COS 424: Interacting with Data

Homework #2 — Due: Friday, March 9, 2007
Boosting and Bagging — written exercises

See the course website for important information about collaboration and late policies, as well as where and when to turn in assignments. Approximate point values are given in brackets. Be sure to show your work and justify your answers.

Problem 1

[15] Suppose we modify the tree growing algorithm presented in class to use the impurity function

$$I(r) = \min\{r, 1 - r\}.$$

Let us call this the *min-error* impurity function.

As usual, for a split in which p_1 positive and n_1 negative examples reach the left branch, and p_2 positive and n_2 negative examples reach the right branch, the weighted impurity of the split will be

$$(p_1 + n_1) \cdot I\left(\frac{p_1}{p_1 + n_1}\right) + (p_2 + n_2) \cdot I\left(\frac{p_2}{p_2 + n_2}\right).$$

a. Suppose that each branch of this split is replaced by a leaf labeled with the more frequent class among the examples that reach that branch. Show that the number of training mistakes made by this truncated tree is exactly equal to the weighted impurity given above. Thus, using the min-error impurity is equivalent to growing the tree greedily to minimize training error.

b. Suppose the dataset looks like the following. There are three $\{0, 1\}$-valued attributes, and one $\{-, +\}$-valued class label y.

a_1	a_2	a_3	y
0	0	0	+
1	1	0	+
0	1	0	+
1	0	1	−
0	0	1	−
0	1	0	−
1	1	0	−
1	1	1	−
1	0	0	−
1	1	0	−

Which split will be chosen at the root when the Gini index impurity function is used? Which split will be chosen at the root when min-error impurity is used? Explain your answers.

c. Under what general conditions on p_1, n_1, p_2 and n_2, will the weighted min-error impurity of the split be strictly smaller than the min-error impurity before making the split (i.e., of all the examples taken together)?

d. What do your answers to the last two parts suggest about the suitability of min-error impurity for growing decision trees?

Problem 2

[8] Suppose AdaBoost is run on m training examples, and suppose on each round that the weighted training error ϵ_t of the t-th weak hypothesis is at most $1/2 - \gamma$, for some number $\gamma > 0$. Show that after

$$T > \frac{\ln m}{2\gamma^2}$$

rounds, the combined hypothesis H must be consistent with the m training examples, i.e., must have zero training error.

Problem 3

[8] For this problem, start by looking at the programming part of this assignment which describes a procedure called "bagging." Suppose we have a training set of m examples, and we create a bootstrap replicate, as prescribed on each round of bagging. As noted elsewhere, some examples may be included in this bootstrap replicate multiple times, and some examples will be omitted from it entirely.

As a function of m, compute the expected fraction of the training set that does *not* appear at all in the bootstrap replicate. What is the limit of this expectation as $m \to \infty$?

Problem 4

[15] In class, we considered the following training dataset:

example	**label**
228	0
67	1
138	1
209	0
156	1
46	0
197	0
6	0
173	1

We noticed that a training example is positive if and only if it is at least 50 and at most 180. Such a rule has the form

$$h(x) = \begin{cases} 1 & \text{if } a \leq x \leq b \\ 0 & \text{else} \end{cases}$$

for some numbers a and b. Let us call such a rule an *interval function.*

a. Describe an algorithm that, given a dataset consisting of n-bit integers, will efficiently (i.e., in polynomial time), find a consistent interval function, if one exists.

b. Suppose that this algorithm finds an interval function that is consistent with all m training examples. First argue that we can assume without loss of generality that this function will be defined by endpoints a and b which are themselves n-bit integers. Then use the generalization-error result proved in class for finite hypothesis spaces to derive an upper bound on the generalization error of this function in terms of m, n and a confidence parameter δ.

c. Suppose that $n = 8$. How large a sample size m is required to ensure that a consistent function will have accuracy at least 99% with confidence 95% (so that $\delta = 0.05$)?

Problem 5

[15] In this problem, we will explore the *Vapnik-Chervonenkis (VC) dimension* which can be used as a measure of complexity for infinitely large hypothesis spaces in much the same way that $\lg |\mathcal{H}|$ can be used for finite hypothesis spaces.

First, some abstract definitions:

Let $\mathcal{H}$ be any space of $\{0, 1\}$-valued functions over some input space X. We say that a set of examples $x_1, \ldots, x_m$ in X are *shattered* by $\mathcal{H}$ if for all bit sequences $b_1, \ldots, b_m$ in $\{0, 1\}$, there exists a function h in $\mathcal{H}$ such that $h(x_i) = b_i$ for all $i = 1, \ldots, m$.

The *VC-dimension* of the class $\mathcal{H}$ is the cardinality of the largest set shattered by $\mathcal{H}$. That is, the VC-dimension is the largest value of m such that there exists a set of examples $x_1, \ldots, x_m$ that are shattered by $\mathcal{H}$.

a. Suppose now (and for the rest of this problem) that $\mathcal{H}$ is the set of all interval functions as defined in the last problem (but in which the endpoints a and b, as well as the examples x may be arbitrary real numbers, not necessarily integers). Find a set of two examples (i.e., real numbers in this case) x_1 and x_2 that are shattered by $\mathcal{H}$.

b. Show that there is *no* set of three examples that are shattered by $\mathcal{H}$.

c. What is the VC-dimension of $\mathcal{H}$? Explain your answer.