CMPT-740 Fall 2003
Foundations of Data Mining
Martin Ester

Assignment 4

Total marks: 60
Due date: October 15, 2003

Assignment 4.1

Consider a hypothesis *h* for some boolean concept learned from a training data set of $n = 100$ examples. The observed training error is 0.17, i.e. the hypothesis misclassifies 17 training examples.

(a) What is the standard deviation and the (two-sided) 95% confidence interval for the true error?
(b) How many training examples would you need to assure that the width of the (two-sided) 95% confidence interval of the true error will be at most 0.1?

Assignment 4.2

We want to develop a decision tree classifier that is scalable to large secondary storage data sets. The most expensive operations of a decision tree classifier are as follows:

- Evaluation of all potential splits and selection of the best one.
- Partitioning of the training data according to the chosen split.

To efficiently support these operations on secondary storage data sets, CH sets can be used. The *CH set* for decision tree node *N* and attribute *A* contains one class histogram (i.e., frequency values for each possible class) for each value of *A* representing all training data records belonging to *N*. The *CHgroup* of a node *N* consists of the set of all CH sets (for the different attributes) of node *N*. We assume that the whole CH group of the root node fits into main memory. Thus, for each node of the decision tree the corresponding CH group can be kept memory resident.

The cost of the growth-phase clearly dominates the cost of the pruning phase and, therefore, we ignore the pruning phase here.

(a) Design an algorithm for decision tree construction from secondary storage training data based on CH sets and CH groups. Provide the pseudo-code for your algorithm.

(b) Analyse the runtime complexity of your algorithm using the number of disk page accesses (reads and writes) as the cost measure. How often do you have to read and / or write the whole training data set?