Machine Learning - Assignment 2

Due Tuesday, October 9, 2001

1. [20 points] Suppose that an attribute splits the set of examples D into subsets D_i, and each subset has p_i positive examples and n_i negative examples. Show that unless the ration $\frac{p_i}{p_i+n_i}$ is the same for all i, the attribute will have a strictly positive information gain.

2. [20 points] We discussed in class that if leaves in a decision tree are impure, we could report either the majority classification, or the probabilities of all classes. Consider now a Boolean classification problem. Show that reporting the majority class minimizes the number of misclassified examples at the leaf, while reporting the probability of each class minimizes the sum of the squared errors at the leaf.

3. [10 points] Mitchell, pg.124, Problem 4.5

4. [20 points] Mitchell, pg.125, Problem 4.10

5. Compute the VC dimension for the following classes of hypothesis. Give a good justification for why your answer is correct.

   (a) [10 points] $H =$ triangles in the x-y plane. Points inside the triangle are classified as positive examples

   (b) [20 points] $H =$ convex polygons in the x-y plane. Points inside the polygon are classified as positive examples.