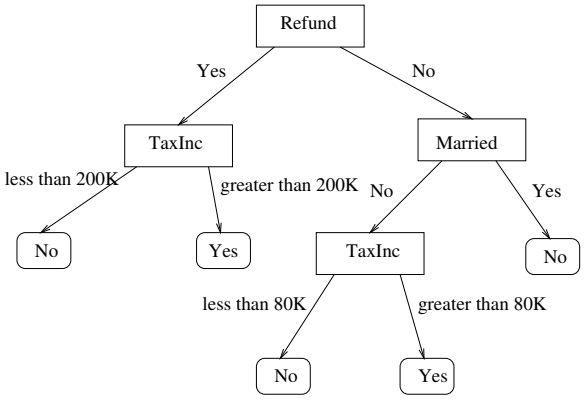
CSci 8980: Data Mining (Fall 2000) Homework #3

Due date: October 4, 2000

1. Consider the following training set and decision tree. You are asked to find rules based on the method similar to C4.5 rules.

Refund	Married	TaxInc	Class
Yes	No	125K	No
No	Yes	100K	No
No	No	70K	No
Yes	Yes	120K	No
No	No	95K	Yes
No	Yes	60K	No
Yes	No	220K	No
Yes	No	250K	Yes
No	No	85K	Yes
No	Yes	75K	No
No	No	90K	Yes



   (a) Find all the rules corresponding to each leaf node of the decision tree.

   (b) Prune each rule according to the method used in C4.5 rules. Instead of using confidence factor, use F1 measure for selecting a term (or condition) to drop. For example, given a rule (A and B) => C. Find recall and precision of this rule, and then compute F1 measure ($F1 = 2 * Recall * Precision/(Recall + Precision)$). Now consider dropping A from this rule, i.e., consider B => C. Find recall and precision of this rule, and then compute F1 measure. Similarly, compute F1 for A => C. If the F1 measure of the original rule is better than these two subset rules, then stop pruning and keep the original rule. Otherwise select the rule with better F1 measure between these two subset rules. Continue pruning with this new rule.

   (c) Following the method used in C4.5 rules, order the classes. Use F1 measure for this process as well.

   (d) Classify the following test samples using the rules.

      i. Refund = No, Married = No, TaxInc = 100K

      ii. Refund = No, Married = Yes, TaxInc = 50K

iii. Refund = Yes, Married = Yes, TaxInc = 90K

2. Compare C4.5 rules and RIPPER (http://portal.research.bell-labs.com/orgs/ssr/people/wcohen/postscript/ml-95-ripper.ps).

   (a) Discuss strengths and weaknesses of both methods.

   (b) Consider a data set that has large difference in the class size (i.e., some classes are much bigger than others). We'd like to find good rules (high precision rules) for small classes. Which method (between C4.5 rules and RIPPER) would be better? Please provide supporting arguments.