CS 412: Spring'21
Introduction To Data Mining

Assignment 2

(Due Tuesday, March 16, 11:59 pm)

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.

   For D_1:

   - **Step 1:** Sort the data. (This step is not required here as the data are already sorted.)
   - **Step 2:** Partition the data into equidepth bins of depth 3.

     Bin 1 : $13, 15, 16$ Bin 2 :$16, 19, 20$ Bin 3 : $20, 21, 22$
     Bin 4 : $22, 25, 25$ Bin 5 :$25, 25, 30$ Bin 6 : $33, 33, 35$
     Bin 7 : $35, 35, 35$ Bin 8 :$36, 40, 45$ Bin 9 : $46, 52, 70$

   - **Step 3:** Calculate the arithmetic mean of each bin.
   - **Step 4:** Replace each of the values in each bin by the arithmetic mean calculated for the bin.

     Bin 1 : $44/3, 44/3, 44/3$ Bin 2 :$55/3, 55/3, 55/3$ Bin 3 : $21, 21, 21$
     Bin 4 : $24, 24, 24$ Bin 5 :$80/3, 80/3, 80/3$ Bin 6 : $101/3, 101/3, 101/3$
     Bin 7 : $35, 35, 35$ Bin 8 :$121/3, 121/3, 121/3$ Bin 9 : $56, 56, 56$

   For D_2:

   - **Step 1:** Sort the data. (This step is not required here as the data are already sorted.)
   - **Step 2:** Partition the data into equidepth bins of depth 3.

     Bin 1 : $5, 10, 11$ Bin 2 : $13, 15, 35$
     Bin 3 : $50, 55, 72$ Bin 4 : $92, 204, 156$

   - **Step 3:** Calculate the arithmetic mean of each bin.

- **Step 4:** Replace each of the values in each bin by the arithmetic mean calculated for the bin.

$$\text{Bin 1}: 26/3, 26/3, 26/3 \qquad \text{Bin 2}: 63/3, 63/3, 63/3$$
$$\text{Bin 3}: 177/3, 177/3, 177/3 \qquad \text{Bin 4}: 452/3, 452/3, 452/3$$

The quality of approximation is better for low variance data.

(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.
For D_1:
Partition the data into equidepth bins of depth 9:

$$\text{Bin 1}: 13, 15, 16, 16, 19, 20, 20, 21, 22$$
$$\text{Bin 2}: 22, 25, 25, 25, 25, 30, 33, 33, 35$$
$$\text{Bin 3}: 35, 35, 35, 36, 40, 45, 46, 52, 70$$

For D_2:
Partition the data into equidepth bins of depth 4:

$$\text{Bin 1}: 5, 10, 11, 13 \qquad \text{Bin 2}: 15, 35, 50, 55 \qquad \text{Bin 3}: 72, 92, 204, 215$$

- (6 points) Equal-width partitioning.
For $D_1 : (70 - 13)/3 = 19$
Partitioning the data into 3 equi-width bins will require the width to be 19.
Then we have:

$$\text{Bin 1}: 13, 15, 16, 16, 19$$
$$\text{Bin 2}: 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36$$
$$\text{Bin 3}: 40, 45, 46, 52, 70$$

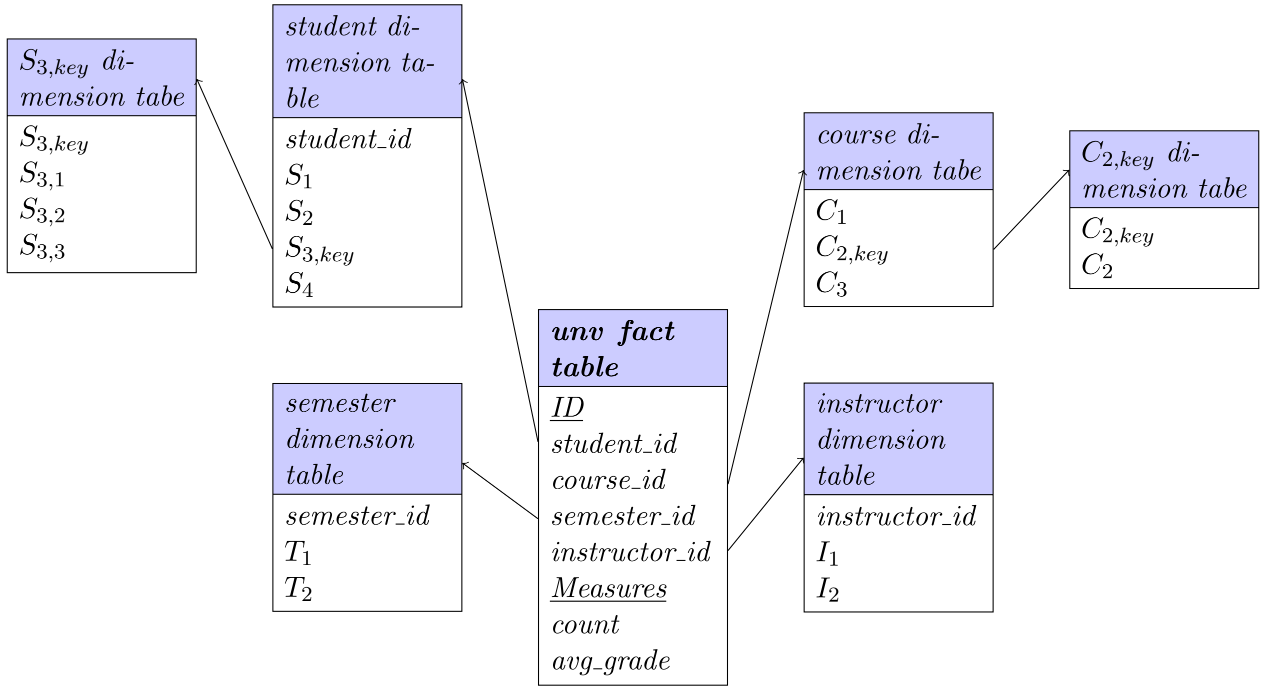
For $D_2 : (215 - 5)/3 = 70$
Partitioning the data into 3 equi-width bins will require the width to be 70.
Then we have:

$$\text{Bin 1}: 5, 10, 11, 13, 15, 35, 50, 55, 72 \qquad \text{Bin 2}: 92 \qquad \text{Bin 3}: 204, 215$$

The quality of approximation is better for low variance data for both methods.

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.



(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.

- Roll-up on course from *course_id* to I_2.
- Roll-up on semester from *semester_id* to *all*.
- Slice for course="CS".

(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)?

$5^4 = 625$. This cube will contain 625 cuboids.

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.

Five closed patterns: $\{a_{10}, a_{11}\} : 4, \{a_{10}, a_{11}, a_{20}\} : 3, \{a_1, a_2, \ldots, a_{12}\} : 3, \{a_1, a_2, \ldots, a_{20}\} : 2, \{a_1, a_2, \ldots, a_{30}\} : 1$.

One max pattern: $\{a_1, a_2, \ldots, a_{30}\} : 1$.

(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?

Three closed patterns: $\{a_{10}, a_{11}\} : 4, \{a_{10}, a_{11}, a_{20}\} : 3, \{a_1, a_2, \ldots, a_{12}\} : 3, \{a_1, a_2, \ldots, a_{20}\} : 2$.

One max pattern: $\{a_1, a_2, \ldots, a_{20}\} : 2$.

(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?

One closed pattern: $\{a_{10}, a_{11}\} : 4$.

One max pattern: $\{a_{10}, a_{11}\} : 4$.

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

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

$$\text{support}(A \Rightarrow B(s,c)) = P(A \cup B) = \frac{4}{11}$$

.

$$\text{confidence}(A \Rightarrow B(s,c)) = P(A \mid B) = \frac{\text{support}(A \cup B)}{\text{support}(A)} = \frac{4}{8} = \frac{1}{2}.$$

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

$C_1 = \{A, B, C, D, E\}$

$L_1 = \{A, B, C, D\}$

$C_2 = \{AB, AC, AD, BC, BD, CD\}$

$L_2 = \{AB, AC, AD, BC, BD\}$

$C_3 = \{ABC, ABD, ACD, BCD\}$

$L_3 = \{ABC\}$

$C_4 = \emptyset$

$L_3 = \emptyset$

Finally resulting in the complete set of frequent itemsets: $\{A, B, C, D, AB, AC, AD, BC, BD, ABC\}$.

(c) (10 points) What is the FP-tree corresponding to transactions in Table 1? Please show all intermediate steps to get full credit.

We first cancel E out since its frequency is 2, which is less than minimum absolute support. Then we sort each itemset according to the frequency of items from high to low.

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

The FP-trees are shown from Figure 1 to Figure 11. See Page 7 to Page 10.

(d) (6 points) Find all frequent itemsets using the FP-Growth algorithm. Please show all intermediate steps to get full credit.

$L = \{\{A:8\}, \{B:7\}, \{C:7\}, \{D:5\}\}$

$CPB(D) = \{\{A, B:1\}, \{A, B, C:1\}, \{A:2\}, \{B:1\}\}$

$CFPT(D) = \{\{A:4, B:2\}, \{B:2\}\}$

Generates FPs: $\{\{A, D:4\}, \{B, D:4\}\}$

$CPB(C) = \{\{A:2\}, \{A, B:3\}, \{B:2\}\}$

$CFPT(C) = \{\{A:5, B:3\}, \{B:2\}\}$

Generates FPs: $\{\{A, C:5\}, \{B, C:5\}, \{A, B, C:3\}\}$

$CPB(B) = \{\{A:4\}\}$

$CFPT(B) = \{\{A:4\}\}$

Generates FPs: $\{\{A, B:4\}\}$

Which finally results in the complete set of frequent itemsets:

$\{\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{A, D\}, \{B, D\}, \{A, C\}, \{B, C\}, \{A, B, C\}\}$.

5. (18 points) Consider the following contingency table corresponding to two itemsets:

	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.

$$Kulc(A, B) = \frac{1}{2}(P(A \mid B) + P(B \mid A)) = \frac{1}{2}\left(\frac{sup(A \cup B)}{sup(A)} + \frac{sup(A \cup B)}{sup(B)}\right) = \frac{1}{2}\left(\frac{a}{a+c} + \frac{a}{a+b}\right).$$

From the formula

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

we know that $Kulc(A, B)$ is independent of d, which is the null transaction. By definition, $Kulc(A, B)$ is null-invariant.

(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.

$$Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)} = \frac{sup(A \cup B)}{sup(A) \times sup(B)} = \frac{a}{(a+c)(a+b)/(a+b+c+d)} = \frac{a(a+b+c+d)}{(a+c)(a+b)}.$$

From the formula

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

we know that when d changes, $Lift(A, B)$ also changes. $Lift(A, B)$ is not free of null-transaction. Hence, $Lift(A, B)$ is not null-invariant.

A, B will be considered as independent if $Lift(A, B) = 1$, which indicates that

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

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

(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?

We have

$$Lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)},$$

and

$$Cosine(A, B) = \frac{P(A \cup B)}{\sqrt{P(A)P(B)}}.$$

The difference is that the denominator of $Lift(A, B)$ is $P(A)P(B)$ while the denominator of $Cosine(A, B)$ is $\sqrt{P(A)P(B)}$.

We also have

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

and

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

The term $a+b+c+d$ cancels out due to square root, making $Cosine(A, B)$ independent of d. Then $Cosine(A, B)$ is free of null-invariant. By definition, $Cosine(A, B)$ is null-invariant.

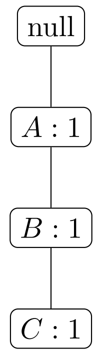


Figure 1: Insert T_1

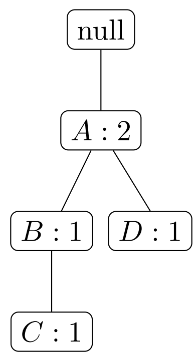


Figure 2: Insert T_2

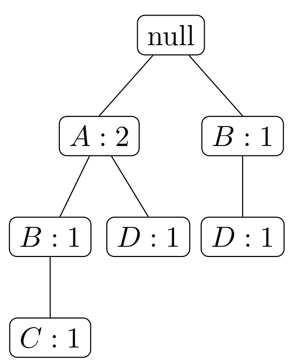


Figure 3: Insert T_3

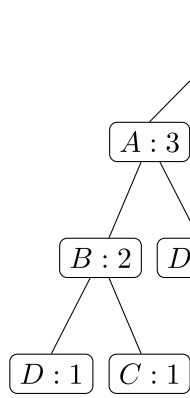


Figure 4: Insert T_4

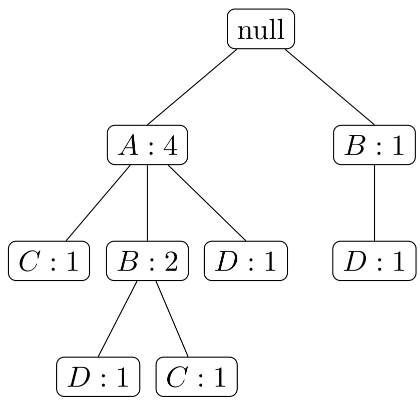


Figure 5: Insert T_5

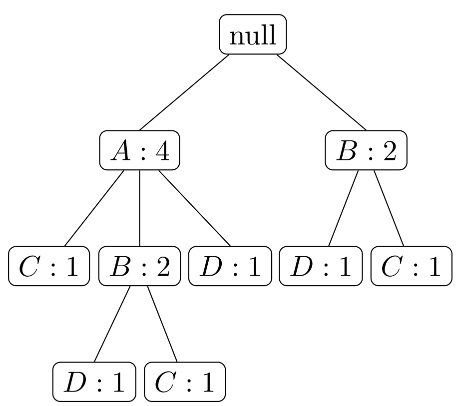


Figure 6: Insert T_6

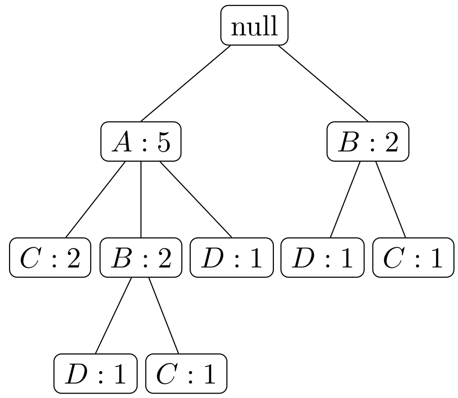


Figure 7: Insert T_7

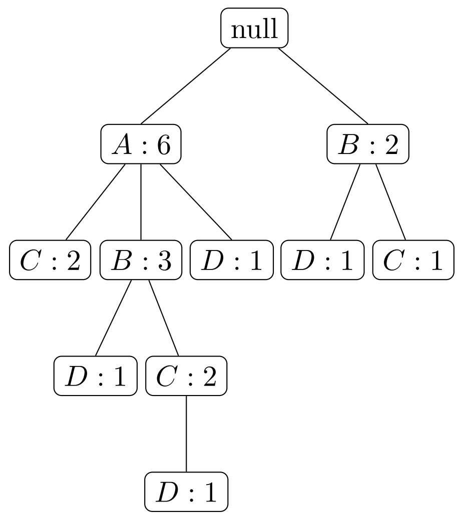


Figure 8: Insert T_8

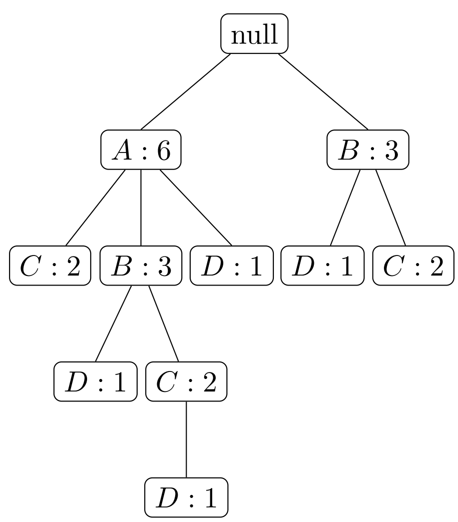


Figure 9: Insert T_9

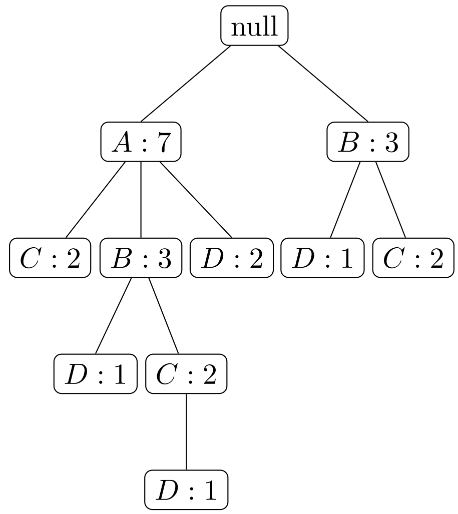


Figure 10: Insert T_{10}

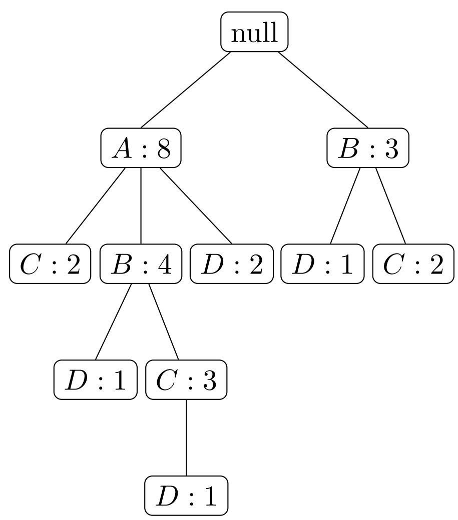


Figure 11: Insert T_{11}