CSCE 496/896: Machine Learning **Fall Semester, 1999**

Homework Assignment 1

September 15, 1999 *Due Date: Friday, October 8*

This is due at the start of class on the due date.

1. Do *any two* of the following
   (a) (25 pts) Problem 1.5 from the text (p. 18)
   (b) (25 pts) Problem 2.10 from the text (p. 50)
   (c) (25 pts) Implement the ID3 algorithm from table 3.1 (p. 56) and run it at least 20 times on randomly generated data sets, where the examples are generated and labeled the same way as described in Problem 2.10. For each of the 20 runs, use a different size for the training set. Then evaluate each of the 20 decision trees on a randomly generated test set of size 100 (use the same test set for each different training set). Then plot the performance on the test set versus the size of the training set. How did increasing the training set size influence generalization error? Did overfitting occur?
2. (10 pts) Do Problem 2.3 on p. 48
3. (5 pts) Do Problem 3.2 on p. 77
4. (15 pts) Do Problem 7.2 on p. 227
5. (15 pts) Let the instance space X be such that each $x \in X$ is represented by n boolean features, and let the concept class C be the power set of X (i.e. an unbiased concept class). Give a *lower bound* on the sample complexity for PAC learning C and argue the correctness of your bound.
6. (5 pts) State how many hours you spent on each problem of this homework assignment (for CSCE 896 students, this includes the next two problems).

The following two problems are only for students registered for CSCE 896. CSCE 496 students who do these will receive extra credit, but the amount will be substantially less than the number of points indicated.

7. (15 pts) Do Problem 3.4 on p. 78
8. (20 pts) Do Problem 7.6 on p. 228