CS 412: Fall'21
Introduction To Data Mining

Assignment 2

(Due Monday, October 11, 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.

- The homework is due at 11:59 pm on the due date. We will be using Gradescope for collecting the homework assignments. You should have been added to the Gradescope page for our class – if not, please email us. 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!

- Please use Slack or Slack first if you have questions about the homework. You can also send us e-mails, and/or come to our office (zoom) hours. If you are sending us emails with questions on the homework, please start subject with "CS 412 Fall'21: " and send the email to *all of us* (Arindam, Yikun, Dongqi, Zhe, and Hang) for faster response.

- The homework should be submitted in pdf format and there is no need to submit source code with details of your computation.

- For each question, you will NOT get full credit if you only give out a final result. Please show the necessary details, including any formulae you use, what the variables mean in the formulae, any derivation or calculation steps, and explanations as appropriate. Consider the following two examples:

Example 1 **Q:** Given a dataset $\mathcal{X} = \{3.1, 4.2, -1\}$, compute the mean?
A: For any set of n numbers $\mathcal{X} = \{x_1, \ldots, x_n\}$, the mean can be computed as $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$. For the given dataset $\mathcal{X}$, the mean is $\mu = \frac{3.1+4.2-1}{3} = 2.1$

Example 2 **Q:** A coin claimed to be unbiased has been tossed 100 times, with 54 heads and 46 tails. What is the χ^2 statistic?
A: For a categorical variable taking k possible values, if the expected values are $e_i, i = 1, \ldots, k$ and the observed values are $o_i, i = 1, \ldots, k$, then the χ^2 statistic can be computed as: $\chi^2 = \sum_{i=1}^{k} \frac{(o_i - e_i)^2}{o_i}$. For the problem, since the coin is claimed to be unbiased, the expected values are $50, 50$. Further, the observed values are $54, 46$. Then, the chi-squared statistic is given by $\chi^2 = \frac{(54-50)^2}{50} + \frac{(46-50)^2}{50} = 0.64$.

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) (3 points) Use smoothing by bin means to smooth both of these datasets, using equal depth binning with a bin depth of 3. Please round the bin means to the nearest integer. Illustrate your steps.

   [Step 1. Sort the data if the raw data is not sorted.]

   [Step 2. Since the depth of a bin is set to be 3, then we use equal-depth binning.]

   - D_1:
     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;
   - D_2:
     Bin 1: 5, 10, 11;
     Bin 2: 13, 15, 35;
     Bin 3: 50, 55, 72;
     Bin 4: 92, 204, 215;

   [Step 3. Use bin means for smoothing.]

   - D_1:
     Bin 1: 15, 15, 15;
     Bin 2: 18, 18, 18;
     Bin 3: 21, 21, 21;
     Bin 4: 24, 24, 24;
     Bin 5: 27, 27, 27;
     Bin 6: 34, 34, 34;
     Bin 7: 35, 35, 35;
     Bin 8: 40, 40, 40;
     Bin 9: 56, 56, 56;
   - D_2:
     Bin 1: 9, 9, 9;
     Bin 2: 21, 21, 21;
     Bin 3: 59, 59, 59;
     Bin 4: 170, 170, 170;

   (b) (3 points) 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.

   [Step 1. Analysis.]

   Since we are approximating each bin by the bin mean, variance of each bin is a measure of error

due to the approximation (because the variance of each bin is eliminated), and we want this to be low. Quality of approximation is good if the variance in each bin is small. Therefore, we consider the average variance of bins.

After aggregation, we have the average variance of bins of equal-depth partitioned D_1: 14.57, and the average variance of bins of equal-depth partitioned D_2: 820.61.

We can observe bins of D_1 having low average variance and bins of D_2 having higher average variance, a latent reason is that the last bin of D_2 carrying large divergence.

(c) (12 points) Partition each of the datasets into three bins by each of the following methods, and comment on the effect of these techniques 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.
- (6 points) Equal-width partitioning.

[Step 1. Sort the data if the raw data is not sorted.]

[Step 2. Use equal-depth binning and equal-width binning to partition raw data.]

- divides into 3 intervals, each interval should contain approximately same number of samples
  D_1(equal-depth):
  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;
- divides into 3 intervals, each interval should contain approximately same number of samples
  D_2(equal-depth):
  Bin 1: 5, 10, 11, 13;
  Bin 2: 15, 35, 50, 55;
  Bin 3: 72, 92, 204, 215;
- width of each interval = (70-13)/3 = 19, the interval of Bin 1 is [13, 32), of Bin 2 is [32, 51), and of Bin 3 is [51, 70]
  D_1(equal-width):
  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;
- width of each interval = (215-5)/3 = 70, the interval of Bin 1 is [5, 75), of Bin 2 is [75, 145), and of Bin 3 is [145, 215]
  D_2(equal-width):
  Bin 1: 5, 10, 11, 13, 15, 35, 50, 55, 72;
  Bin 2: 92;
  Bin 3: 204, 215;

[Step 3. Analysis.]

Since we are approximating each bin by the bin mean, variance of each bin is a measure of error due to the approximation (because the variance of each bin is eliminated), and we want this to be low. Quality of approximation is good if the variance in each bin is small. Therefore, we consider the average variance of bins.

After aggregation, we have the average variance of bins of

- equal-depth partitioned D_1: 48.83
- equal-depth partitioned D_2: 1460.02
- equal-width partitioned D_1: 40.47
- equal-width partitioned D_2: 184.61

Therefore, we can see that equal-width achieves lower average variance (i.e., approximation error) for assigning the a few extreme values to the last bin.

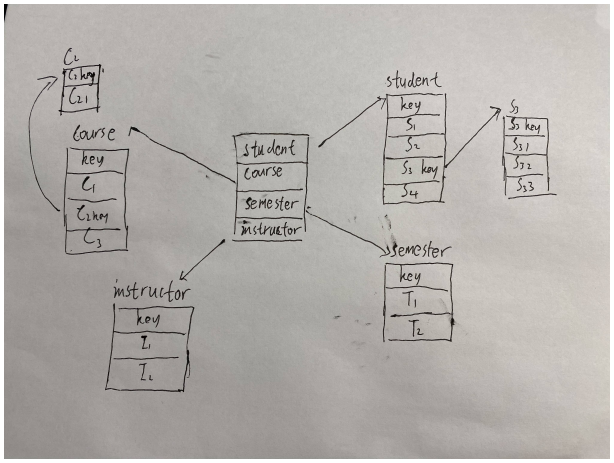
2. (18 points) Suppose that a data warehouse for Big Ten universities consists of the four dimensions: *student, course*, *semester*, and *instructor*, and two measures: *avg grade* and *count*. Each dimensions has a concept hierarchy, e.g., for *student*, the concept hierarchy is "*student* < *major* < *college* < *university* < *all* "; for *courses*, the concept hierarchy is "*course* < *department* < *college* < *university* < *all* ", etc.

   At the lowest conceptual level, i.e., for a given *student*, *course*, *semester*, *instructor* combination, the *avg grade* measure stores the actual course grade of the student. At higher levels of the concept hierarchy for one or more dimensions, *avg grade* stores the average grade for the given combination.

   (a) (5 points) Assume that *student* has attributes $S_1, S_2, S_{3,key}, S_4$, where $S_{3,key}$ has attributes $S_{3,key}, S_{3,1}, S_{3,2}, S_{3,3}$; *course* has attributes $C_1, C_{2,key}, C_3$, where $C_{2,key}$ has attributes $C_{2,key}, C_{2,1}$; *semester* has attributes T_1, T_2, and *instructor* has attributes I_1, I_2. Draw a *snowflake schema* diagram for the data warehouse, where the dimension tables will be based on the attributes of the dimensions.

   As described by Figure 1

Figure 1: For Question 2(a)



   (b) (5 points) Starting with the base cuboid [*student, course, semester, instructor*], what specific OLAP operations (e.g., roll-up, drill down, etc.) should you perform in order to list the average grade of Computer Science courses (i.e., department = "Computer Science" for dimension *course*) for each student.

   Roll up on course (from "course" to "department"); Slice for department = "computer science"

   (c) (8 points) If each of the four dimensions in the data warehouse has five levels (including all), e.g., "*student* < *major* < *status* < *university* < *all* " for *student*, how many cuboids will the data cube contain (including the base and apex cuboids)? Clearly justify your answer.

For each dimension associated with 5 levels, cuboid has 5 possible choices, which are the 4 levels and the "all".

Therefore, for a 4-dimension data cube, where each dimension has five levels (including all), the total number of cuboids is

$$5 \times 5 \times 5 \times 5 = 625.$$

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, for a minimum (absolute) support of 1, i.e., $minsup = 1$, how many closed patterns and maximal pattern(s) do we have and what are they? Clearly justify your answer.

   There are 5 closed patterns: CP_1 : "$\{a_1, a_2, \ldots, a_{20}\} : 2$", CP_2 : "$\{a_1, a_2, \ldots, a_{12}\} : 3$", CP_3 : "$\{a_1, a_2, \ldots, a_{30}\} : 1$", CP_4 : "$\{a_{10}, a_{11}\} : 4$", CP_5 : "$\{a_{10}, a_{11}, a_{20}\} : 3$"

   There is 1 maximal pattern: MP_1 : "$\{a_1, a_2, \ldots, a_{30}\} : 1$"

   (b) (6 points) For TDB1, for a minimum (absolute) support of 2, i.e., $minsup = 2$, how many closed patterns and maximal pattern(s) do we have and what are they? Clearly justify your answer.

   There are 4 closed patterns: CP_1 : "$\{a_1, a_2, \ldots, a_{12}\} : 3$", CP_2 : "$\{a_1, a_2, \ldots, a_{20}\} : 2$", CP_3 : "$\{a_{10}, a_{11}\} : 4$", CP_4 : "$\{a_{10}, a_{11}, a_{12}\} : 3$"

   There is 1 maximal pattern: MP_1 : "$\{a_1, a_2, \ldots, a_{20}\} : 2$"

   (c) (6 points) For TDB1, for a minimum (absolute) support of 4, i.e., $minsup = 4$, how many closed patterns and maximal pattern(s) do we have and what are they? Clearly justify your answer.

   There is 1 closed patterns: CP_1 : "$\{a_{10}, a_{11}\} : 4$"

   There is 1 maximal pattern: MP_1 : "$\{a_{10}, a_{11}\} : 4$"

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

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

   [Step 1. Calculate the relational support of A $\cup$ B: $s = 4$ (absolute support), $s = 4/11 = 0.36$ (relative support)]

   [Step 2. Calculate confidence of A $\rightarrow$ B: $c = sup(A, B)/sup(A) = 4/8 = 0.5$]

   (b) (8 points) Find all frequent itemsets using Apriori algorithm. Please use the alphabetical ordering A, B, C, D, E for the candidate generation (self-join and pruning) phase. Please show all intermediate steps (like how to scan and filter the data, how to generate candidates with the ordering you are referring to, and if any pruning tricks are available, until you reach your final result) to get full credit.

   [Step 1. Scan and filter to get frequent F1]

   - {A}: 8
   - {B}: 7
   - {C}: 7
   - {D}: 5

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

Table 1: Transaction Database.

[Step 2. Generate candidate C2]

- {A, B}
- {A, C}
- {A, D}
- {B, C}
- {B, D}
- {C, D}

[Step 3. Scan and filter to get frequent F2]

- {A, B}: 4
- {A, C}: 5
- {A, D}: 4
- {B, C}: 5
- {B, D}: 3

[Step 4. Generate candidate C3]

- {A, B, C}
- {A, B, D}
- {A, C, D}
- {B, C, D}

[Step 5. Scan and filter to get frequent F3] {A, B, C}: 3

[Step 6. Cannot generate C4, then, F4 is empty. Terminate.]

(c) (10 points) What is the FP-tree corresponding to transactions in Table 1?
Requirements: Please insert transactions in the order of $T_1, T_2, \ldots, T_{11}$. You need to draw three FP-Trees after inserting T_1, T_5, and T_{11} to get full credit.

[Step 1. scan DB once, find single item frequent pattern]

- {A}: 8
- {B}: 7
- {C}: 7

- {D}: 5

[Step 2. Sort frequent items in frequency descending order, f-list]

- F-list = A-B-C-D

[Step 3. Scan DB again, find the ordered frequent itemlist for each transaction]

TID	Items	Ordered, frequent itemlist
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

[Step 4. For each transaction, insert the ordered frequent itemlist into a FP-tree]

- After inserting T_1, we have Figure 2a.
- After inserting T_5, we have Figure 2b.
- After inserting T_{11}, we have Figure 2c.

(d) (6 points) Find all frequent itemsets using the FP-Growth algorithm.
Requirements: Please draw a table for the conditional database derived from the FP-tree in the previous question (e.g., see `Page 51` of Slides 05) to get full credit.

[Step 1. Find Conditional Pattern Base of D, C, and B, respectively]

[Step 2. Find Conditional FP-tree of D, C, and B, respectively]

[Step 3. Draw table as Figure 3 to find frequent patterns]

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

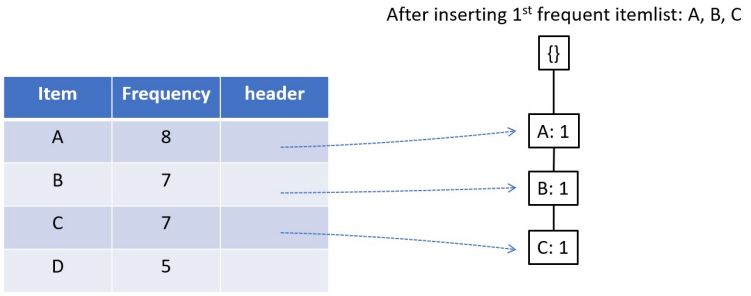
(a) (6 points) What is $Kulc(A, B)$, the Kulczynski measure between A and B? Show that $Kulc(A, B)$ is null invariant.
Based on the definition of Kulczynski measure we have:

$$Kulc(A, B) = \frac{1}{2}\left(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)}\right)$$

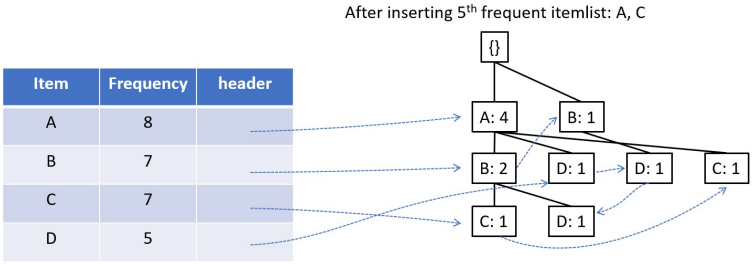
In our case we have $s(A) = \frac{a+c}{a+b+c+d}$, $s(B) = \frac{a+b}{a+b+c+d}$, $s(A \cup B) = \frac{a}{a+b+c+d}$. Hence, the Kulczynski measure can be represented as:

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

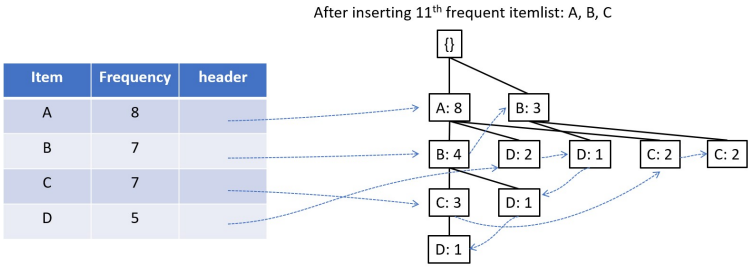
from which we observe that d (i.e., the number of null transactions) is not involved in the formula and it indicates that $Kulc(A, B)$ is null invariant.



(a) After inserting T1



(b) After inserting T5



(c) After inserting T11

Figure 2: Answer of Q4(c).

(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.

From the definition of Lift we have

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

where the d (i.e., the number of null transactions) is involved and it indicates that the measure is not null invariant. From the definition of Lift we know when $s(A \cup B) = s(A) \times s(B)$ A, B will be considered independent. In our case, the condition should be:

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

$$b \times c = a \times d$$

Item	Conditional Patten Base	Conditional FP-tree	Frequent Patterns Generated
B	{A:4}	<A:4>	{A,B:4}
C	{A,B:3}, {A:2}, {B:2}	<A:5, B:3>	{A,C:5}, {B,C:5}, {A,B,C:3}
D	{A,B,C:1}, {A,B:1}, {A:2}, {B:1}	<A:4>	{A,D:4}, {B,D:3}

Figure 3: Answer of Q4(d)

(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?

Based on the definition of Cosine we know:

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

from which we observe that there is no d (i.e., the number of null transactions) involved and it indicates that the measure is not null invariant. The only difference between $Lift(A, B)$ and $Cosine(A, B)$ is the order of the term $(a + b + c + d)$ which is 2 in $Lift(A, B)$ and is 1 in $Cosine(A, B)$. Hence, the term $(a+b+c+d)$ can be eliminated in $Cosine(A, B)$ to make it null invariant but cannot be eliminated in $Lift(A, B)$.