# CS 412: Spring'21
# Introduction To Data Mining

# Take-Home Midterm

**(Due Tuesday, March 23, 10:00 am)**

**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 10 am, Tue, March 23. We will be using Compass for collecting the homework assignments. Please submit your answers via Compass (`http://compass2g.illinois.edu`). 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 DO NOT have to submit 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, you can use slack or campaswire. However, since the midterm needs to be submitted within 24 hours, please try to do your best in answering the questions based on your own understanding, in case responses are delayed.

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? Clearly explain your answer.

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

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

   (f) (3 point) How is a quantile-quantile plot different from a quantile plot? Clearly explain.

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) (3 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) (3 points) Under the null hypothesis, i.e., 'Buy Beer' and 'Buy Diaper' are independent, what is the expected number for 'Buy Beer' and 'Not But Diaper'?

   (c) (4 points) What is the $\chi^2$ statistic for the contingency table? Show steps of your calculation.

   (d) (4 points) At a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.

   (e) (4 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. What is the $\chi^2$ statistic for this updated contingency table? Show steps of your calculation.

   (f) (4 points) For the updated contingency table, at a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.

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 $min\_conf = 0.6$.

      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.

| 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.

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

(b) (12 points) A manager at a grocery store is interested in only the frequent patterns (i.e., itemsets) that satisfy certain constraints. For the following cases, state the *type of constraint*[1] 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 (Table 3) has 4 transactions, and we will consider frequent sequential pattern mining with (absolute) minimum support of 3. List all the frequent sub-

| Sequence_ID | Sequence |
|---|---|
| $S_1$ | $\langle a(abc)(ac)d(cf)\rangle$ |
| $S_2$ | $\langle (ad)c(bc)(ae)\rangle$ |
| $S_3$ | $\langle (ef)(ab)(df)cb\rangle$ |
| $S_4$ | $\langle eg(af)cbc\rangle$ |

Table 3: A sequence database.

sequences starting with the following prefixes and show details of your calculations:

(a) (12 points) Frequent subsequences starting with `a`.

(b) (6 points) Frequent subsequences starting with `c`.

(c) (6 points) Frequent subsequences starting with `d`.

(d) (3 points) Frequent subsequences starting with `b`.

(e) (3 points) Frequent subsequences starting with `e`.

(f) (3 points) Frequent subsequences starting with `f`.

[1]The type of constraints will be from our discussion on constraint-based pattern mining in class, e.g., Section 6.3 in the text book, and related in-class discussions.

(g) (3 points) Are there other frequent subsequences in the database not covered by the above? If your answer is 'yes', list one such frequent subsequence. If your answer is 'no', clearly explain why not.