The Hong Kong Polytechnic University
Department of Computing
COMP 578 Data Warehousing and Data Mining
MScECT, Semester 1, 03-04
Mid-term Quiz
Answer ALL questions
Time Allowed: 60 minutes
(Aids: A standard calculator)

1. The company management of an internet book store decided to record all user sessions on their server so that they can try to mine for interesting patterns in the data. Assume that each such session consists of an ordered list of pages accessed by a user as shown in Table 1. Based on the data in the table, the book store would like to know if they can predict whether or not a person will buy computer books, business books or "does not buy any book" based on the "date of week", the "length of session" and the "number of pages visited in a session". Use the ID3 to complete the construction of the following decision tree and try to find the answer for the bookstore.

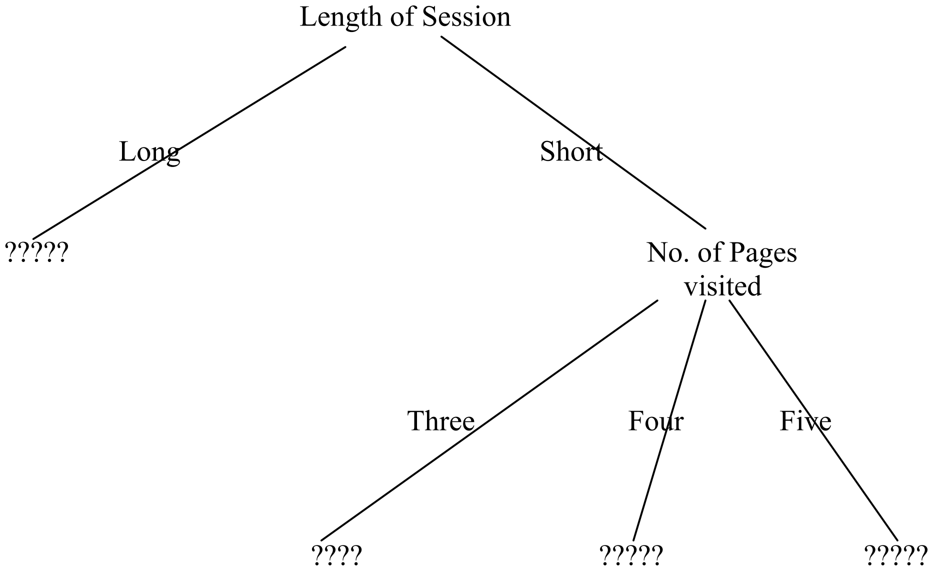


Table: Pages visited during a user session

Date of week	**Session number**	**Length of session**	**Sequence of pages visited (s_i)**	**Types of books bought**
Monday	0013	Long	B→C→C	Did not buy any book
Monday	0024	Short	A→B→A→E→F	Bought computer books
Monday	0035	Short	B→C→D	Bought business books
Monday	0014	Short	B→C→B→D→C	Bought business books
Monday	0085	Long	A→B→C	Did not buy any book
Tuesday	0099	Long	A→D→E	Bought computer books
Tuesday	0102	Long	A→B→C	Did not buy any book
Tuesday	1011	Long	A→B→A→B	Bought business books
Tuesday	1339	Short	B→D→F→D	Did not buy any book
Tuesday	2021	Long	A→C→D→A	Bought business books

Given the testing data set below, what is the classification accuracy of your decision tree?

Date of week	Session number	Length of session	Sequence of pages visited (s_i)	Types of books bought
Monday	0019	Short	B→C→C	Bought computer books
Tuesday	0034	Long	A→B→E→F	Did not buy any book
Monday	0055	Short	B→C→D	Bought business books
Tuesday	0004	Short	B→C→B→C→F	Bought business books
Monday	004	Long	A→B→C→B	Did not buy any book

2. A friend of yours is a manager of a large international bank. He would like to discover what criteria their loan officers have been using when approving loans. The data consists of a number of customer records each of which is characterized by 5 attributes: Age (with domain in the interval of [0, 100]), Income (with domain in the interval of [50K, 100K]), Marital Status (with domain={Married, Single}), Account Balance (with domain={High, Medium, Low}), and Loan Size (with domain=[100K, 200K].

Record No.	Age	Income	Marital Status	Account Balance	Loan Size
1	56	92K	M	High	160K
2	47	88K	M	Low	139K
3	32	90K	S	Low	199K
4	26	72K	S	High	178K
5	66	66K	M	Low	125K

a) Consider only the attributes of Age, Income and Loan Size, group the above record into two clusters using the k-means algorithm and using record No. 1 and 2 to be their respective initial centers. Based on your results, is there any evidence to support your friend's belief that older clients that earn more income are given relatively larger loans by the loan officers?

b) After interviewing the Loan Officers, you believe that patterns underlying the data set can become more obvious when the quantitative variables of Age and Income are transformed into qualitative variables according to the following rules:

- If Age=[0, 40] then Age=Young.
- If Age=[36, 55] then Age=Middle.
- If Age=[56, 100] then Age=Old.
- If Income=[50K, 75K] then Income=Low.
- If Income=[75K, 100K] then Income=High.
- If Loan Size=[100, 140] then Loan Size=Small.
- If Loan Size=[140, 170] then Loan Size=Medium
- If Loan Size=[170, 200] then Loan Size=Large.

Given these transformed discrete qualitative attributes and the Marital Status and Account Balance, use the Condorset to determine if there is any interesting inherent grouping in the data. Show you steps and discuss your results.

3. Suppose that you are given a set of transaction data by a supermarket as shown below.

a) By setting the minimum support to 25% and minimum confidence to 40%, use the Apriori algorithm to find all 3-item interesting association rules in the data set.

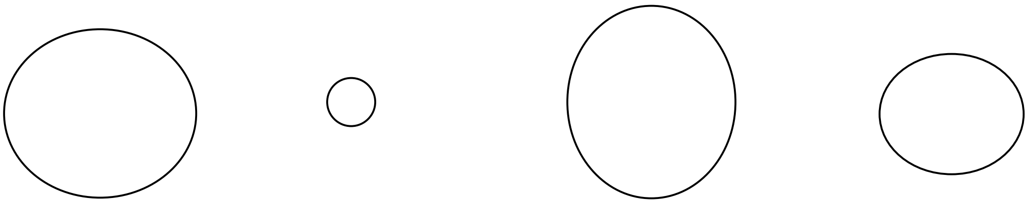
b) Assume that a user sets the lift ratio to 1.75, which rules did you discover for a) above are still interesting?

c) Please explain if the choice of minimum support and minimum confidence is suitable. If not, what do you suggest should be used.

Table

Customer	Items
1	Orange, Coke, apple, diapers
2	Pepsi, diapers, lemon
3	Pepsi, apple, orange, Coke
4	Apple, orange, lemon
5	Diapers, apple, Coke
6	Orange, Pepsi, lemon
7	Coke, orange, diapers, apple
8	Lemon, apple, Coke, orange, diapers

4. A clothing database maintained by a fashion outlet contains data that describe the different kinds of clothing it sells. Among the attributes in the database, there are those related to "patterns". Many of these patterns are inherently fuzzy and are best described with the concepts of fuzzy sets. For example, each of the following figures can be considered a "circular" pattern. With the fuzzy set notations introduced in our classes, develop a membership function for a "circle". (Examples of patterns that are considered "circle" are given below.)



***** END *****

***** PLEASE RETURN THIS QUESTION PAPER TOGETHER WITH YOUR ANSWER SHEETS *****