

# Computer Science 463/563 Assignment 2

Dana Angluin

January 21, 2009

## Instructions

This assignment is to be done (preferably in Latex) and handed in by class time on Friday, January 30. (For a quick introduction to Latex, search “latex tutorial”. You may sign up for a CPSC 463 course account if you need access to Latex or other software for the course.)

This assignment may be discussed in groups, but should be written up individually; please credit all the people (other than the instructor or TA) and written references (online or not) that you consulted in doing these problems.

## Problem 1

Empirical check. We saw in class that if  $D$  is any probability distribution over the unit square and the unknown target concept is an axis-parallel rectangle  $R$ , then if we draw

$$m \geq \frac{4}{\epsilon} \ln \frac{4}{\delta} \tag{1}$$

points and label them using  $R$ , then with probability at least  $(1 - \delta)$ , the smallest axis-parallel rectangle  $S$  containing the positive points misclassifies at most  $\epsilon$  of the probability mass of the points.

This property is true in the worst case, that is, for any distribution  $D$ . How pessimistic is this bound if  $D$  is something “nice”, say the uniform distribution over the unit square? Write a program to estimate the average number of sample points you need to reduce the misclassification rate of  $S$  to at most  $\epsilon$ , when  $D$  is the uniform distribution over the unit square, and compare the results to the bound in (1). How pessimistic is the bound in this case?

Please include a listing of your program in your answer for this problem.

## Problem 2

Neighborly points. Given a sample of labeled points and a new point  $p$  to label, the 1-nearest neighbor algorithm finds the sample point closest to  $p$  (in Euclidean distance) and predicts the label of  $p$  to be the label of that nearest neighbor.

(a) If  $r$  is a nonnegative real number, then  $[0, r]$  denotes the closed interval of the real line consisting of all real numbers  $x$  such that  $0 \leq x \leq r$ . Suppose  $D$  is an arbitrary probability distribution on the nonnegative real numbers and the target concept is some interval  $[0, r]$ . Show that the 1-nearest neighbor algorithm achieves PAC learning. That is, there exists some polynomial  $q$  such that if we draw  $m \geq q(1/\epsilon, 1/\delta)$  samples and use the 1-nearest neighbor algorithm, then with probability at least  $(1 - \delta)$  the 1-nearest neighbor algorithm misclassifies at most  $\epsilon$  of the probability mass of the points.

(b) (Required of graduate students, optional for undergraduates.) If  $r$  is a real number, let  $L_r$  denote the set of points in the plane below (and including) the line  $y = r$ . That is,

$$L_r = \{(x, y) : y \leq r\}.$$

Show that the 1-nearest neighbor algorithm does NOT achieve PAC learning for the class of target concepts  $L_r$ . What can you conclude about using the 1-nearest neighbor algorithm for learning axis-parallel rectangles?

### Problem 3

VC-dimension. Consider points  $(x, y)$  in the plane, and a concept class  $C$  consisting of unions of two finite line segments. Find the VC-dimension of  $C$ . What can you conclude about the relationship between the VC-dimension of a concept class  $C$  and the VC-dimension of the class  $C'$  of unions of two concepts from  $C$ ?