| **CS 412: Introduction to Data Mining** | **Spring 2018** |
|---|---|
| Homework 2 | |
| *Handed Out: February* 21, 2018 | *Due: March* 7, 2018 *11:59 pm* |

# 1 General Instructions

- This assignment is due at 11:59 PM on the due date. We will be using Compass (http://compass2g.illinois.edu) for collecting the non-programming part of this assignment. Contact TAs if you face technical difficulties in submitting the assignment. We shall NOT accept any late submission!
- The non-programming part of homework **MUST** be submitted in pdf format. Handwritten answers are not acceptable. Name your pdf file as YourNetid-HW2.pdf
- For Question 1 and Question 2, you need to explain the logic of your answer/result for every subquestion. A result/answer without any explanation will not receive any points.
- The programming part of this assignment will be hosted on hackerrank (https://www.hackerrank.com/) as a programming contest. To participate in this contest, please use the same hackerrank account that you used for HW1. The contest framework will allow you to verify the correctness of your submission based on a set of sample test cases. We may use additional test cases to grade your submission.
- It is OK to discuss the problems with the TAs and your classmates, however, it is NOT OK to work together or share code. 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. Also feel free to send TAs emails and come to office hours.

# 2 Question 1 (1 point)

A data cube, $C$, has $d$ dimensions, and $n$ nonempty cells in the base cuboid. Assume that there are no concept hierarchies associated with the dimensions.

1. What could be the maximum number of cells (including both base cells and aggregate cells) in the data cube, $C$?
2. What could be the minimum number of cells (including both base cells and aggregate cells) in the data cube, $C$?

# 3 Question 2 (1 point)

Assume you are exploring a dataset with five attributes $\{A, B, C, D, M\}$, where $M$ is a continuous attribute, and the remaining attributes are binary. You are looking at a view $V_1$ that shows the SUM of $M$, for the two possible values of $A$. Now, if you perform two sequential operations on $V_1$: (i) **drill-down** by choosing an attribute from $\{B, C, D\}$ for de-aggregation, and (ii) **slicing** by choosing a filter criterion on one of the attributes from $\{A, B, C, D\}$, you would get a new view $V_2$. Now, consider another user who is looking at the same original view $V_1$. If s/he performs the same set of operations on $V_1$ (same de-aggregation attribute and filter criterion), however, in reverse order (first slicing, then drill-down), and gets a view $V_3$, can $V_2$ and $V_3$ be different? Explain why the two views can (can not) be different.

# 4 Question 3 (6 points)

Participate in the programming contest hosted at hackerrank:
https://www.hackerrank.com/contests/hw-2-sp-2018/.

**Please pay special attention to the output format. There will be no manual grading for the programming problem. So you need to strictly follow the formatting rules.**