CSci6903: Assignment Set # 1
Handed Out on: September 29, 1999
Due Date: October 13, 1999

- Problem 2.4, p.48 of Mitchell's Text Book.
- Problem 3.4, p.78 of Mitchell's Text Book
- C5.0 is a decision-tree learning system successor of C4.5 and ID3. Like its predecessors, C5.0 can be used to learn how to classify unlabeled data from a set of labeled data. Details about C5.0 can be found at http://www.rulequest.com/see5-info.html and a free demo version can be downloaded. After downloading the system, read carefully the C5.0 tutorial located at: http://www.rulequest.com/see5-unix.html. (Please, note that the actual C5.0 software package has been ordered and will be provided shortly. Further details will follow in class, but until then, please use the demo version).
  1. Consider the Sonar Data Set provided in the C5.0 package. Divide this data set randomly into two subsets: the Training and the Testing Sets (C5.0 provides an option (-S) for random sampling the data). The Training Set should contain 2/3 of the data while the Testing Set should contain 1/3. Estimate the accuracy of C5.0 on both the Training and the Testing Sets. Are the results obtained on the Training Set predictive of those obtained on the Testing Set? Repeat this experiment three times with different random partitions. Do you get the same results every time? [Please, note that in future applications, you will be required to perform n-fold cross-validation experiments. This is not required here, because n-fold cross-validation has not yet been formally covered in class (but you are welcome to carry on such experiments if you wish!)]
  2. Experiment with the softening threshold options as well as the pruning options (-c and -m) to see whether either of them can improve C5.0's accuracy rate. Try different degrees of pruning. Repeat each experiment three times with different partitions. Discuss your experimental results. How well do your pruning experiments support Occam's Razor?
  3. Once again divide the Sonar Data Set into a 2/3-1/3 partition. Put the 1/3 partition aside for testing, but modify the 2/3 partition by creating imbalanced training sets as follows: decrease the set of positive training examples (say, the set of mines) by 1/8, 2/8 ... 7/8 of their original size. Test the trees induced by C5.0 from those 7 different training sets on the reserved (and balanced) testing set. Repeat each experiment three times by generating new 2/3-1/3 partitions every time. Discuss your experimental results. In particular, how does an imbalance in the training set affect the accuracy of C5.0?
  4. For each of the 7 cases considered in the previous experiments, re-balance the training set by decreasing the set of negative training examples (say, the set of rocks) by the same amount as the positive training set (the set of mines). i.e., if the positive training set was decreased by 1/8, then also decrease the negative training set by 1/8. Do the same thing for 2/8, 3/8, ... 7/8. Test the trees induced by C5.0 from those 7 different training sets on the reserved (and balanced) testing set. Repeat each experiment three times by generating new 2/3-1/3 partitions every time. Discuss your experimental results. In particular 1) was it helpful or harmful to re-balance the training sets? 2) How do your results compare with those obtained on the full balanced training set?