CSCI 4502/5502: Data Mining
Homework 4

Due at **12:30pm on Thursday, Mar 12, 2015**. Submit one file electronically at moodle: "**LastName_FirstName_Homework4.pdf**". Make sure to include your name, student id, and the Honor Code Pledge (`http://honorcode.colorado.edu/student-information/honor-code-pledge`).

1. The Apriori algorithm uses a hash tree data structure to efficiently count the support of candidate itemsets. Consider the hash tree for candidate 3-itemsets shown in Figure 6.32.

   (a) How many candidate 3-itemsets are there in total? (i.e., without using hash tree, how many candidate 3-itemsets do we need to check for each transaction?)

   (b) Given a transaction that contains items {1, 3, 4, 6, 8}, which of the hash tree leaf nodes will be visited when finding the candidate 3-itemsets contained in the transaction?

   (c) Use the visited leaf nodes in part (b) to determine the candidate itemsets that are contained in the transaction {1, 3, 4, 6, 8}.

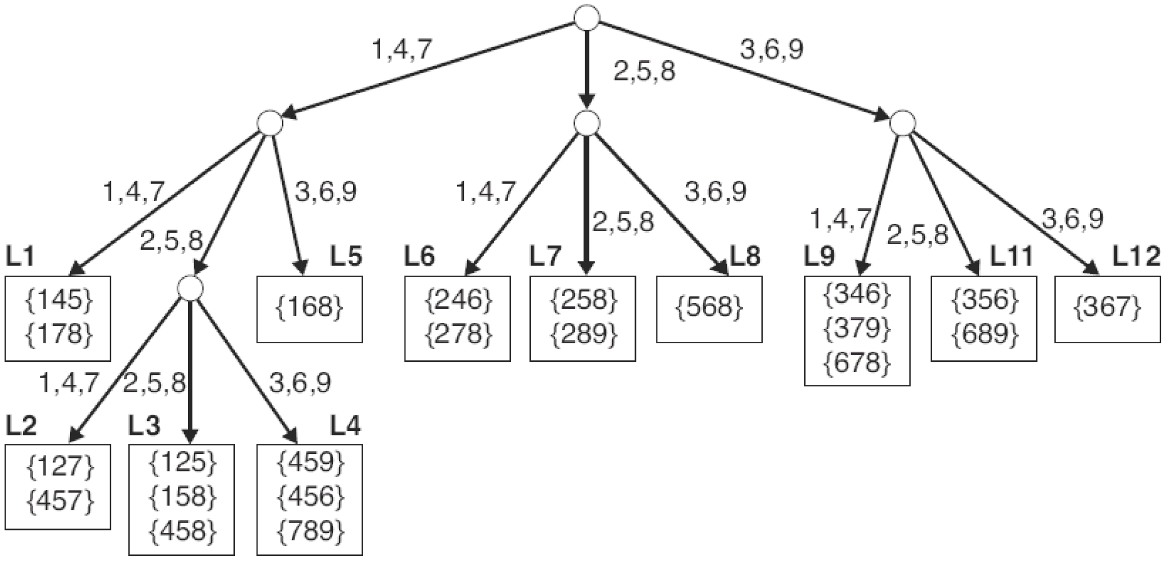

Figure 6.32. An example of a hash tree structure.

2. For each of the following statements, determine if it is true or false and briefly explain why. **Note: Task (c) is required for CSCI 5502 students and 5-point extra credit for CSCI 4502 students.**

   (a) Rule $v \notin S$ is monotonic but not antimonotonic.

   (b) Rule $V \subset S$ is monotonic but not antimonotonic.

   (c) Rule $avg(S) \geq v$ can be converted into a monotonic rule.

3. Consider the traffic accident data set shown in the following table.

Weather Condition	*Driver's Condition*	*Traffic Violation*	*Seat Belt*	*Crash Severity*
Bad	Sober	None	Yes	Minor
Good	Sober	Disobey stop sign	Yes	Minor
Bad	Sober	Disobey stop sign	Yes	Minor
Good	Alcohol-impaired	Exceed speed limit	No	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Alcohol-impaired	Disobey stop sign	Yes	Minor
Good	Alcohol-impaired	Exceed speed limit	Yes	Major
Bad	Alcohol-impaired	None	Yes	Major
Good	Sober	Disobey traffic signal	Yes	Major
Good	Alcohol-impaired	Non	No	Major
Bad	Sober	Disobey traffic signal	No	Major
Good	Sober	Exceed speed limit	Yes	Major

Let *Seat Belt* be the class label.

(a) Using information gain as the attribute selection measure, construct the first level of the decision tree.

(b) If gain ratio is used as the attribute selection measure, will the first level of the decision tree be different from above? **Note: This task is required for CSCI 5502 students and 10-point extra credit for CSCI 4502 students.**

(c) Given a traffic accident with the values "Bad", "Sober", "None", and "Major" for the attributes *Weather Condition*, *Driver's Condition*, *Traffic Violation*, and *Crash Severity*, respectively, how would a naïve Bayesian classifier determine whether *Seat Belt* was used or not?