Introduction to Machine Learning, Fall 2013

Problem Set 4: VC dimension & Decision trees

Due: Thursday, October 17, 2013 by 11am (*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) **VC-dimension**

   (a) Show that the VC-dimension of a finite hypothesis set H is at most $\log_2 |H|$.

   (b) Show that the VC-dimension of the set of all closed balls in $\mathbb{R}^d$, that is sets of the form $\{\vec{x} \in \mathbb{R}^d : ||\vec{x} - \vec{x}_0||_2 \leq r\}$ for some $\vec{x}_0 \in \mathbb{R}^d$ and $r \geq 0$, is at most $d + 2$.

   Hint: Recall that the VC-dimension of the set of linear classifiers $\vec{w} \cdot \vec{x} \geq b$ in dimension n (i.e., $\vec{w} \in \mathbb{R}^n$) is $n + 1$. Construct a feature mapping to reduce this hypothesis class to a subset of linear classifiers of dimension $d + 1$, and then apply this result.

   (c) Consider the hypothesis class $\mathcal{H}_\alpha$ defined on the real line $x \in \mathbb{R}$ and parameterized by a single parameter α, given by $\mathcal{H}_\alpha = \{x : x \in [\alpha, \alpha + 1] \text{ or } x \in [\alpha + 2, +\infty]\}$. Show that the VC-dimension of $\mathcal{H}_\alpha$ is exactly 3.
   (Recall that to prove that the VC-dimension is 3, you must (i) demonstrate a set of three points that are shattered by the hypothesis class, and (ii) demonstrate that any set of four or more points *cannot* be shattered by the hypothesis class.)

2. (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 12 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. Your answer must be a **binary tree** *(any branch on x_1 must have an associated threshold).*

(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.*)

Acknowledgements: Problem 1 is adapted from an assignment by Mehryar Mohri.