CSCI 4502/5502: Data Mining

Midterm Exam (Fall 2018)

Student Name:

Email Address:

Honor Code Pledge: On my honor as a University of Colorado Boulder student I have neither given nor received unauthorized assistance.

_____ _____

Student Signature *Date(mm/dd/yyyy)*

_____

Row Assignment

Instructions

1. This is a closed-book exam.
2. No calculator is allowed, write out your computation steps without doing the final calculation.
3. No smartphone, no computer is allowed. Turn off your cellphone or mute it.
4. Sit in the row as specified in the "Row Assignment" above.
5. Write down your name, email address, and sign the Honor Code Pledge.
6. The total exam time is 75 minutes, from 9:30am to 10:45am.
7. If you think there is ambiguity in a question, state your assumptions and answer accordingly.
8. Do not start the exam until being told by the proctor.

1. Determine if the following statements are true or false. Briefly explain why.

   (a) $MIN(S.price) \leq v$ is a monotonic constraint, where $MIN(S.price)$ is the minimum item price in itemset S.

   (b) If A and B are positively correlated, it implies that B is caused by A.

   (c) Outliers are caused by errors and should be discarded.

   (d) The decision tree classification method cannot support incremental data.

2. Provide a brief answer for each of the following questions.

   (a) Data mining aims to find interesting patterns. Identify three properties that an interesting pattern should have.

   (b) Identity three differences between OLTP and OLAP.

   (c) Why cannot the k-means clustering algorithm find clusters of arbitrary shape?

   (d) When using classification to detect cancer, what is the difference between false positive and false negative?

3. Consider the 2×2 contingency table summarizing a student population with respect to playing chess and riding bike.

   (a) Compute the $lift$ value. $lift(A, B) = \frac{P(A \cup B)}{P(A)P(B)}$.

   (b) Compute the χ^2 value. $\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$ and $e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{N}$.

   (c) Are both correlation measures null-invariant? Briefly explain why.

	chess	$\neg$ chess
bike	500	1200
$\neg$ bike	800	2500

4. Consider the market basket transactions shown in the table.

   (a) Let *min_support = 50%*, find all frequent itemsets using the Apriori algorithm.

   (b) Draw the corresponding FP-tree for this data set. (**Note: This task is required for CSCI 5502 students, and 5-point extra credit for CSCI 4502 students.**)

TID	Items
1	A, F, I, N, S
2	C, F, N, S, R
3	A, N, S, O
4	I, N, S, R
5	C, F, I, N, S, R
6	C, N, O
7	F, I, N, S, R
8	A, F, S
9	C, N, S, R
10	A, F, S, O, R

5. Consider the training examples shown in the table for a binary classification problem of the class label "Play Golf". **Information Gain**: $p_i = |C_{i,D}|/|D|, Info(D) = -\sum_{i=1}^{m} p_i log_2(p_i)$, and $Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j)$ **Bayes' Theorem**: $P(H|X) = P(X|H)P(H)/P(X)$.

   (a) Write out the steps for computing the information gain of the "Outlook" attribute. Given the information gain values of two attributes, why do we select the attribute with higher information gain for decision tree classification?

   (b) Using naive Bayes classifier, write out the steps for classifying "Play Golf" as Yes or No for $X = (Outlook = Sunny, Temp = Hot, Humidity = normal, Windy = False)$.

ID	Outlook	Temp	Humidity	Windy	Play Golf
1	Rainy	Hot	High	False	No
2	Rainy	Hot	High	True	No
3	Overcast	Hot	High	False	Yes
4	Sunny	Mild	High	False	Yes
5	Sunny	Cool	Normal	False	Yes
6	Sunny	Cool	Normal	True	No
7	Overcast	Cool	Normal	True	Yes
8	Rainy	Mild	High	False	No
9	Rainy	Cool	Normal	False	Yes
10	Sunny	Mild	Normal	False	Yes
11	Rainy	Mild	Normal	True	Yes
12	Overcast	Mild	High	True	Yes
13	Overcast	Hot	Normal	False	Yes
14	Sunny	Mild	High	True	No

6. Consider the following set of one-dimensional points: {70, 150, 180, 40, 80, 210, 30, 60}.

   (a) Show the first round of *k-means* clustering method to generate two clusters, assuming the initial centroids are 50 and 120, respectively.

   (b) Identify a different pair of initial centroids that results in two clusters that are different from that of (a)

   (c) What is the computation complexity of k-means clustering? Briefly explain why. (**Note: This task is required for CSCI 5502 students, and 5-point extra credit for CSCI 4502 students.**)

7. (**Note: This task is optional and 5-point extra credit for all students.**) Consider the hash tree for candidate 3-itemsets shown in Figure 6.32.

   (a) Given a transaction that contains items {2, 3, 5, 6, 8}, which of the hash tree leaf nodes will be visited when finding the candidate 3-itemsets contained in the transaction?

   (b) What are the three bottlenecks of the Apriori algorithm? Which bottleneck is addressed by the hash tree method?

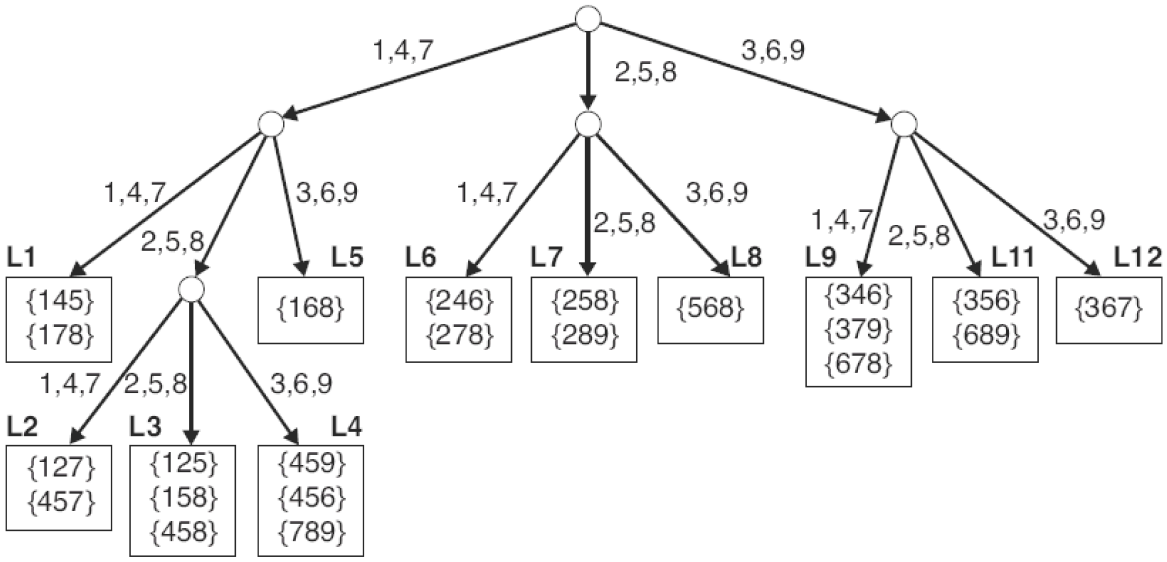

Figure 6.32. An example of a hash tree structure.