CSCI3220 Algorithms for Bioinformatics 2018-2019 First Term
The Chinese University of Hong Kong

# Assignment 5

Due date: 12 December 2018 (Wed) 23:59 Full mark: 100
Expected time spent: 1-3 hours

Aims:
1. Get familiar with concepts and data structures related to data clustering.
2. Think about some other aspects of clustering algorithms not directly covered in class.
3. Find ways to make tedious calculations easy and less error-prone.

**Description:**
In this assignment, you will work out some details of using the quad tree structure in hierarchical clustering. You are encouraged to use some ways (which could be, but not necessarily, writing a small program) to do some tedious calculations. You will then implement the min-heap data structure. Finally, you will answer some questions related to the k-means clustering method.

**Questions:**

1. This question is about hierarchical clustering. You are given the following gene expression data matrix $X$ of five genes $g_1$-$g_5$ measured in six samples $s_1$-$s_6$.

| $X$ | $s_1$ | $s_2$ | $s_3$ | $s_4$ | $s_5$ | $s_6$ |
|---|---|---|---|---|---|---|
| $g_1$ | 4 | 6 | 5 | 1 | 2 | 8 |
| $g_2$ | 1 | 4 | 6 | 8 | 1 | 3 |
| $g_3$ | 2 | 9 | 3 | 5 | 1 | 6 |
| $g_4$ | 8 | 5 | 2 | 1 | 3 | 4 |
| $g_5$ | 7 | 1 | 1 | 3 | 2 | 9 |

   (a) Produce the pair-wise distance matrix among the five genes, where the distance between any two genes $g_i$ and $g_j$ is defined as their squared Euclidean distance, $d(g_i, g_j) = \frac{\sum_{k=1}^{6}(X_{ik}-X_{jk})^2}{6}$, in which $x_{ik}$ is the expression level of gene $g_i$ in sample $s_k$. Give distance values up to 1 decimal point. (10%)

   (b) Suppose you want to perform agglomerative hierarchical clustering of the five genes based on the distance matrix in Part a, using a quad tree to index the distance values. Show the initial quad tree, with the different levels of the tree clearly indicated. (10%)

   (c) For each of the following link types, show the updated quad tree after the first merge of two clusters: i) single-link; ii) average-link; iii) complete link. For the two merging clusters, the original entries for the one with the smaller index will be used to store the distances of the new cluster. (15%)

2. Write a computer program called **MinHeap** in C, C++, Java or Python that builds a min-heap for a list of integers and remove the smallest integer one by one until the heap is empty. Specifically, the input includes the followings:
   - The number of integers on the list
   - The integers, one per line, which can be assumed to be all unique

   The program should treat the input integers as an array that follows their input order and turn it into a heap. The program should print to standard output (stdout) this initial heap and the updated heap after removing each smallest value. Each heap should be printed as a single comma-delimited list on a separate line.

For example, if your implementation is Java, below is a typical run of the program (based on an example from the lecture notes):

```
>java MinHeap
10
13
5
10
4
11
1
9
12
8
6
1,4,9,5,6,10,13,12,8,11
4,5,9,8,6,10,13,12,11
5,6,9,8,11,10,13,12
6,8,9,12,11,10,13
8,11,9,12,13,10
9,11,10,12,13
10,11,13,12
11,12,13
12,13
13
```

The main part of your program is expected to contain no more than 100 lines of code.

Your program will be graded based on i) whether it can be compiled/run successfully, ii) whether it follows the input/output formats as specified above, iii) its accuracy on a number of test cases and iv) whether the program is well-documented with appropriate comments added to explain the meaning of the code. (50%)

3. This question is about k-means.

   (a) Suppose the data set contains an outlier point that is very different from all the other points. Describe how it would affect the clusters produced by k-means. (5%)

   (b) Suppose the data set contains a column with values much larger in magnitude than the other columns. Describe how it would affect the clusters produced by k-means. (5%)

   (c) There is another clustering method that is similar to k-means, but instead of assigning every point to the cluster with the closest representative, it assigns every point to each cluster in a fuzzy manner according to its distance to the cluster representative. For example, suppose k=2 and a point has a distance of 10 from the representative of cluster 1 and a distance of 20 from the representative of cluster 2, it is assigned a membership value of (1/10)/[(1/10)+(1/20)] = 2/3 to cluster 1 and a membership value of (1/20)/[(1/10)+(1/20)] = 1/3 to cluster 2. Describe one advantage and one disadvantage of this method as compared to k-means. (5%)

   (d) [Optional] In practice, the number of clusters is often not known. Propose a way to determine an appropriate value of k based on the input data set. (bonus 5%)

**Submission:**

For the programming question, you should first submit your program to our **online judge system**. Please see tutorial notes set 1 for explanations.

For the non-programming question, give all your answers in a single file named <ID>_asmt5.<ext>, where <ID> is your student ID and <ext> is either doc, docx or pdf. We prefer pdf files because it has better portability. Then put your files for both the programming and non-programming questions in a zip file named <ID>_asmt5.zip and submit it to Blackboard.

Both your written and source code files should contain the following header. Contact Kevin before submitting the assignment if you have anything unclear about the guidelines on academic honesty.

```
CSCI3220 2018-19 First Term Assignment 5

I declare that the assignment here submitted is original except for source
material explicitly acknowledged, and that the same or closely related material
has not been previously submitted for another course. I also acknowledge that I
am aware of University policy and regulations on honesty in academic work, and
of the disciplinary guidelines and procedures applicable to breaches of such
policy and regulations, as contained in the following websites.

University Guideline on Academic Honesty:
http://www.cuhk.edu.hk/policy/academichonesty/

Student Name: <fill in your name>
Student ID  : <fill in your ID>
```

**Marking Scheme and Notes:**

1. Remember to submit your assignment by 23:59pm of the due date. We may not accept late submissions.

2. For the written part, if you submit multiple times, **ONLY** the content and time-stamp of the **latest** one before the submission deadline will be considered. For the program, the version submitted to the online judge system with the highest score will be graded.

**University Guideline for Plagiarism**

Please pay attention to the university policy and regulations on honesty in academic work, and the disciplinary guidelines and procedures applicable to breaches of such policy and regulations. Details can be found at http://www.cuhk.edu.hk/policy/academichonesty/. With each assignment, students will be required to submit a statement that they are aware of these policies, regulations, guidelines and procedures.