Introduction to Machine Learning, Fall 2012

Problem Set 3: Decision trees & boosting

Due: Thursday, October 25, 2012 by 11am (in class, *before* class begins)

Important: *See problem set policy on the course web site.* You must show **all** of your work and be rigorous in your writeups to obtain full credit.

---

1. (15 points) **Decision Trees**

   We are writing a nature survival guide and need to provide some guidance about which mushrooms are poisonous and which are safe. (Caution - example only - do not eat any mushrooms based on this table.) We gather some examples of both types of mushroom, collected in a table, and decide to train a binary decision tree to classify them for us (two children per node, i.e., each decision chooses some variable to split on, and divides the data into two subsets). We have one real-valued feature (size) and two discrete-valued features (spots and color). Recall that we do binary splits on a real-valued variable by finding the threshhold with the highest information gain (see lecture 11 slides).

y = Poisonous?	x_1 = size (real-valued)	x_2 = spots?	x_3 = color
N	1	N	White
N	5	N	White
N	2	Y	White
N	2	N	Brown
N	3	Y	Brown
N	4	N	White
N	1	N	Brown
Y	5	Y	White
Y	4	Y	Brown
Y	4	Y	Brown
Y	1	Y	White
Y	1	Y	Brown

   Do this problem by hand and show all of your work.

   (a) What is the entropy of the target variable, "poisonous"?

   (b) What is the first attribute a decision tree trained using the entropy or information gain method we discussed in class would use to classify the data?

   (c) What is the information gain of this attribute?

   (d) Draw the full decision tree learned from this data set (no pruning, no bound on its size).

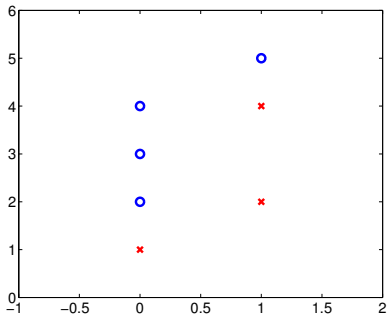
   (e) Now consider the following data, where we wish to predict the target variable Y. Suppose we train a decision tree (again using information gain, and again with no pruning or bound on size).

Y	A	B	C
0	0	0	0
0	0	0	1
0	0	1	0
0	0	1	0
0	0	1	1
1	0	1	1
0	1	0	0
1	1	0	1
1	1	1	0
0	1	1	1
1	1	1	1

What would be the *training* error of our classifier? Give as a percentage, and explain why. (*Hint: you can do this by inspection; there are no significant calculations required.*)

2. (15 points) **Boosting**

Consider the following classification problem:



You wish to use boosting to learn a classifier where the weak learners are decision stumps, i.e. defined by a choice of a single feature: a threshold on the value of x_1 (horizontal axis) or x_2 (vertical axis). To choose each weak learner, the boosting algorithm directly minimizes the weighted classification error over all possible decision stumps.

Find $h_3(x)$, the classifer after running three iterations of AdaBoost (i.e., finding the first three weak learners). At each step, compute the weighted error of the weak learner and use this to compute the resulting classifier weight α_i, and the new example weights. *Show all of your work.* To be consistent for grading purposes, choose the feature x_1, i.e. a vertical line, for the first weak learner.

What is the error of the overall (ensemble) classifier at each stage, i.e., the training error of $h_1(x), h_2(x)$, and $h_3(x)$?

Acknowledgements: These problems are adapted from an assignment developed by Alex Ihler at UC Irvine.