CS 412: Spring'21
Introduction To Data Mining

Assignment 2

(Due Thursday, March 18, 11:59 pm)

- Feel free to talk to other students of the class while doing the homework. We are more concerned that you learn how to solve the problems 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 Slack first if you have questions about the homework. You can also use CampusWire, send us e-mails, and/or come to our office hours. If you are sending us emails with questions on the homework, please cc all of us (Arindam, Carl, Jialong, and Qi) for faster response.
- The homework is due at 11:59 pm on the due date. We will be using Compass for collecting the homework assignments. Please submit your answers via Compass (`http://compass2g.illinois.edu`). Please do NOT email a hard copy of your write-up. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions!
- The homework should be submitted in pdf format.
- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary details, calculation steps, and explanations as appropriate.
- All the data can be download from Compass2g.

1. (18 points) Consider the following two datasets D_1, D_2 with sets of observations respectively on age of employees in company *AllElectronics* and salary of employees (as multiple of $\$1k$) in company *AllQuantums*:

   D_1: {13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70}

   D_2: {5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215}

   (a) (6 points) Use smoothing by bin means to smooth both of these datasets, using a bin depth of 3. Illustrate your steps. Comment on the effect of this technique for the given datasets in terms of the quality of approximation based on the variance of the bin.
   (1 points) Correct answer for D1:
   Bin 1: 14, 14, 14
   Bin 2: 18, 18, 18
   Bin 3: 21, 21, 21
   Bin 4: 24, 24, 24
   Bin 5: 26, 26, 26
   Bin 6: 33, 33, 33
   Bin 7: 35, 35, 35
   Bin 8: 40, 40, 40
   Bin 9: 56, 56, 56
   (1 point) Intermediate steps or sufficient explanation for D1
   (1 points) Correct answer for D2:
   Bin 1: 8, 8, 8
   Bin 2: 21, 21, 21
   Bin 3: 59, 59, 59
   Bin 4: 170, 170, 170
   (1 point) Intermediate steps or sufficient explanation for D2
   (2 point) Comment on the effect of techniques: a better approximation is achieved if data have low variance in each bin.

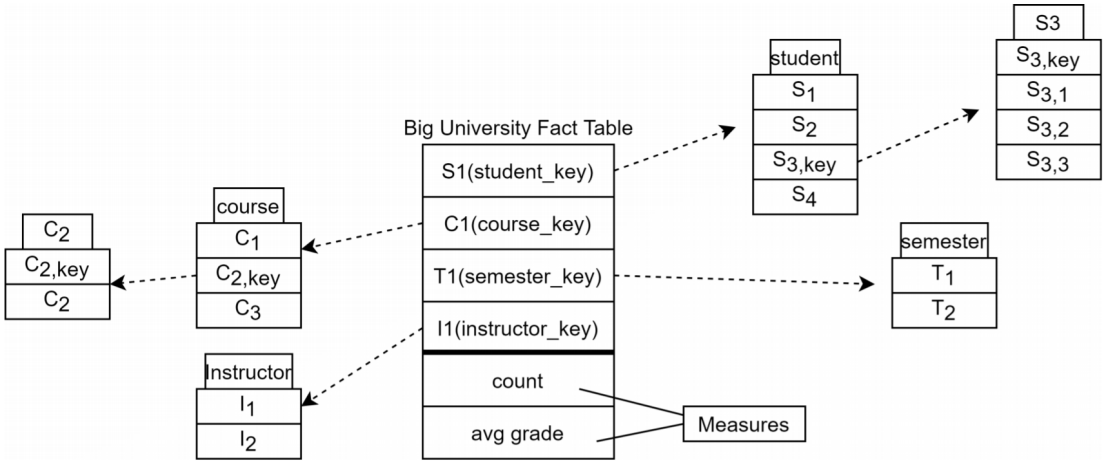
   (b) (12 points) Partition each of the datasets into three bins by each of the following methods, and comment on the effect of these techniquea for the given datasets in terms of the quality of approximation based on the variance of the bin:

   - (6 points) Equal-frequency (equal-depth) partitioning.
     (1 points) Correct answer for D1:
     Bin 1: 13, 15, 16, 16, 19, 20, 20, 21, 22
     Bin 2: 22, 25, 25, 25, 25, 30, 33, 33, 35
     Bin 3: 35, 35, 35, 36, 40, 45, 46, 52, 70
     (1 point) Intermediate steps or sufficient explanation for D1
     (1 points) Correct answer for D2:
     Bin 1: 5, 10, 11, 13
     Bin 2: 15, 35, 50, 55
     Bin 3: 72, 92, 204, 215
     (1 point) Intermediate steps or sufficient explanation for D2
     (2 point) Comment on the effect of techniques: equal-frequency has better approximation for high variance data. The outlier can't influence the partitioning a lot.

- (6 points) Equal-width partitioning.
  (1 points) Correct answer for D1:
  Bin 1: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30
  Bin 2: 33, 33, 35, 35, 35, 35, 36, 40, 45, 46
  Bin 3: 52, 70
  (1 point) Intermediate steps or sufficient explanation for D1
  (1 points) Correct answer for D2:
  Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72
  Bin 2: 92
  Bin 3: 204, 215
  (1 point) Intermediate steps or sufficient explanation for D2
  (2 point) Comment on the effect of techniques: equal-width has better approximation for low variance data. Few outliers cause less empty bins.

2. (18 points) Suppose that a data warehouse for Big University consists of the four dimensions: *student, course*, *semester*, and *instructor*, and two measures: *count* and *avg grade*. At the lowest conceptual level (e.g., for a given student, course, semester, and instructor combination), the *avg grade* measure stores the actual course grade of the student. At higher conceptual levels, avg grade stores the average grade for the given combination. Assume that *student* has attributes $S_1, S_2, S_{3,key}, S_4$, where $S_{3,key}$ has attributes $S_{3,1}, S_{3,2}, S_{3,3}, S_{3,key}$; *course* has attributes $C_1, C_{2,key}, C_3$, where $C_{2,key}$ has attributes $C_2, C_{2,key}$; *semester* has attributes T_1, T_2, and *instructor* has attributes I_1, I_2.

   (a) (5 points) Draw a *snowflake schema* diagram for the data warehouse, where the dimension tables will be based on the attributes of the dimensions.
   (5 point) 1 point for each dimension table and 1 point for the fact table.



   (b) (5 points) Starting with the base cuboid [*student, course, semester, instructor*], what specific OLAP operations (e.g., roll-up from semester to year) should you perform in order to list the average grade of CS courses for each Big University student.
   (2.5 point) Roll up on course from course_id to department and slice for department=CS. Get full credits if you mention slice on course to get CS courses.
   (2.5 point) Roll up to all for semester and instructor.

(c) (8 points) If each dimension has five levels (including all), such as "*student* < *major* < *status* < *university* < *all*" for *student*, how many cuboids will this cube contain (including the base and apex cuboids)?
(8 point) The number of cuboids $T = \prod_{i=1}^{n} L_i$, where n is the number of dimension and L_i is the level number for each dimension. Since $n = 4$ and $L_i = 5$, we have $T = 5^4 = 625$.

3. (18 points) Suppose we have a transaction database, TDB1, with the following transactions:
$T_1 == \{a_1, a_2, \ldots, a_{12}\}, T_2 = \{a_{10}, a_{11}, a_{20}\}, T_3 = \{a_1, a_2, \ldots, a_{20}\}, T_4 = \{a_1, a_2, \ldots, a_{30}\}$

(a) (6 points) For TDB1, how many closed patterns and maximal pattern(s) do we have and what are they if the minimum (absolute) support is 1? Justify your answer.
(4 points) closed patterns: $\{a_1, a_2, \ldots, a_{30}\} : 1$, $\{a_1, a_2, \ldots, a_{20}\} : 2$, $\{a_{10}, a_{11}, a_{20}\} : 3$, $\{a_1, a_2, \ldots, a_{12}\} : 3$, $\{a_{10}, a_{11}\} : 4$. -1 for each pattern missing, 0 score if only one pattern is correct.
(1 point) max pattern: $\{a_1, a_2, \ldots, a_{30}\} : 1$
(1 point) if you provide the support of each pattern or any kind of justifications besides the pattern list.

(b) (6 points) For TDB1, how many closed patterns and maximal pattern(s) do we have and what are they if the minimum (absolute) support is 2?

(4 points) closed patterns: $\{a_1, a_2, \ldots, a_{20}\} : 2$, $\{a_{10}, a_{11}, a_{20}\} : 3$, $\{a_1, a_2, \ldots, a_{12}\} : 3$, $\{a_{10}, a_{11}\} : 4$. -1 for each pattern missing.
(1 point) max pattern: $\{a_1, a_2, \ldots, a_{20}\} : 2$
(1 point) if you provide the support of each pattern or any kind of justifications besides the pattern list.

(c) (6 points) For TDB1, how many closed patterns and maximal pattern(s) do we have and what are they if the minimum (absolute) support is 4?

(2 points) closed patterns: $\{a_{10}, a_{11}\} : 4$.
(2 points) max pattern: $\{a_{10}, a_{11}\} : 4$.
(2 points) if you provide the support of each pattern or any kind of justifications besides the pattern list.

4. (28 points) Giving the following transaction database, we will focus on frequent pattern mining with minimum absolute support of 3.

TID	Items
T_1	A,B,C
T_2	A,D,E
T_3	B,D
T_4	A,B,D
T_5	A,C
T_6	B,C
T_7	A,C
T_8	A,B,C,D,E
T_9	B,C
T_{10}	A,D
T_{11}	A,B,C

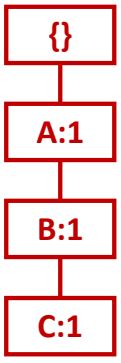
(a) (4 points) For an association rule $A \Rightarrow B(s, c)$, calculate its support s and confidence c.

(2 points) s = $P(A \cup B) = \frac{4}{11}$.
(2 points) c = $\frac{P(A \cup B)}{P(A)} = \frac{\frac{4}{11}}{\frac{8}{11}} = 0.5$.

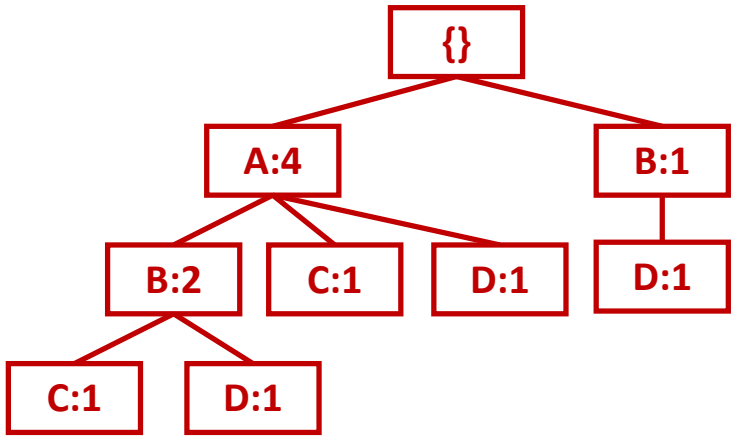
(b) (8 points) Find all frequent itemsets using Apriori algorithm. Please show all intermediate steps to get full credit.

(4 points), -1 for each missing, Frequent itemsets are $\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, C\}, \{A, D\}, \{B, C\}, \{B, D\}, \{A, B, C\}$.
(4 points) Show the support table for the database at each scan(length)

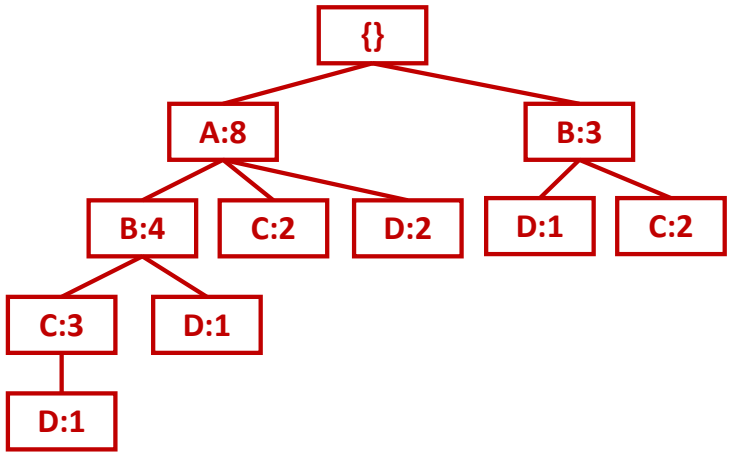
(c) (10 points) What is the FP-tree corresponding to transactions in Table 1? **Requirements:** Please insert transactions in the order of $T_1, T_2 ... T_{11}$. You need to demonstrate three FP-Trees after inserting 1-st, 5-th and the last transaction (T_{11}) to get the **full credit**.



(3 points) 1st FP-Tree



(3 points) 2nd FP-Tree



(4 points) Final FP-Tree

(d) (6 points) Find all frequent itemsets using the FP-Growth algorithm. **Requirements:** Please draw a table of conditional database derived from the FP-tree in the previous question (see Page.50 of slides 05FPBasic(Feb23)) to get the **full credit**.

2 points for each line below (i.e. each projected database)

Item	conditional database
B	{A:4}
C	{A,B:3}, {A:2}, {B:2}
D	{A,B,C:1}, {A,B:1}, {A:2}, {B:1}

The frequent itemsets are the same as (b). Correct steps but wrong answers will receive -1 score.

5. (18 points) Consider the following contingency table corresponding to two itemsets:
Note that common mistakes and their corresponding points are listed below the answer. (e.g. (1 point) on a 4 point problem would mean -3)

	A	¬A	Σ_{row}
B	a	b	a+b
¬B	c	d	c+d
Σ_{col}	a+c	b+d	a+b+c+d

(a) (6 points) What is $Kulc(A, B)$, the Kulczynski measure between A and B? Show that $Kulc(A, B)$ is null invariant.

(2 points) $Kulc(A,B) = \frac{1}{2}\left(\frac{\frac{a}{a+b+c+d}}{\frac{a+c}{a+b+c+d}} + \frac{\frac{a}{a+b+c+d}}{\frac{a+b}{a+b+c+d}}\right) = \frac{1}{2}\left(\frac{a}{a+c} + \frac{a}{a+b}\right)$

(1 point) If the answer is not simplified/fully-calculated to cancel out the total.

(4 points) Show the Kulczynski measure is null-invariant. It is null invariant because d (the number of nulls) cancels out of its formula.

(1 point) If the answer is not the full condition.

(b) (8 points) What is $Lift(A, B)$? Show that $Lift(A, B)$ is not null invariant. Based on $Lift(A, B)$, when will A, B be considered independent, in terms of the entries in the contingency table.

(2 points) $Lift(A,B) = \frac{\frac{a}{a+b+c+d}}{(\frac{a+c}{a+b+c+d})(\frac{a+b}{a+b+c+d})} = \frac{a(a+b+c+d)}{(a+c)(a+b)}$

(1 point) If the answer is not simplified to cancel out the total (note: one total $a + b + c + d$ should remain).

(2 points) Show that Lift is not null invariant. It is not null-invariant because d is part of its formula.

(1 point) If the correct version of lift is given but it having d (and therefore being null-variant) is not specified.

(4 points) Show when A and B are considered independent based on lift.

(2 points) For condition that is only one case of full answer.

(1 point) For saying $P(A \cup B) = P(A)P(B)$ (definition of independence) and not $ad = bc$.

(1 point) For having the correct equation and not solving. (e.g. Lift = 1)

(3 points) For not simplifying fully.

(3 points) You were otherwise correct except you misunderstood $\cup$ notation in this context.

$$Lift(A,B) = 1 = \frac{a(a+b+c+d)}{(a+c)(a+b)}$$

$$(a+c)(a+b) = a(a+b+c+d)$$

$$a^2 + ab + ac + bc = a^2 + ab + ac + ad$$

$$bc = ad$$

(c) (4 points) What is the difference between $Lift(A, B)$ and $Cosine(A, B)$? Why does such a difference make $Cosine(A, B)$ null invariant?
(2 points) $Cosine(A, B)$ takes the square root of the denominator.

(2 points) Explain why the difference makes Cosine null invariant.

$$Cosine(A, B) = \frac{\frac{a}{a+b+c+d}}{\sqrt{(\frac{a+c}{a+b+c+d})(\frac{a+b}{a+b+c+d})}} = \frac{a}{\sqrt{(a+c)(a+b)}}$$

The square root causes d to cancel out so $Cosine(A, B)$ is null invariant.