Homework 2

CS 412: An Introduction to Data Warehousing and Data Mining
Fall 2018

Handed In: Oct 9th, 2018

- Feel free to talk to other members of the class when doing the homework. We are more concerned that you learn how to solve the problem than that you demonstrate that you solved it entirely on your own. You should, however, write down your solution yourself. Please try to keep the solution brief and clear.
- Please use Piazza first if you have questions about the homework. Also feel free to send us e-mails and come to office hours.
- The homework is due at 11:59 PM on the due date. We will be using Compass (`http://compass2g.illinois.edu`) for collecting homework assignments. Please do NOT hand in a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment. We do NOT accept late homework!
- The homework should be submitted in pdf format. If you use additional source code for solving problems, you are required to submit them and use the file names to identify the corresponding questions. For instance, 'Problem1.netid.py' refers to the python source code for Problem 1, replace netid with your netid. Compress all the files (pdf and source code files) into one zip file. Submit the compressed file ONLY. (If you did not use any source code, submitting the pdf file without compression will be fine)
- For each question, you will NOT get full credit if you only give out a final result. Necessary calculation steps are required. If the result is not an integer, round your result to 3 decimal places.

Problem 1. (**23** points total)
Suppose the base cuboid of a data cube contains two cells:
$(a_1, a_2, a_3, a_4, ..., a_9, a_{10}) : 1, (a_1, b_2, a_3, b_4, ..., a_9, b_{10}) : 1$
where $a_i \neq b_i$ for any i.

(a) (3 points) How many cuboids are there in this data cube?

(b) (5 points) How many (nonempty) closed cells are there in this data cube?

(c) (5 points) How many (nonempty) aggregate cells are there in this data cube?

(d) (5 points) How many (nonempty) aggregate closed cells are there in this data cube?

(e) (5 points) If we set minimum support = 2, how many (nonempty) aggregate cells are there in the corresponding iceberg cube?

Problem 2. (**25** points total)
We have the following data cube measures. Which of them are algebraic measures? Explain each of your selection and non-selection.

(a) (5 points) Standard deviation;

(b) (5 points) Average of Min and Max;

(c) (5 points) Sum of the largest 50 values;

(d) (5 points) Sum of the largest $\lfloor \frac{n}{1000} \rfloor$ values, where n is the number of data points in the current cell and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x;

(e) (5 points) Mode, if the data is guaranteed to be binary.

Problem 3. (**33** points total)
A database has 5 transactions listed in Table 1. Let $min\_sup = 0.6$ and $min\_conf = 0.7$.

Table 1:

trans_id	item_bought
001	{H, A, D, B, C}
002	{D, A, E, F}
003	{C, D, B, E}
004	{B, A, C, H, D}
005	{B, G, C}

(a) (5 points) List the frequent k-itemset for the largest k;

(b) (3 points) Show an itemset S, where (1) $\forall S_0 \subset S$ ($S_0 \neq \emptyset$), S_0 is frequent, and (2) S is not frequent;

(c) (5 points) List all the closed patterns;

(d) (5 points) List all the max-patterns;

(e) (5 points) List all the strong association rules (with support and confidence) with the following shape:

$x \in \{001, 002, ..., 005\}$, $buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3)$. $[s, c]$.

(f) (5 points) Now we want to mine frequent patterns using FP-Growth. We first find single item frequent patterns and sort them in frequency descending order. To break ties, we assume the order is $B - C - D - A$. Construct the FP-tree;

(g) (5 points) Show A's conditional (i.e., projected) database.

Problem 4. (**19** points total)
Suppose we are interested in only the frequent patterns satisfying certain constraints. For example, in Table 1, each item has its price. The price information is listed in Table 2. We still have $min\_sup = 0.6$.

Table 2:

item	A	B	C	D	E	F	G	H
price	10	20	40	30	90	90	30	50

(a) (5 points) Find all the *frequent* patterns S in Table 1 satisfying $sum(S.price) \geq 45$ (i.e., the sum of the price of all the items in S is no less than 45);

(b) (8 points) Is the constraint $sum(S.price) \geq 45$ monotonic or anti-monotonic? How about $sum(S.price) \leq 45$? Can you find an efficient method to mine *frequent* patterns with $sum(S.price) \leq 45$?

(c) (6 points) Let us focus on two other constraints $avg(S.price) \geq 30$ (i.e., the average price of all the items in S is no less than 30) and $avg(S.price) \leq 30$. Are they convertible? If so, how to convert them to anti-monotonic cases?