Machine Learning - Assignment 2

Due Tuesday, October 8, 2002

1. [20 points] We discussed in class that if the leaves of a decision tree are impure, then we could report either the majority classification, or the probability of all classes. Consider now a concept learning problem, in which the classes are labeled 0 and 1. Suppose you have a set of instances, for which you can either report the majority class, or the proportion of instances in each class. Show the reporting the majority class minimizes the number of misclassification, whereas reporting the probability minimizes the mean squared error.

2. [30 points] Many decision tree algorithms convert the resulting tree into a set of rules.

   (a) Give an example of a small decision tree that requires an exponential blow-up if expressed as a rule set, or state why this is impossible.

   (b) Most pruning techniques work by replacing a subtree of a decision tree by a single leaf. Explain how you would do this transformation to the rule set resulting from the decision tree.

   (c) Suppose we convert the decision tree into a rule set and then remove one of the antecedents of a rule. How would you convert the resulting rules set back into a tree? Hint: think about what boolean function this rule set represents, and then how you would represent this function using a tree.

   (d) Discuss the relative merits of decision trees and rule set when learning a classification function from scratch, and when changing an existing classification function.

3. [20 points] Imagine that we have a source emitting symbols according to probabilities $p_1, \ldots p_n$. The entropy of the source is then:

$$H(P) = -\sum_{i=1}^{n} p_i \log_2 p_i$$

This is the average number of bits required to specify words from the source, *assuming that the probability distribution is known.* Now assume we do not know the real probability distribution, but we make a guess at it, $Q = \{q_1, \ldots q_n\}$. The *cross-entropy of P given Q* is defined as:

$$H_Q(P) = -\sum_{i=1}^{n} p_i \log_2 q_i$$

This is the average number of bits for encoding a symbol if we assume that emissions happen according to Q. Show that $H(P) \leq H_Q(P)$. Explain what this inequality means.

4. [30 points] This question is meant to give you a little bit of experience using C4.5, the standard decision tree construction algorithm. A link to a great web page regarding C4.5 is posted on the course schedule.

   Your task is the following:

   (a) Download and compile C4.5

   (b) Choose one data set from the UCI repository (link is on the web page)

   (c) Design and run an experiment to test the effect of pruning on the size and accuracy of the decision tree.

   (d) Write a short report explaining what the experiment consisted of, what the results were, and a plot or table of the size and error (or accuracy) vs. the pruning parameter.