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 endpoints (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.

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'?

   (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'?

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

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

   (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) $Lift(A, B)$.

ii. (3 points) $Jaccard(A, B)$.

iii. (3 points) $Cosine(A, B)$.

iv. (3 points) $Kulczynski(A, B)$.

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.

   ii. (6 points) List all the association rules (with support and confidence) for the following type of rules:
   $\forall x \in transaction, \quad buys(x, item_1) \wedge buys(x, item_2) \Rightarrow buys(x, item_3) \quad [s, c]$.

   (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.

      (ii) (6 points) The sum of the price of all the items with profit over \$5 in each pattern is at least \$200.

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?

   (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 $\langle bb \rangle$?

   (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 $\langle (bd) \rangle$?

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.

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

(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.

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

(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.