CS 2756, Spring 2022
Homework 4
Due: April 6, 2022, 11:59 PM, Gradescope

Instructions:

- The main submission of HW 4 must be uploaded on Gradescope. Your submission will be graded via Gradescope platform.
- Your PDF file should be named as "lastname_firstname_hw4". For example, if your name is Jane Doe, name your file as "doe_jane_hw4".
- Please write your name (First Name, Last Name) legibly, as it appears on Canvas. Please also include your PITT email.
- Type out your solutions on a separate blank file (using Word, Latex, etc.), and don't forget to convert the file to PDF extension before submitting. You should not type out on the original pdf file.
- Only a subset of the questions will be graded. However, such questions are not determined a priori, therefore please do your best to answer all the questions correctly.

Question 1
Consider the market basket transactions shown in Table 1.

Table 1: Market basket transactions.

Transaction ID	Item Bought
1	{Beer, Diapers}
2	{Milk, Diapers, Bread, Butter}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Milk, Beer, Diapers, Eggs}
6	{Beer, Cookies, Diapers}
7	{Milk, Diapers, Bread, Butter}
8	{Bread, Butter, Diapers}
9	{Bread, Butter, Milk}
10	{Beer, Butter, Cookies}

(a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

(b) What is the maximum size of frequent itemsets that can be extracted (assuming minsup >0)?

(c) Given that this dataset has 7 distinct items, write an expression for the maximum number of size-3 itemsets that can be derived.

(d) What is the support count for {Bread}, {Milk}, {Bread, Milk}?

e) What is the confidence of the rule {Bread} -> {Milk} and {Milk}->{Bread}?

f) In general, rules {a} –> {b} and {b} –> {a}can have the different confidence.
If you know that the rules {a} –> {b} and {b} –> {a} have the same confidence, what can you say about the relation between support of {a} and support of {b}?

Question 2

Suppose {A,B,D,E} is a frequent itemset and {B,C,D} is NOT a frequent itemset. Given this information, we can be sure that certain other itemsets are frequent and sure that certain itemsets are NOT frequent. Other itemsets may be either frequent or not.

Which of the following statements are correct?

Give a one sentence explanation if you believe any statement is incorrect.

a) {A, E} can be either frequent or not frequent.

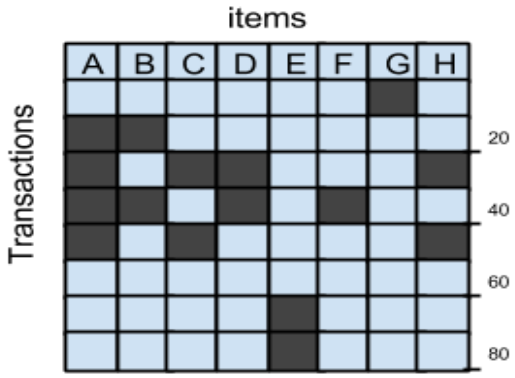
b) {A,B,D,E,F} can be either frequent or not frequent.

c) {A,B,C,D,E,F} can be either frequent or not frequent.

Question 3

The dataset below contains 8 items and 80 transactions. Dark-colored cells indicate the presence of items, and lightly shaded cells indicate the absence of items. We apply the Apriori algorithm to extract frequent itemsets with minsup = 25% (i.e., itemsets must contain at least 20 transactions).

Indicate whether the following statements are "true" or "false." Give an explanation if your answer is "false." An example of two statements with their corresponding answers is shown below.



Statement	True/False. Explanation if false
{A} is a closed itemset	True
{A} is a maximal itemset	True/False: False, {A, B} is frequent.

Statements:

a) {A} is a closed frequent itemset

b) {A,C} is a closed itemset

c) {A,B,D} is a frequent itemset

d) {A,D} is a maximal itemset

Question 4

Itemsets	Support
{X}	20
{X,Y}	16
{X,Z}	14
{X,W}	14
{X,Y,W}	14
{X,Y,Z}	10

Consider a market-basket transaction data set that has four items X, Y, Z, and W. The table on the right shows the support of some itemsets, while the support of other itemsets is unknown. Label each itemset listed below with the following letter(s):

• C if it is a closed itemset,

• N if it is not a closed itemset, and

• I if the information is not enough to judge whether or not it is closed.

a. {X}
b. {Y}
c. {X, Y}
d. {X, W}
e. {X, Z}

Question 5

Consider a market basket data set that has only six items (p,q,r,s,t,w). The following is a set of all frequent 3-itemsets in this data:

{p, q, r}, {p, q, s}, {p, r, s}, {p, r, t}, {p, s, t}, {p, s, w}, {q, s, t}, {r, s, t}.

a. List all candidate 4-itemsets obtained by a candidate generation procedure using the Fk-1 x F1 merging strategy.

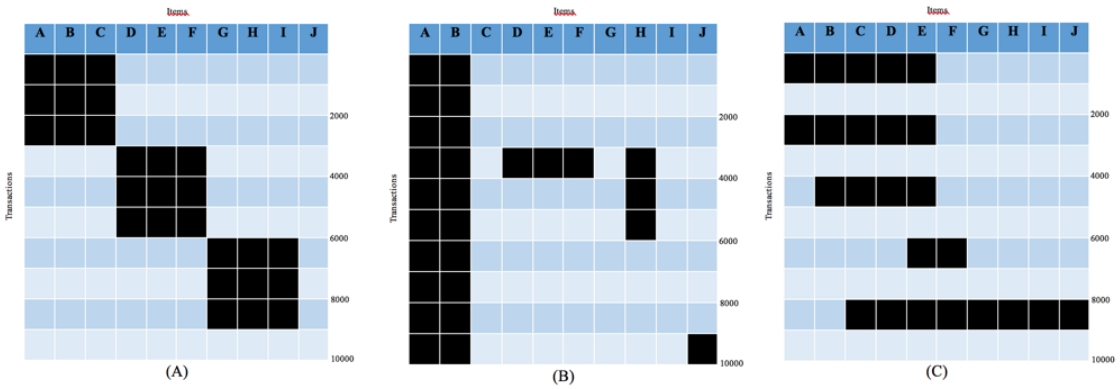
b. List all candidate 4-itemsets obtained by the candidate generation procedure in Apriori (Fk-1x Fk-1).

c. List all candidate 4-itemsets that survive the candidate pruning step of the Apriori algorithm.

d. Based on the list of frequent 3-itemsets given above, is it possible to generate a frequent 5-itemset? State your reason clearly.

Question 6

Each dataset below contains 10 items and 10000 transactions. Dark cells indicate the presence of items and white (and grey) cells indicate the absence of items. We will apply the Apriori algorithm to extract frequent item sets with minsup = 20% (i.e. Itemsets must contain at least 2000 transactions).



(a) Which dataset will produce the largest number of frequent itemsets?
Answer (A/B/C):

(b) Which dataset will produce the longest frequent itemset?
Answer (A/B/C):

(c) Which dataset will produce frequent itemsets with highest support?
Answer (A/B/C):

(d) Which dataset will produce the least number of maximal frequent itemsets?
Answer (A/B/C):

(e) Which dataset will produce the least number of closed itemsets?
Answer (A/B/C):

Question 7

Consider a binary data set representing the words that are contained in a document. The rows are the documents, the columns are the words, and the entry corresponding to the ith row (document) and jth column (word) is a 1 if the word is present in the document and 0 if it is not.

a) Choose **yes** or **no** to indicate which of the following properties an interestingness measure should possess to be useful for evaluating whether a pair of words in this data set are strongly related, i.e., if the two words co-occur in many documents.

	word1	**word2**	**word3**	**...**
document1	1	1	0	...
document2	0	0	1	...
document3	1	0	1	...
...	...	...	...	...

i. Invariant under scaling: Yes No

ii. Invariant under inversion: Yes No

iii. Invariant under null addition: Yes No

b. Based on your answers above, would you prefer the cosine measure or correlation for this task? Briefly justify your answer.

Question 8

Consider the data set shown in Table 1. Suppose we are interested in extracting the following association rule:

$\{\alpha_1 <= \text{Age} < \alpha_2, \text{Play Chess} = \text{Yes}\}$ *implies* $\{\text{Enjoy Solving Puzzles} = \text{Yes}\}$

Table 1: Data set

Age	Plays Chess	Enjoys Solving Puzzles
10	Yes	Yes
12	Yes	Yes
15	Yes	No
18	Yes	No
20	Yes	Yes
22	No	No
26	No	No
28	Yes	Yes
36	No	No
39	No	Yes
41	No	No
47	No	Yes

To handle the continuous attribute, we apply the equal-frequency approach with 3, 4, and 6 intervals. Categorical attributes are handled by introducing as many new asymmetric binary attributes as the number of categorical values. Assume that the support threshold is 10% and the confidence threshold is 70%.

(a) Suppose we discretize the Age attribute into 3 equal-frequency intervals. Find a pair of values for $\alpha 1$ and $\alpha 2$ that satisfy the minimum support and minimum confidence requirements.

(b) Repeat part (a) by discretizing the Age attribute into 4 equal-frequency intervals.

(c) Repeat part (a) by discretizing the Age attribute into 6 equal-frequency intervals.

(d) From the results in part (a), (b), and (c), discuss how the choice of discretization intervals will affect the rules extracted by association rule mining algorithms.

Question 9

Consider the following frequent 3-sequences:

<{r}, {s, t}>

<{p}, {s, t}>

<{p}, {r}, {s}>

<{p}, {r}, {t}>

<{p},{r,t}>

<{s, t}, {w}>

<{p,r,s}>

<{r,s,t}>

a) Generate all the candidate 4-sequences from the given frequent 3-sequences, using the method for candidate generation described in the book. For every 4-sequence generated, also write down the corresponding 3-sequences that were merged to generate the 4-sequence.

b) Find out the candidate 4-sequence that would survive the candidate pruning.

Question 10

Assume that the following seven 3-sequences are the only frequent 3-sequences generated by the Apriori algorithm in a sequence data set:
<{A} {C} {D}>
<{B} {C} {D}>
<{A} {C} {E}>
<{A} {D, E}>
< {A, B} {E}>
<{A} {C, D}>
<{C} {D, E}>

(a) Is it possible to merge <{A, B} {E}> and <{A} {C} {E}>, and generate a 4-sequence? If yes, write down the sequence. **If no, explain briefly**.

(b) <{B} {C} {D, E}> is one of the candidate 4-sequence generated by merging <{B} {C} {D}> and <{C} {D, E}>. Will this be pruned in the candidate pruning step? **Explain briefly**.