CS 412: Fall'21
Introduction To Data Mining

Take-Home Midterm

(Due Friday, October 15, 06:00 pm)

General Instructions

- You will have to answer the questions yourself, you cannot consult with other students in class. It is an open book exam, so you can use the textbook and the material shared in class, e.g., slides, lectures, etc.
- The take-home midterm will be due at 6 pm, Fri, October 15. We will be using gradescope for the midterm. Contact the TAs if you are having technical difficulties in submitting the assignment. We will NOT accept late submissions.
- Your answers should be typeset and submitted as a pdf. You cannot submit a hand-written and scanned version of your midterm. You can only do math equations, tables, and figures by hand, then scan or take a picture, and include that in your typeset answers.
- You DO NOT have to submit (python) code for any of the questions.
- For the questions, 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.
- If you have clarification questions, please use slack (preferred) or email all of us (instructor + 4 TAs). However, since the midterm needs to be submitted within 24 hours, please ask the question as early as possible (ideally, within 1-2 hours after the midterm is posted).

1. (18 points) This question considers summarization and visualization of probability distributions:

   (a) (3 points) Describe what a five-number summary of a distribution is.

   (b) (3 points) Describe what boxplots are and explain how boxplots incorporate the five-number summary.

   (c) (3 point) Can two different distributions have the exact same boxplot? If yes, briefly explain why and give an example. If no, briefly explain why not.

   (d) (3 points) Describe what quantile-quantile plots are.

   (e) (6 points) Consider the quantile-quantile (QQ) plot comparing prices of items sold in two stores, Store A and Store B, with Store A in the x-axis and Store-B in the y-axis. We assume that the price range of items sold in both the stores is the same, e.g., it is $[0, a]$ for some fixed $a > 0$, so the QQ-plot goes from $(0, 0)$ to (a, a).

      i. (3 points) Can the QQ plot stay entirely above the $y = x$ line except for the end-points (0,0) and (a,a)? Clearly explain your answer.

      ii. (3 points) Let m_A and m_B respectively denote the median price of items in Store A and Store B. Professor Median claims that if the point (m_A, m_B) lies below the $y = x$ line, then the average price of items in Store A is more than the average price of items in Store B. Do you agree with Professor Median? Clearly explain your answer.

(a) The five-number summary of a distribution is composed by five numbers, min, Q1 (25th percentile), median, Q3 (75th percentile), max.

(b) Boxplot is a technique to represent data with a box. Specifically, the Q1 and Q3 compose the two ends of the box, and the median is marked as a line within the box. Max and min are at the outside of the box and are connected by whiskers with the box.

(c) Yes they can. The boxplot describes the exactly same information as the five-number summary, which are min, Q1 (25th percentile), median, Q3 (75th percentile), max. All these five statistics are based on the order of all the data. In order to keep them fixed but change some of the data, we need to make sure the position of the above five numbers are not changed in all the data. As an example, we can select any number $a \in [Q3, max]$ and modify the number as $a + \Delta a$ but ensure that $a + \Delta a \in [Q3, max]$. Then the positions of the above five numbers keep unchanged and the boxplot is exactly the same as the one before modification.

(d) Quantile-quantile plot shows the quantiles of of a univariate distribution against the corresponding quantiles of another univariate distribution.

(e) i. Yes it can. Assume a data distribution where the prices for all the items in Store-B are $\{0, a, \ldots, a\}$ and the prices for all the items in Store-A are $\{0, \ldots, 0, a\}$. Hence, store-A and store-B share the same min and max but all the other quantities of store-B are larger than the corresponding quantities of store-A.

   ii. No. The comparison between medians from two sets of data cannot leads to the comparison between means. The claim from Professor Median is the same as 'the larger median leads to the larger mean'. We can easily come up with some counterexamples (assume $a = 10$) such as Store-A: $1, 1, 5, 5, 5$, Store-B:$4, 4, 4, 10, 10$. We get $median(A) > median(B)$ but $mean(A) < mean(B)$.

2. (22 points) Table 1 is a summary of customers' purchase history of diapers and beer. In particular, for a total of 1000 customers, the table shows how many bought both Beer and Diapers, how many bought Beer but not Diapers, and so on. For the problem, we will treat both 'Buy Beer' and 'Buy Diaper' as binary attributes.

	Buy Diaper	Not Buy Diaper
Buy Beer	100	400
Not Buy Beer	300	200

Table 1: Contingency table for Beer and Diaper sales.

(a) (2 points) Under the null hypothesis, i.e., 'Buy Beer' and 'Buy Diaper' are independent, what is the expected number for 'Buy Beer' and 'Buy Diaper'?

$400 * 500/1000 = 200$

(b) (2 points) Under the null hypothesis, i.e., 'Buy Beer' and 'Buy Diaper' are independent, what is the expected number for 'Buy Beer' and 'Not Buy Diaper'?

$500 * 600/1000 = 300$

(c) (3 points) What is the χ^2 statistic for the contingency table? Show steps of your calculation.

$\mathbb{E}[Notbuybeer\&NotbuyDiaper] = 600 * 500/1000 = 300$
$\mathbb{E}[BuyDiaper\&NotbuyBeer] = 400 * 500/1000 = 200$
$\chi^2 = \sum_{i=1}^{4} = \frac{(o_i - e_i)^2}{e_i} = 166.67$

(d) (3 points) At a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.

When 1 degree of freedom and $\alpha = 0.05$, the value needed to reject null hypothesis is 3.84. But the χ^2's statistic is larger than it. So, null hypothesis is rejected.

(e) (12 points) Consider an updated contingency table where the entry for 'Not Buy Beer' and 'Not Buy Diaper' is 20,000 instead of 200, and all other entries are the same. Compute the following measures for A='Buy Beer' and B='Buy Diaper' from the updated contingency table:

i. (3 points) $Lift(A, B)$. $\frac{s(A \cup B)}{s(A) \times s(B)} = \frac{100/20800}{500/20800 \times (400/20800)} = 10.4$

ii. (3 points) $Jaccard(A, B)$. $\frac{s(A \cup B)}{s(A) + s(B) - s(A \cup B)} = \frac{100}{100+400+300} = 0.125$

iii. (3 points) $Cosine(A, B)$.
$\frac{s(A \cup B)}{\sqrt{s(A) \times s(B)}} = \frac{100/20800}{\sqrt{500/20800 \times (400/20800)}} = 0.224$

iv. (3 points) $Kulczynski(A, B)$. $\frac{1}{2}(\frac{s(A \cup B)}{s(A)} + \frac{s(A \cup B)}{s(B)}) = \frac{1}{2}(100/20800/(500/20800) + 100/20800/(400/20800)) = 0.225$

3. (24 points) This question considers frequent pattern mining and association rule mining.

(a) (12 points) A transaction database (Table 2) has 5 transactions, and we will consider frequent pattern and association mining with (relative) minimum support $min\_sup = 0.6$ and (relative) minimum confidence $min\_conf = 0.6$.

Customer	Items Bought
C_1	{H, A, D, B, C}
C_2	{D, A, E, F}
C_3	{C, D, B, E}
C_4	{B, A, C, H, D}
C_5	{A, G, C}

Table 2: A transaction database.

i. (6 points) What is the frequent k-itemset for the largest k? Explain your answer. If there are more than one, it is sufficient to mention (and explain) only one.

We can use the Apriori algorithm to find the largest frequent itemset, where we follow the alphabetical ordering for the candidate generation, and relative support is 0.6 (i.e., absolute support is 3).

Customer	Items Bought
C_1	{A, B, C, D, H}
C_2	{A, D, E, F}
C_3	{B, C, D, E}
C_4	{A, B, C, D, H}
C_5	{A, C, G}

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

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

[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, C}: 3
- {A, D}: 3
- {B, C}: 3
- {B, D}: 3
- {C, D}: 3

[Step 4. Generate candidate C3]

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

[Step 5. Scan and filter to get frequent F3]

- {B, C, D}: 3

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

The largest $k = 3$, and the corresponding itemset is only {B, C, D}.

ii. (6 points) List all the association rules (with support and confidence) for the following type of rules:
$\forall x \in transaction, \quad buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3) \quad [s, c]$.
A rule is an association rule iff its support and confidence greater than or equal to the min_sup and min_conf, respectively.
[Step 1.] From the Q3(a)i. we know that only {B, C, D} has the qualified support. Therefore, $item_1$, $item_2$, and $item_3$ can only be chosen from {B, C, D}.
[Step 2.] Confidence of $X \Rightarrow Y = \sup(X, Y)/\sup(X)$, such that

- $buys(x, B) \wedge buys(x, C) \Rightarrow buys(x, D)$, [s=0.6, c=1]
- $buys(x, B) \wedge buys(x, D) \Rightarrow buys(x, C)$, [s=0.6, c=1]
- $buys(x, C) \wedge buys(x, D) \Rightarrow buys(x, B)$, [s=0.6, c=1]

(b) (12 points) A manager at a grocery store is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. Recall that constraint-based pattern mining can work with different types of constraints: pattern anti-monotonic, pattern monotonic, convertible, data anti-monotonic, or pattern/data succinct. For the following cases, state the *type of constraint* for *every constraint* in each case and discuss how to mine such patterns most efficiently.

(i) (6 points) The average price of all the items in each pattern is greater than \$50.
constraint $c_1 : avg(S.price) > 50$, types: convertible and then pattern (anti-)monotonic, and data anti-monotonic

To mine efficiently: 1)Start mining from c_1 by [Convertible], we first re-order the items in the transaction, according to the price-descending order; 2) Continue mining with c_1 by [Pattern Anti-Monotonic], we can apply the pattern anti-monotonicity, such that if a certain pattern violates the average condition (i.e., ≤ 50), we can stop growing this pattern to investigate its superset.

(ii) (6 points) The sum of the price of all the items with profit over \$5 in each pattern is at least \$200.
For a qualified transaction S, there are two constraints.
$c_1 : max(S.profit) > 5$, types: data succinct, data anti-monotonic, pattern monotonic
$c_2 : sum(S.price) \geq 200$, types: pattern monotonic, data anti-monotonic

To mine efficiently: 1) Start mining from c_1 by [Data Succinct], according to data succinctness, we start with only the item whose profit is over \$5 and remove transactions with low-profit items only (data anti-monotone is realized simultaneously); 2) Continue mining with c_2 by [Data Anti-Monotonic] remove unqualified transactions OR Continue mining with c_2 by [Pattern Monotonic], which means we can apply the pattern monotonicity, such that if a certain pattern meets the sum constraint, so does any of its supersets, we do not need to check it and its supersets again in the future.

4. (36 points) A sequence database SDB_1 (Table 3) has 5 transactions, and we will consider frequent sequential pattern mining with (absolute) minimum support of 2.

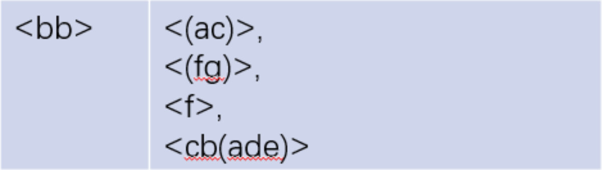
(a) (3 points) What are the length-1 frequent sequential patterns in SDB_1 and what are their supports?
a:3,b:5,c:4,d:3,e:3,f:2

Sequence_ID	Sequence
S_1	$\langle(bd)cb(ac)\rangle$
S_2	$\langle(bf)(ce)b(fg)\rangle$
S_3	$\langle(ah)(bf)abf\rangle$
S_4	$\langle(be)(ce)d\rangle$
S_5	$\langle a(bd)bcb(ade)\rangle$

Table 3: A sequence database SDB_1.

-0.5 pts for each missing or wrong support

(b) (5 points) What is the projected database for prefix $\langle bb\rangle$? Is $\langle bb\rangle$ a length-2 frequent pattern in SDB_1? What is the support of $\langle bb\rangle$?



<bb>	<(ac)>, <(fg)>, <f>, <cb(ade)>

Yes. Support is 4.

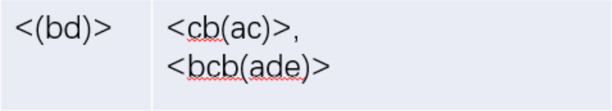
3 pts - correct projected db

2 pts - correct judgement and support

-1 pts for each missing or wrong element in projected database

-1 pts if there is wrong format of elements

(c) (5 points) What is the projected database for prefix $\langle(bd)\rangle$? Is $\langle(bd)\rangle$ a length-2 frequent pattern in SDB_1? What is the support of $\langle(bd)\rangle$?



<(bd)>	<cb(ac)>, <bcb(ade)>

Yes. Support is 2.

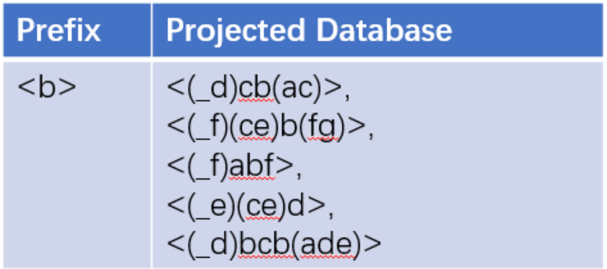
3 pts - correct projected db

2 pts - correct judgement and support

-2 pts for each missing or wrong element in projected database

-1 pts if there is wrong format of elements

(d) (4 points) What is the projected database for prefix $\langle b\rangle$?



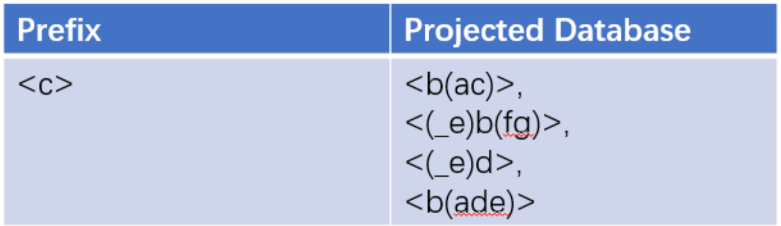
Prefix	Projected Database
<b>	<(_d)cb(ac)>, <(_f)(ce)b(fg)>, <(_f)abf>, <(_e)(ce)d>, <(_d)bcb(ade)>

-1 pts for each missing or wrong answer
-1 pts if there is wrong format of elements

(e) (8 points) Using the projected database for prefix $\langle b \rangle$, find all frequent subsequences starting with b. Please list your answer in lexicographically ascending order.
b, ba, bb, bc, bd, (bd), be, bf, (bf), bba, bbc, bbf, bca, bcb, bcd, b(ce), (bd)a, (bd)b, (bd)c, (bf)b, (bf)f, bcba, (bd)ba, (bd)bc, (bd)ca, (bd)cb, (bd)cba, (bf)bf

-0.5 pts for each missing or wrong answer

(f) (4 points) What is the projected database for prefix $\langle c \rangle$?



Prefix	Projected Database
<c>	<b(ac)>, <(_e)b(fg)>, <(_e)d>, <b(ade)>

-1 pts for each missing or wrong answer
-1 pts if there is wrong format of elements

(g) (7 points) Using the projected database for prefix $\langle c \rangle$, find all frequent subsequences starting with c. Please list your answer in lexicographically ascending order.
c, ca, cb, cd, (ce), cba

-1 pts for each missing or wrong answer