welcome to Kruskal's Clustering!

There are 5 hotspots.

How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter 0 to exit.)

<2>

Station  1:
Coordinates: (2.00, 2.67)
Hotspots: {1,2,3}

Station 2:
Coordinates: (7.50, 7.50)
Hotspots: {4,5}

Inter-clustering distance: 5.00

How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter 0 to exit.)

<-1>

Entry not valid.

How many emergency stations would you like?
(Enter a number between 1 and n to place the emergency stations.
Enter 0 to exit.)

<1>

Station  1:
Coordinates: (4.20, 4.60)
Hotspots: {1,2,3,4,5}

Inter-clustering distance: Not applicable.

How many emergency stations would you like?
(Enter a number between 1 and n to place the emergency stations.
Enter 0 to exit.)

<0>
Thank you for using Kruskal's Clustering. Bye.
************************************************************************************
```

The following is the output for the same input file for the pairs assignment.

**********************PAIRS ASSIGNMENT*******************************

```
Hello and welcome to Kruskal's clustering!

There are 5 hotspots.

The weighted graph of hotspots:

   0.00   1.41   4.47   9.22   9.22
   1.41   0.00   3.16   7.81   7.81
   4.47   3.16   0.00   5.00   5.39
   9.22   7.81   5.00   0.00   1.41
   9.22   7.81   5.39   1.41   0.00


How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter -1 to automatically select the number of emergency stations.
Enter 0 to exit.)

<2>

Station  1:
Coordinates: (2.00, 2.67)
Hotspots: {1,2,3}

Station 2:
Coordinates: (7.50, 7.50)
Hotspots: {4,5}

Inter-clustering distance: 5.00

How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter -1 to automatically select the number of emergency stations.
Enter 0 to exit.)

<-1>

Number of stations: 3, 4

Station  1:
Coordinates: (1.50, 1.50)
Hotspots: {1,2}

Station 2:
```

```
Coordinates: (3.00,5.00)
Hotspots: {3}

Station 2:
Coordinates: (7.50, 7.50)
Hotspots: {4,5}


InterCD/IntraCD = 2.24

How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter -1 to automatically select the number of emergency stations.
Enter 0 to exit.)

<1>

Station  1:
Coordinates: (4.20, 4.60)
Hotspots: {1,2,3,4,5}

Inter-clustering distance: Not applicable.

How many emergency stations would you like?
(Enter a number between 1 and 5 to place the emergency stations.
Enter -1 to automatically select the number of emergency stations.
Enter 0 to exit.)

<0>

Thank you for using Kruskal's Clustering. Bye.
```

****

Note that when the user chose option -1 in the Pairs Assignment, both 3 and 4 was shown as the number of emergency stations that achieves the maximum ratio $\frac{InterCD}{IntraCD}$, and that subsequently stations were reported only for the first one of them (3 stations).

SUBMISSION
Submission mode will be announced separately on Blackboard.

ASSESSMENT CRITERIA

INDIVIDUAL ASSIGNMENT

1. Graph construction: 15 marks
   a. 10 marks for the correct algorithm and graph
   b. 3 marks for the correct presentation of the graph
   c. 2 marks for the well-commented and clear implementation

2. Kruskal's Algorithm: 50 marks
   a. 40 marks for the correct algorithm
   b. 5 marks for the correct presentation
   c. 5 marks for the well-commented and clear implementation

3. Positioning of emergency stations: 15 marks.

4. Calculation of inter-clustering distance: 10 marks.

5. Overall presentation and interaction with the user: 10 marks.

PAIRS ASSIGNMENT

6. Graph construction: 10 marks
   a. 5 marks for the correct algorithm and graph
   b. 3 marks for the correct presentation of the graph
   c. 2 marks for the well-commented and clear implementation

7. Kruskal's Algorithm: 40 marks
   a. 30 marks for the correct algorithm
   b. 5 marks for the correct presentation
   c. 5 marks for the well-commented and clear implementation

8. Positioning of emergency stations: 10 marks.

9. Calculation of inter-clustering distance: 10 marks.

10. Finding the number of clusters to maximize $\frac{InterCD}{IntraCD}$: 20 marks.

11. Overall presentation and interaction with the user: 10 marks.