6.824 Artificial Intelligence, Fall 1993
Problem Set 1

Due: Wednesday, September 23

Problem 1

Consider an instance space where each instance is an object with n attributes $x_1, x_2, \ldots x_n$ where each x_i takes on real number values in the closed interval $[0, 1]$. In this problem we consider conjunctive concepts, i.e., concepts defined by conjunctions of pure atoms for linear attributes where a pure atom is a formula of the form $r_1 \leq x_i \leq r_2$, $r_1 < x_i \leq r_2$, $r_1 \leq x_i < r_2$, or $r_1 < x_i < r_2$ where r_1 and r_2 are real numbers.

(a) Describe the procedures needed to implement Mitchel's algorithm for this concept space. In other words, describe procedures for updating the most general and most specific versions when a negative and positive example are given.

(b) Show that, for this concept space, the set of most general versions maintained by Mitchell's algorithm (i.e., the upper fringe of the version space) can grow exponentially large. More specifically, give a set of negative examples such that the number of different minimal concepts in this concept space that cover those examples is exponential in n, the number of attributes in the instance space.

Problem 2

Now consider an instance space where each instance is a structure with n attributes $x_1, x_2, \ldots x_n$ where each attribute ranges over the two values T and F. Each instance can be viewed as a truth assignment to the attributes — each instance x assigns the feature x_i either the value T or the value F.

For this instance space a concept can be defined by a Boolean formula built from the attributes. For example, the formula $x_1 \wedge \neg x_2$ is true of an instance x if and only if x assigns the feature x_1 the value T and the feature x_2 the value F. Any Boolean combination of feature names determines a concept.

(a) How many instances are there in the instance space defined by n Boolean attributes.

(b) How many concepts are there for the instance space defined by n Boolean attributes. In answering this question take an *extensional* view of concepts, i.e., assume that two Boolean formulas which define the same set of instances are different representation of the same concept.

(c) Suppose that we are trying to learn some fixed target concept. We are given a set of instances, each selected at random according to some fixed but unknown probability distribution on the instance space. Each selected instance is classified as either a positive or negative example according to the fixed target concept. Give the number of instances sufficient to imply that an arbitrary Boolean concept consistent with these instances will, with probability at least $1-\delta$, incorrectly classify at most a fraction ϵ of the instance space. Your answer should be an expression in terms of δ, ϵ, and the number n of attributes. (This is called the *sample complexity* required for PAC learning for this concept space.)

Problem 3

In this problem we again consider the instance space defined by n Boolean attributes. However we restrict concepts to those that can be expressed as Boolean combinations of literals that involve at most k Binary Boolean connectives each of which is either $\vee$ or $\wedge$. For example, $x_1 \vee (x_2 \wedge \neg x_3)$ is a Boolean combination of literals containing two binary Boolean connectives.

(a) Show that for fixed k the number of examples sufficient for PAC-learning over the concept space of k-binary operators is polynomial in $1/\epsilon, 1/\delta$, and n. Don't worry about getting tight bounds on the sample complexity required for PAC learning — just give some polynomial that is sufficient. Show the derivation of your polynomial.

(b) Explain why the polynomial sample complexity you gave in part **a** does not immediately imply that this concept space is PAC learnable in polynomial time.

Problem 4

Consider an instance space in which each instance is an integer between 1 and n. Consider a concept space in which each concept is a disjunction of pure literals, i.e., a formula of the form $p_1 \leq i \leq q_1 \vee p_2 \leq i \leq q_2 \vee \ldots \vee p_k \leq i \leq q_k$ where i ranges over instances (integers from 1 to n) and each p_i and q_i are integers between 1 and n.

(a) Arguing only from the cardinality of the concept space show that the number of samples sufficient for PAC-learning this concept space is polynomial in $1/\epsilon$, $1/\delta$, k and $\log n$.

(b) Give a polynomial time algorithm for PAC learning for this concept space (it should be polynomial in $1/\epsilon$, $1/\delta$, k and $\log n$).

(c) Show that for a fixed k this concept space has a bounded VC-dimension that is independent of n. Use this fact to give a number of samples sufficient for PAC learning that is indendent of n.