CSCI 5502 - Data Mining - Homework 4
Name: Krishna Chaitanya Sripada
Student ID: 104375417
Honor Pledge: On my honor, as a University of Colorado at Boulder student, I have neither given nor received unauthorized assistance on this work.

Ans 1

(a) The number of candidate 3-itemsets in total are 22 i.e., L1 has 2, L2 has 2, L3 has 3, L4 has 3, L5 has 1, L6 has 2, L7 has 2, L8 has 1, L9 has 3, L11 has 2 and L12 has 1. So a total of 22 candidate 3-itemsets.

Without using hash tree, the number of candidate 3-itemsets we need to check for each transaction is $\binom{k}{3}$ where 'k' is the number of items. Here k = {1,2,3,4,5,6,7,8,9}. Therefore, the number of candidate 3-itemsets = $\binom{9}{3} = 84$.

(b) The list of candidate 3-itemsets for the items {1,3,4,6,8} are {1,3,4}, {1,3,6}, {1,3,8}, {1,4,6}, {1,4,8}, {1,6,8}, {3,4,6}, {3,4,8}, {3,6,8}, {4,6,8}. The hash tree leaf nodes that need to visited are : L1, L5, L9 and L12 i.e, for transactions with prefix {1,3}, the L5 node is visited. For transactions with prefix {1,4}, the L1 node is visited. The transactions with prefix {1,6}, the L5 node is visited. For transactions with prefix {3,4}, the L9 node is visited. For transactions with prefix {3,6}, the L12 node is visited and finally for transactions with prefix {4,6}, the L5 node is visited.

(c) The candidate item sets contained in the transaction {1,3,4,6,8} are {1,6,8} and {3,4,6}.

Ans 2

(a) Rule $v \notin S$ is monotonic but not antimonotonic.
Antimonotonic Example: Let v = 4 and S = {1,2,4}. Now itemset S violates the rule $v \notin S$. Let $S^{'}$ be the superset= {1,2,4,5}. Here the superset $S^{'}$ also violates the rule $v \notin S$. Thus the rule is antimonotonic.

Not Monotonic Example: Let v = 5 and S = {1,2,4}. Now itemset S satisfies the rule $v \notin S$. Let $S^{'}$ be the superset= {1,2,4,5}. Here the superset $S^{'}$ does not satisfy the rule $v \notin S$. Thus the rule is not monotonic.

Therefore, the rule $v \notin S$ is monotonic but not antimonotonic is **false**.

(b) Rule $V \subset S$ is monotonic but not antimonotonic.
Monotonic Example: Let V = {2,3} and S = {2,3,6,7}. Now itemset S satisfies the rule $V \subset S$. Let $S^{'}$ be the superset = {2,3,6,7,8,9} also satisfies the rule. Thus the rule is monotonic.

Not Antimonotonic Example: Let V = {2,3} and S = {4,5,6,7}. Now itemset S violates the rule $V \subset S$. Let $S^{'}$ be the superset = {4,5,6,7,2,3}. Here the superset $S^{'}$ does not violate the rule. Thus the rule is not antimonotonic.

Therefore, the rule $V \subset S$ is monotonic but not antimonotonic is **true**.

(c) Rule $avg(S) \geq v$ can be converted into a monotonic rule.
Monotonic Example: Let v = 25 and S = {25,50,75,100,125,150} where the items in the itemset S are ordered in an ascending order, the $avg(S) = 87.5$ satisfies the rule. Let $S^{'}$ be the superset = {25,50,75,100,125,150,175,200} where the items in the itemset $S^{'}$ are also ordered in an ascending order, the $avg(S) = 112.5$ also satisfies the rule.

Therefore, the rule $avg(S) \geq v$ can be converted into a monotonic rule is **true**.

Ans 3

(a) Given "Seat Belt" is the class label. Now using information gain as the attribute selection measure, we calculate the following:

Expected information needed to classify a tuple in our dataset is,

$$Info(D) = -\frac{8}{12}\log_2(\frac{8}{12}) - \frac{4}{12}\log_2(\frac{4}{12})$$
$$= 0.39 + 0.528 = 0.918 \text{ bits}$$

Now, we need to compute the expected information requirement for each attribute.

(i) The first attribute is "Weather Condition". The expected information needed to classify a tuple in the dataset if the tuples are partitioned according to Weather Condition is,

In "Bad" category, there are 3 yes tuples and 2 no tuples and in the "Good" category, there are 5 yes tuples and 2 no tuples. Therefore,

$$
\begin{aligned}
Info_{WC}(D) &= \tfrac{5}{12} * (-\tfrac{3}{5}\log_2 \tfrac{3}{5} - \tfrac{2}{5}\log_2 \tfrac{2}{5}) + \tfrac{7}{12} * (-\tfrac{5}{7}\log_2 \tfrac{5}{7} - \tfrac{2}{7}\log_2 \tfrac{2}{7}) \\
&= 0.404 + 0.502 = 0.906 \text{ bits}
\end{aligned}
$$

Hence, the gain in information for such a partitioning would be,

$$Gain(WeatherCondition) = Info(D) - Info_{WC}(D) = 0.918 - 0.906 = 0.012 \text{ bits.}$$

(ii) The second attribute is "Driver's Condition". The expected information needed to classify a tuple in the dataset if the tuples are partitioned according to Driver's Condition is,

In "Sober" category, there are 5 yes tuples and 2 no tuples and in "Alcohol-impaired" category, there are 3 yes tuples and 2 no tuples. Therefore,

$$
\begin{aligned}
Info_{DC}(D) &= \tfrac{7}{12} * (-\tfrac{5}{7}\log_2 \tfrac{5}{7} - \tfrac{2}{7}\log_2 \tfrac{2}{7}) + \tfrac{5}{12} * (-\tfrac{3}{5}\log_2 \tfrac{3}{5} - \tfrac{2}{5}\log_2 \tfrac{2}{5}) \\
&= 0.502 + 0.404 = 0.906 \text{ bits}
\end{aligned}
$$

Hence, the gain in information for such a partitioning would be,

$$Gain(Driver'sCondition) = Info(D) - Info_{DC}(D) = 0.918 - 0.906 = 0.012 \text{ bits.}$$

(iii) The third attribute is "Traffic Violation". The expected information needed to classify a tuple in the dataset if the tuples are partitioned according to Traffic violation is,

In "None" category, there are 2 yes tuples and 1 no tuple. In "Disobey stop sign" category, there are 3 yes tuples and 0 no tuples. In "Exceed speed limit", there are 2 yes tuples and 1 no tuple. In "Disobey traffic signal" category, there are 1 yes tuple and 2 no tuples.

$$
\begin{aligned}
Info_{TV}(D) &= \tfrac{3}{12} * (-\tfrac{2}{3}\log_2 \tfrac{2}{3} - \tfrac{1}{3}\log_2 \tfrac{1}{3}) + \tfrac{3}{12} * (-\tfrac{3}{3}\log_2 \tfrac{3}{3}) + \tfrac{3}{12} * (-\tfrac{2}{3}\log_2 \tfrac{2}{3} - \tfrac{1}{3}\log_2 \tfrac{1}{3}) + \tfrac{3}{12} * (-\tfrac{1}{3}\log_2 \tfrac{1}{3} - \tfrac{2}{3}\log_2 \tfrac{2}{3}) \\
&= 0.229 + 0 + 0.229 + 0.229 = 0.687 \text{ bits}
\end{aligned}
$$

Hence, the gain in information for such a partitioning would be,

$$Gain(TrafficViolation) = Info(D) - Info_{TV}(D) = 0.918 - 0.687 = 0.231 \text{ bits.}$$

(iv) The fourth attribute is "Crash Severity". The expected information needed to classify a tuple in the dataset if the tuples are partitioned according to Crash Severity is,

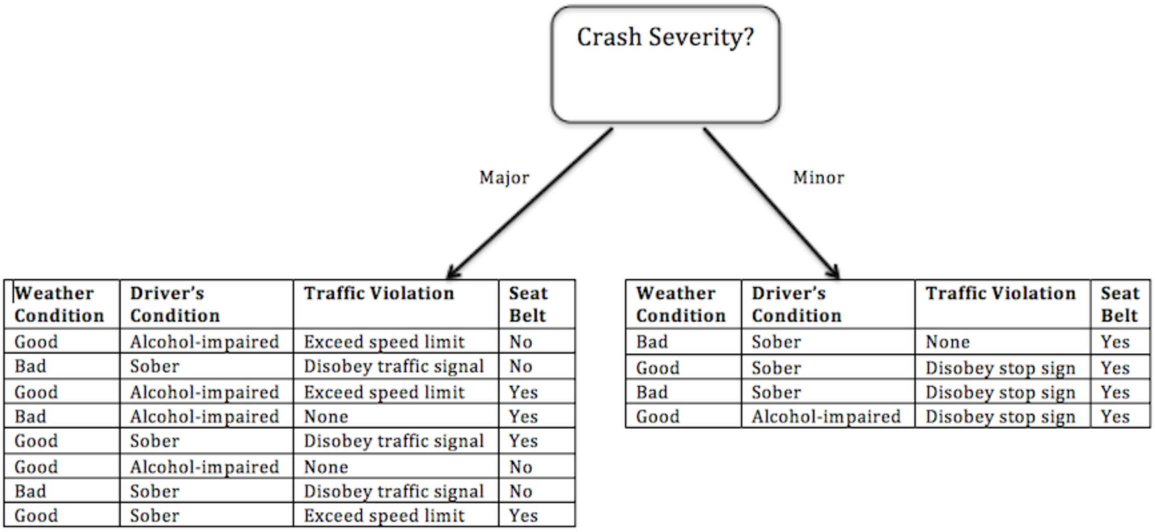
In "Minor" category, there are 4 yes tuples and 0 no tuples. In "Major" category, there are 4 yes tuples and 4 no tuples.

$$
\begin{aligned}
Info_{CS}(D) &= \tfrac{4}{12} * (-\tfrac{4}{4}\log_2 \tfrac{4}{4}) + \tfrac{8}{12} * (-\tfrac{4}{8}\log_2 \tfrac{4}{8} - \tfrac{4}{8}\log_2 \tfrac{4}{8}) \\
&= 0 + 0.666 = 0.666 \text{ bits}
\end{aligned}
$$

Hence, the gain in information for such a partitioning would be,

$$Gain(CrashSeverity) = Info(D) - Info_{CS}(D) = 0.918 - 0.666 = 0.251 \text{ bits.}$$

Because "Crash Severity" has the highest information gain among the attributes, it is selected as the splitting attribute.



Weather Condition	Driver's Condition	Traffic Violation	Seat Belt
Good	Alcohol-impaired	Exceed speed limit	No
Bad	Sober	Disobey traffic signal	No
Good	Alcohol-impaired	Exceed speed limit	Yes
Bad	Alcohol-impaired	None	Yes
Good	Sober	Disobey traffic signal	Yes
Good	Alcohol-impaired	None	No
Bad	Sober	Disobey traffic signal	No
Good	Sober	Exceed speed limit	Yes

Weather Condition	Driver's Condition	Traffic Violation	Seat Belt
Bad	Sober	None	Yes
Good	Sober	Disobey stop sign	Yes
Bad	Sober	Disobey stop sign	Yes
Good	Alcohol-impaired	Disobey stop sign	Yes

Figure 1 :First Level Decision Tree

(b) Since we are using gain ratio as the attribute selection measure, the following is the information calculated:
(i) For the first attribute “Weather Condition” splits the data into two partitions “Bad” and “Good” containing 5 and 7 tuples respectively, therefore,

$$\begin{aligned} SplitInfo_{WC}(D) &= -\tfrac{5}{12}\log_2\tfrac{5}{12} - \tfrac{7}{12}\log_2\tfrac{7}{12} \\ &= 0.526 + 0.453 = 0.97 \end{aligned}$$

$$GainRatio(WeatherCondition) = \tfrac{0.012}{0.97} = 0.012$$

(ii) For the second attribute “Driver’s Condition” splits the data into two partitions “Sober” and “Alcohol-impaired” containing 7 and 5 tuples respectively, therefore,

$$\begin{aligned} SplitInfo_{DC}(D) &= -\tfrac{7}{12}\log_2\tfrac{7}{12} - \tfrac{5}{12}\log_2\tfrac{5}{12} \\ &= 0.453 + 0.526 = 0.97 \end{aligned}$$

$$GainRatio(Driver'sCondition) = \tfrac{0.012}{0.97} = 0.012$$

(iii) For the third attribute “Traffic Violation” splits the data into 4 partitions “None”, “Disobey stop sign”, “Exceed speed limit”, “Disobey traffic signal” containing 3, 3, 3, 3 tuples respectively, therefore,

$$\begin{aligned} SplitInfo_{TV}(D) &= -4 * \tfrac{3}{12}\log_2\tfrac{3}{12} \\ &= 2 \end{aligned}$$

$$GainRatio(TrafficViolation) = \tfrac{0.231}{2} = 0.1155$$

(iv) For the fourth attribute “Crash Severity” splits the data into 2 partitions “Major” and “Minor” containing 4 and 8 tuples respectively, therefore,

$$\begin{aligned} SplitInfo_{CS}(D) &= -\tfrac{4}{12}\log_2\tfrac{4}{12} - \tfrac{8}{12}\log_2\tfrac{8}{12} \\ &= 0.528 + 0.389 = 0.917 \end{aligned}$$

$$GainRatio(CrashSeverity) = \tfrac{0.251}{0.917} = 0.2737$$

Since the “Crash Severity” has the maximum gain ratio, it is selected as the splitting attribute.

Even on using gain ratio measure, we see that the first level of the decision tree is **not** different from the information gain measure.

(c) Let C_1 correspond to the class Seat Belt = yes and C_2 correspond to Seat Belt = no. The tuple we wish to classify is,
X = (Weather Condition = Bad, Driver’s Condition = Sober, Traffic Violation = None, Crash Severity = Major)
We need to maximize $P(\mathbf{X}|C_i)P(C_i)$, for i = 1,2. Therefore,

$$P(SeatBelt = yes) = \tfrac{8}{12} = 0.67$$

$$P(SeatBelt = no) = \tfrac{4}{12} = 0.33$$

The conditional probabilities needed are:

$$P(WeatherCondition = Bad|SeatBelt = yes) = \tfrac{3}{8} = 0.375$$

$$P(WeatherCondition = Bad|SeatBelt = no) = \tfrac{2}{4} = 0.5$$

$$P(Driver'sCondition = Sober|SeatBelt = yes) = \tfrac{5}{8} = 0.625$$

$$P(Driver'sCondition = Sober|SeatBelt = no) = \tfrac{2}{4} = 0.5$$

$$P(TrafficViolation = None|SeatBelt = yes) = \tfrac{2}{8} = 0.25$$

$$P(TrafficViolation = None|SeatBelt = no) = \tfrac{1}{4} = 0.25$$

$$P(CrashSeverity = Major|SeatBelt = yes) = \tfrac{4}{8} = 0.5$$

$$P(CrashSeverity = Major|SeatBelt = no) = \tfrac{4}{4} = 1$$

Using these probabilities, we obtain,

$$\begin{aligned} P(\mathbf{X}|SeatBelt = yes) = P(WeatherCondition = Bad|SeatBelt = yes) * P(Driver'sCondition = Sober|SeatBelt = \\ yes) * P(TrafficViolation = None|SeatBelt = yes) * P(CrashSeverity = Major|SeatBelt = yes) \\ = 0.375 * 0.625 * 0.25 * 0.5 \\ = 0.029 \end{aligned}$$

Similarly,

$$P(\mathbf{X}|SeatBelt = no) = 0.5 * 0.5 * 0.25 * 1 = 0.0625$$

To find the class, C_i that maximizes $P(\mathbf{X}|C_i)P(C_i)$, we compute,

$$P(\mathbf{X}|SeatBelt = yes)P(SeatBelt = yes) = 0.029 * 0.67 = 0.01943$$
$$P(\mathbf{X}|SeatBelt = no)P(SeatBelt = no) = 0.0625 * 0.33 = 0.020625$$

Thus naive Bayesian classifier determines that the Seat Belt was **not** used.