CSci 8980: Data Mining (Fall 2000) Homework #2

Due date: September 25, 2000

1. Compare the following three methods for tree pruning. Provide a brief description of their strength and weakness.

   (a) Stop splitting node if it contains less than 1% of transactions.

   (b) Use half of training data for building tree and use the other half for validating pruning.

   (c) MDL-based pruning.

2. This problem illustrates how the missing attribute values are handled by the C4.5 decision-tree algorithm.

   You are given the training set tabulated below, where "?" denotes the missing values. The final column in this set gives weight of each instance.

X	Y	Class	Weight
P	T	0	1.0
?	T	1	1.0
Q	T	0	1.0
Q	F	1	1.0
R	F	0	1.0
P	F	1	1.0
R	T	0	1.0
P	?	1	1.0
Q	?	0	1.0

   (a) Modify the definition of Entropy (equation 3.3 on page 57 of Mitchell's book) by using the weighted class frequency p_i of Class i, defined as $p_i = (\sum_{t \in C_i} w_t)/(\sum_t w_t)$, where w_t is the weight of instance t. Also, modify the $|S_v|$ and $|S|$ definitions in equation 3.4 to $|S_v| = \sum_{t \in S_v} w_t$ and $|S| = \sum_{t \in S} w_t$.

   Every splitting decision is evaluated using only the training instances where attribute value is known. If training instance t with weight w has a missing value for the splitting attribute A, then t is assigned to all the partitions with modified weights. Modified weight of t in partition S_v is given by $w * |S_v|/|S|$ (using the above definitions, calculated over instances where value of A is known).

   Given this description, calculate the best split decision for the given training set. Assume that X is split three-way (three branches). Partition the training set using the best decision, and show the set of training instances for each partition. Grow one more level of the tree in a similar manner, and stop the induction phase. Assign each leaf node a class with highest weight. Weight of a class C at leaf L, w_C^L, is the sum of weights of instances at L bearing class label C. Weight of the leaf, w^L is the sum of weights of all training instances at L.

   (b) When a new instance with weight w needs to be classified and it has a missing value for the splitting attribute, the instance is sent to all the branches of the node with modified weights $w * |S_v|/|S|$ ($|S_v|$ and $|S|$ are the same as used in the induction phase). If an instance with modified weight w' reaches the leaf L assigned class C, then that instance is predicted to be of class C with probability $w' * w_C^L/w^L$. If class

C is predicted for that instance at more than one leaf nodes, then the probabilities are added. Finally, the new instance is assigned the most probable class.

Given this description of classifying unseen cases and the tree that was grown in part (a) above, predict the class of the following case.

X	Y	Weight
?	?	1.0
R	?	1.0
?	F	1.0

3. Ex. 6.1 of Mitchell's book.

   Hints: Note that "the two tests are independent" means that they are conditionally independent.

   First compute P($\oplus$,$\oplus$) using the techniques similar to the one described on page 158 of Mitchell's book. (Note that P($\oplus$, $\oplus$) $\neq$ P($\oplus$)P($\oplus$).) Then compute P(cancer$|\oplus$, $\oplus$).

4. Ex. 6.2 of Mitchell's book.

5. Given a table containing training samples and test samples t_1 = (T, T, F) and t_2 = (T, F, T), compute P($+|t_1$), P($-|t_1$), P($+|t_2$), and P($-|t_2$).

Instance	a_1	a_2	a_3	classification
1	T	T	T	+
2	T	T	T	+
3	F	F	T	+
4	F	F	F	+
5	T	F	T	-
6	T	F	F	-
7	F	T	F	-
8	F	T	F	+

6. Consider a table containing training samples.

Instance	a_1	a_2	a_3	a_4	classification
1	1	0	0	1	+
2	1	0	0	1	+
3	0	0	0	1	+
4	0	0	1	1	+
5	0	0	1	0	-
6	0	0	1	1	-
7	0	1	1	0	-
8	0	1	1	0	-

   (a) Find estimates for conditional probabilities (P($a_1|+$), P($a_1|-$), P($a_2|+$), P($a_2|-$), ...) by a simple fraction and then predict the class of a test sample (a_1 = 1, a_2 = 1, a_3 = 0, a_4 = 1) using Naive-Bayesian classification.

(b) Find estimates for conditional probabilities using m-estimate of probability discussed in page 179-180 of Mitchell's book and then predict the class of the test sample. Use $p = 0.25$ and $m = 4$.

(c) Comment on the two methods for computing estimates for conditional probabilities in the context of this example, and also in the general scenario.

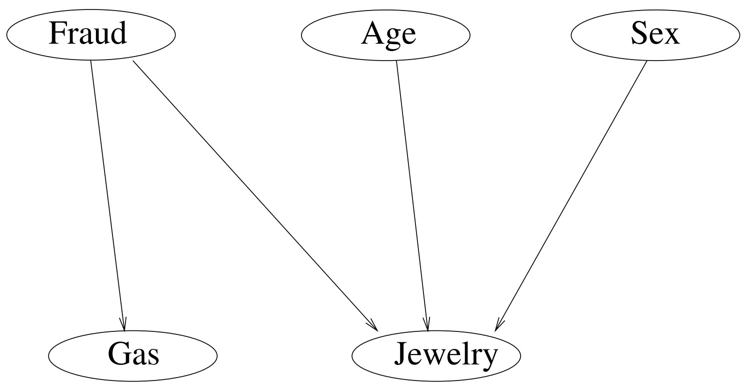
7. A Bayesian network shown below (from "A tutorial on learning with Bayesian networks.", D. Heckerman, 1999) illustrates credit card fraud. There are five variables in this network. The probability of the Jewelry variable is dependent upon the values of the Fraud, Age, and Sex variables. Likewise, the value of the Gas variable is dependent upon the value of the Fraud variable. The Fraud, Age, and Sex variables themselves are not conditioned on any variable.

   Note that '*' is a shorthand for "any state". For the computation of probabilities, show your work for computing the probabilities.

P(f=yes) = 0.00001

P(a=<30) = 0.25
P(a=30-50) = 0.40

P(s=male) = 0.5



P(g=yes | f=yes) = 0.2
P(g=yes | f=no) = 0.01

P(j=yes | f=yes,a=*,s=*) = 0.05
P(j=yes | f=no,a=<30,s=male) = 0.0001
P(j=yes | f=no,a=30-50,s=male) = 0.0004
P(j=yes | f=no,a>50,s=male) = 0.0002
P(j=yes | f=no,a=<30,s=female) = 0.0005
P(j=yes | f=no,a=30-50,s=female) = 0.0002
P(j=yes | f=no,a>50,s=female) = 0.0001

(a) Compute P(j=yes, g=yes, s=male, a=<30, f=yes).

(b) Compute P(j=yes, g=yes, s=female, a=>50, f=yes).

(c) Compute P(j=no, g=yes, a=>50, f=no).

(d) Compute P(f=yes | j=yes, g=no, s=male, a=<30).

(e) Compare the above probability to the prior probability of fraud (i.e., P(Fraud=yes)). What can you say about a transaction that has (j=yes, g=no, s=male, a=<30)?