COS 511: Foundations of Machine Learning

Homework #1 — Due:
PAC Learning — February 17, 2003

General comments: Be sure to read the collaboration and late policies on the course web site. In particular, you should attempt to solve these problems on your own before joining a group to solve them together. Also, from time to time, especially if asked to do so by students, I will post hints for some of the problems on the course web site which you can choose to look at, or not (but again, try to solve them first without the hints).

It is a very good idea to start early on the problem sets, at least to read them over so that you can be thinking about them in the background.

Be sure to show your work and justify your answers in a mathematically rigorous fashion. Approximate point values are shown in brackets.

Problem 1

Probability warm-up.

a. [5] Prove that if X and Y are independent real-valued random variables, then $\mathrm{E}[XY] = \mathrm{E}[X]\,\mathrm{E}[Y]$. Also prove that this equality need not hold when X and Y are not independent. (As in class, you can assume that X and Y have a discrete (i.e., countable) range, or set of possible values.)

b. [5] Suppose one card is picked at random from each of four ordinary decks of cards. What is the probability that an ace of spades is among the four selected cards? Answer this question with an *approximate* upper bound using the union bound, and also give an *exact* answer. Evaluate your answers numerically.

c. [5] Suppose now that all four decks of cards are shuffled together and a set of four cards are randomly picked from this huge deck. Now what is the probability that an ace of spades is among the four selected cards? Give an *exact* answer. Can the union bound be used also in this case? If so, what approximate answer does it give? If not, why not? As before, evaluate your answers numerically.

Problem 2

[15] Consider the following learning problem: Let the domain be $X = \mathbb{R}$, and let $\mathcal{C} = \mathcal{C}_s$ be the class of concepts defined by unions of s intervals. That is, each concept c is defined by real numbers $a_1, b_1, \ldots, a_s, b_s$ where $c(x) = 1$ if and only if $x \in [a_1, b_1] \cup \cdots \cup [a_s, b_s]$.

Describe an efficient algorithm that learns the class $\mathcal{C}_s$ for every s, assuming that s is known ahead of time to the learner. You should describe a single algorithm that works for all $\mathcal{C}_s$, provided that s is known so that the learner can choose the number of examples needed as a function of ϵ, δ and s. You can use any hypothesis space you wish. Prove that your algorithm is PAC (i.e., produces a hypothesis with error at most ϵ with probability at least $1 - \delta$), and derive an exact expression for the number of examples needed. Also argue briefly that your algorithm runs in time polynomial in $1/\epsilon$, $1/\delta$ and s.

Problem 3

The Occam's razor result proved in class only applies to finite $\mathcal{H}$. Suppose now that $\mathcal{H}$ is discrete, i.e., either finite or countably infinite. Let $g : \mathcal{H} \to (0, 1]$ be any function such that

$$\sum_{h \in \mathcal{H}} g(h) \leq 1.$$

Although g may look a bit like a probability distribution, you should *not* think of it as one. It is just a function — any function — whose positive values happen to add up to a number not bigger than one.

Let m be the number of given examples (each chosen at random, as usual, from some unknown distribution D).

a. [15] Prove that, with probability at least $1 - \delta$,

$$\mathrm{err}_D(h) \leq \frac{\ln(1/g(h)) + \ln(1/\delta)}{m}$$

for all $h \in \mathcal{H}$ that are consistent with the observed data. As usual, $\mathrm{err}_D(h) = \Pr_{x \sim D}[h(x) \neq c(x)]$, and c is the target concept.

b. [5] Suppose hypotheses in $\mathcal{H}$ are represented by bit strings and that $|h|$ denotes the number of bits needed to represent h. Show how to choose g to prove that

$$\mathrm{err}_D(h) \leq O\left(\frac{|h| + \ln(1/\delta)}{m}\right)$$

for all $h \in \mathcal{H}$ that are consistent with the observed data (with probability at least $1 - \delta$). Give explicit constants (in other words, give a bound that does not use $O(\cdot)$ notation).

c. [5] How does the bound in (b) reflect the intuition that "simpler" hypotheses should be prefered to more "complex" ones? How does the bound in (a) reflect the intuition that prior knowledge helps learning?