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.
   (3 points) Minimum value, first quartile, median value, third quartile, maximum value. Get half of points if anyone is missing or wrong.

   (b) (3 points) Describe what boxplots are and explain how boxplots incorporate the five-number summary.
   (3 points) A boxplot is a graphic display of five-number summary. The ends of the box are the first and third quartiles of the data. The line within the box marks the median of the data. The two lines outside the box extends to the minimum and maximum values.

   (c) (3 point) Can two different distributions have the exact same boxplot? Clearly explain your answer.
   (3 points) Yes. Two different distributions can have the same five-number summary. For example, distribution A [1,2,3,4,5,6,7,8] and distribution B [1,2,3.5,4.5,4.5,6,7.5,8] have the same boxplot but different distributions.

   (d) (3 points) Describe what quantile plots are.
   (3 points) A quantile plot displays all of the data and plots quantile information. For a data x_i sorted in increasing order, f_i indicates that approximately 100 f_i % of the data are below or equal to the value x_i.

   (e) (3 points) Describe what quantile-quantile plots are.
   (3 points) A quantile-quantile plot graphs the quantiles of one univariate distribution against the corresponding quantiles of another.

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

   (3 points) Quantile-quantile plot compares the relative distribution of two data sets and the quantile plot only shows one data set distribution.

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'?

(2 points) Correct answer: 200.
(1 points) Show the calculation or explain the steps.

$$e_{ij} = \frac{count(A = a_i) * count(B = b_i)}{n}$$

(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'?
(2 points) Correct answer: 300.
(1 points) Show the calculation or explain the steps.

$$e_{ij} = \frac{count(A = a_i) * count(B = b_i)}{n}$$

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

(2 points) Correct answer for χ^2 statistic: 166.667
(2 points) Show the calculation or explain the steps.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

(d) (4 points) At a significance level of $\alpha = 0.05$, are these two variables 'Buy Beer' and 'Buy Diaper' independent? Explain your answer.
(4 points) No. The degree of freedom is (number of row - 1) * (number of columns -1) = (2-1) * (2-1) = 1. When df = 1 and $\alpha = 0.05$, p = 3.841. Because 3.841 < 166.667, we reject the hypothesis test. So these two variables are not independent.

(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 χ^2 statistic for this updated contingency table? Show steps of your calculation.
(1 points) Correct answer for expected value of each item: Table 2.
(1 points) Show the calculation or explain the steps.

$$e_{ij} = \frac{count(A = a_i) * count(B = b_i)}{n}$$

(1 points) Correct answer for χ^2 statistic: 887
(1 points) Show the calculation or explain the steps.

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

	Buy Diaper	Do Not Buy Diaper
Buy Beer	9.62	490.38
Do Not Buy Beer	390.38	19909.62

Table 2: Expected value for each item.

(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.
(4 points) No. The degree of freedom is (number of row - 1) * (number of columns -1) = (2-1) * (2-1) = 1. When df = 1 and $\alpha = 0.05$, p = 3.841. Because 3.841 < 887, we reject the hypothesis test. So these two variables are not independent.

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

(a) (12 points) A transaction database (Table 3) 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 3: 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.
k=3 (2 points), {B,C,D} (2 points). Any kind of intermediate steps like Apriori or Fp-Growth. (2 points)

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

A. $buys(x, B) \wedge buys(x, C) \Rightarrow buys(x, D)$. (2 points)
B. $buys(x, C) \wedge buys(x, D) \Rightarrow buys(x, B)$. (2 points)
C. $buys(x, B) \wedge buys(x, D) \Rightarrow buys(x, C)$. (2 points)

(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.

[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.

(i) (6 points) The average price of all the items in each pattern is greater than \$50. Convertible constraints (3 points). Using Fp-growth on descending ordered transactions. (3 points)

(ii) (6 points) The sum of the price of all the items with profit over \$5 in each pattern is at least \$200. succinct (2 points), anti-monotone (2 points), remove transactions with profit less than \$5. As the size of frequent itemset grows, we don't need to check the sum of the price if its sub-pattern already satisfies this condition. (2 points)

4. (36 points) A sequence database (Table 4) has 4 transactions, and we will consider frequent sequential pattern mining with (absolute) minimum support of 3. List all the frequent subsequences starting with the following prefixes and show details of your calculations:

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 4: A sequence database.

Each subsequence is worth a uniform number of points (e.g. for (a) they are worth 2.4). Half this number of points will be subtracted for incorrect subsequences until 0 is reached. Without explanation you will get −20% on a through f.

(a) (12 points) Frequent subsequences starting with a. <a>, <ab>, <ac>, <acb>, <acc>

(b) (6 points) Frequent subsequences starting with c. <c>, <cc>, <cb>

(c) (6 points) Frequent subsequences starting with d. <d>, <dc>

(d) (3 points) Frequent subsequences starting with b. <b>, <bc>

(e) (3 points) Frequent subsequences starting with e. <e>

(f) (3 points) Frequent subsequences starting with f. <f>

(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.

No. Only subsequences starting with <g> would not be covered. However, g only has a support of 1, so any sequence containing it cannot be frequent.

(2 points) You have the right idea but your explanation doesn't explain why g can't work.