CSCI 4502/5502: Data Mining

Midterm Exam (Fall 2019)

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. Sit in the row as specified in the "Row Assignment" above.
2. Write down your name, email address, and sign the Honor Code Pledge.
3. The total exam time is 75 minutes, from 9:30am to 10:45am, or 11:00am to 12:15pm.
4. This is a closed-book exam.
5. No calculator is allowed, write out your computation steps without doing the final calculation.
6. No smartphone, no computer is allowed. Turn off your cellphone or mute it.
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) Given two association rules $A \Rightarrow B$ and $B \Rightarrow A$, they always have the same confidence value.

   (b) $MAX(S.price) \geq v$ is an anti-monotonic constraint, where $MAX(S.price)$ is the maximum item price in itemset S.

   (c) The DBSCAN clustering method can only identify clusters of convex shapes.

   (d) Decision tree induction is a supervised-learning method.

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

   (a) What are the four different views of data mining?

   (b) What information is shown in a box plot?

   (c) What is the key difference between k-means clustering and Expectation Maximization?

   (d) How are contextual outlier different from collective outliers?

3. Consider the 2×2 contingency table summarizing a consumer population with respect to buying books and buying a laptop.

   (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) Let v be the computed $lift$ value and x be the computed χ^2 value, how do you determine whether buying books is correlated with buying a laptop, positively or negatively?

	books	$\neg$ books
laptop	2000	1000
$\neg$ laptop	3000	4000

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

   (a) Let *min_support = 40%*, 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	B, H, M, Z
2	B, K , R
3	B, H, R, X
4	M, Z
5	B, H, R, X
6	H, K, P
7	B, H, K, R, X
8	H, M
9	B, H, R, X
10	P, H, X
11	H, Z
12	B, K, R, X

5. Consider the training examples shown in the table for a binary classification problem of the class label "Windy". **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 "Temp" attribute. Given the information gain values of two attributes, do we select the attribute with higher information gain or lower information gain for decision tree classification?

   (b) Using naive Bayes classifier, write out the steps for classifying "Windy" as True or False for $X = (Outlook = Overcast, Temp = Mild, Humidity = High, PlayGolf = No)$.

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: {35, 69, 9, 78, 9, 23, 81, 57, 15, 48}.

   (a) Show the first round of *k-means* clustering method to generate two clusters, assuming the initial centroids are 30 and 60, respectively.

   (b) Given two different sets of initial centroids, does the k-means method always generate the same clustering results when it terminates?

   (c) Show the first four rounds of agglomerative hierarchical clustering for the list of data points above. (**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 {1, 4, 5, 8, 9}, which of the hash tree leaf nodes will be visited when finding the candidate 3-itemsets contained in the transaction?

   (b) Is the hash tree method more efficient than the Apriori algorithm? Explain why.

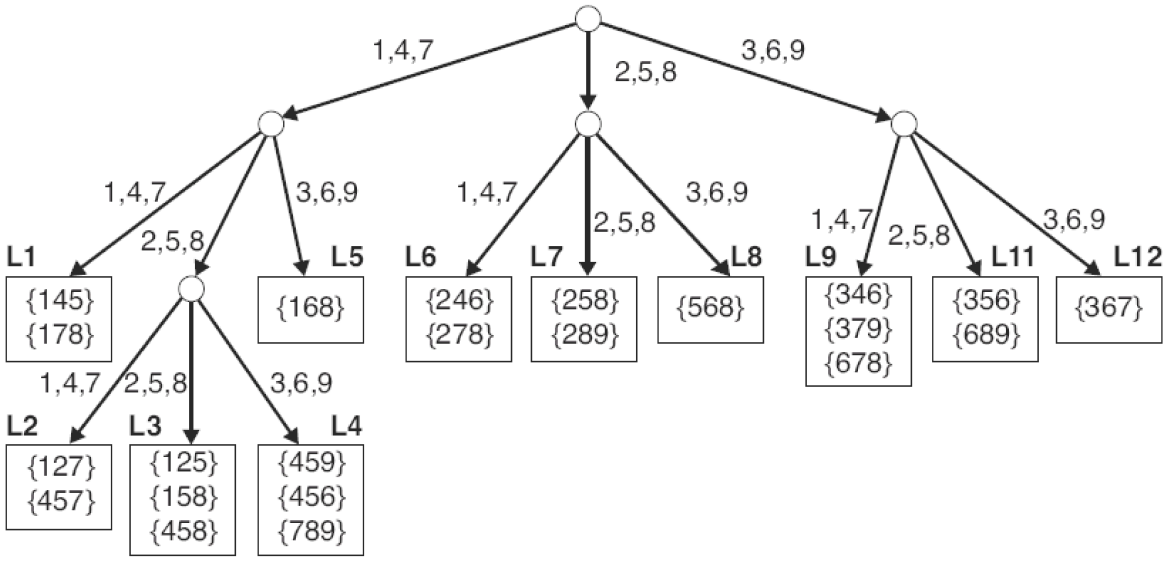

Figure 6.32. An example of a hash tree structure.