CS322 Fall 1999
Module 11 (Decision Tree Learning)

Assignment 11

Due: 1:30pm, Friday 26 November 1999.

Question 1

In electronic commerce applications we want to make predictions about what a user will do. Consider the following made-up data used to predict whether someone will ask for more information (*more_info*) based on whether they accessed from an educational domain (*edu*), whether this is a first visit (*first*), whether they have bought goods from an affiliated company (*bought*), and whether they have visited a famous online information store (*visited*).

Example	*bought*	*edu*	*first*	*visited*	*more_info*
e_1	false	true	false	false	true
e_2	true	false	true	false	false
e_3	false	false	true	true	true
e_4	false	false	true	false	false
e_5	false	false	false	true	false
e_6	true	false	false	true	true
e_7	true	false	false	false	true
e_8	false	true	true	true	false
e_9	false	true	true	false	false
e_{10}	true	true	true	false	true
e_{11}	true	true	false	true	true
e_{12}	false	false	false	false	true

We want to use this data to learn the value of *more_info* as a function of the values of the other variables.

Suppose we measure the error of a decision tree as the number of misclassified examples. The optimal decision tree from a class of decision trees is an element of the class with minimal error.

(a) Give the optimal decision tree with only one node. What is the error of this tree?

(b) Give the optimal decision tree of depth 2 (i.e., the root node is the only node with children). For each node in the tree give the examples that are filtered to that node. What is the error of this tree?

(c) Give the decision tree that is produced by the top-down induction algorithm run to completion, where we split on the attribute that reduces the error the most. For each node in the tree specify which examples are filtered to that node. As well as drawing the tree, give the tree in the format of question 2 of assignment 7 (i.e., in terms of $if(Att, Then, Else)$).

(d) Give two instances that don't appear in the examples above and show how they are classified. Use this to explain the bias inherent in the tree (how does the bias give you these particular predications?).

(e) How can overfitting occur in the learned network? Explain in terms of this example.

Question 2

For each question in this assignment, say how long you spent on it. Was this reasonable? What did you learn?