Machine Learning, 11, 23-35 (1993)
© 1993 Kluwer Academic Publishers, Boston. Manufactured in The Netherlands.

Active Learning Using Arbitrary Binary Valued Queries

S.R. KULKARNI KULKARNI@EE.PRINCETON.EDU
Department of Electrical Engineering, Princeton University, Princeton, NJ 08544

S.K. MITTER AND J.N. TSITSIKLIS MITTER@LIDS.MIT.EDU, JNT@LIDS.MIT.EDU
Laboratory for Information and Decision Systems, M.I.T., 77 Mass. Ave., Cambridge, MA 02139

Editor: David Haussler

Abstract. The original and most widely studied PAC model for learning assumes a passive learner in the sense that the learner plays no role in obtaining information about the unknown concept. That is, the samples are simply drawn independently from some probability distribution. Some work has been done on studying more powerful oracles and how they affect learnability. To find bounds on the improvement in sample complexity that can be expected from using oracles, we consider active learning in the sense that the learner has complete control over the information received. Specifically, we allow the learner to ask arbitrary yes/no questions. We consider both active learning under a fixed distribution and distribution-free active learning. In the case of active learning, the underlying probability distribution is used only to measure distance between concepts. For learnability with respect to a fixed distribution, active learning does not enlarge the set of learnable concept classes, but can improve the sample complexity. For distribution-free learning, it is shown that a concept class is actively learnable iff it is finite, so that active learning is in fact less powerful than the usual passive learning model. We also consider a form of distribution-free learning in which the learner knows the distribution being used, so that "distribution-free" refers only to the requirement that a bound on the number of queries can be obtained uniformly over all distributions. Even with the side information of the distribution being used, a concept class is actively learnable iff it has finite VC dimension, so that active learning with the side information still does not enlarge the set of learnable concept classes.

Keywords: PAC-learning, active learning, queries, oracles

1. Introduction

The PAC learning model (e.g., see Blumer et al., 1986; Valiant, 1984) provides a framework for studying the problem of learning from examples. In this model, the learner attempts to approximate an unknown concept from a set of positive and negative examples of the concept. The examples are drawn from some unknown probability distribution, and the same distribution is used to measure the distance between concepts. After some finite number of examples, the learner is required only to output with high probability a hypothesis close to the true concept. A collection of concepts, called a concept class, is said to be learnable if a bound on the number of examples needed to achieve a certain accuracy and confidence in the hypothesis can be obtained uniformly over all concepts in the concept class and all underlying probability distributions.

One goal of studying such a formal framework is to be able to characterize in a precise sense the tractability of learning problems. For the PAC model, Blumer et al. (1986) showed that a concept class is learnable iff it has finite VC dimension, and they provided upper

and lower bounds on the number of examples needed in this case. The requirement that a concept class have finite VC dimension is quite restrictive. There are many concept classes of practical interest with infinite VC dimension that one would like to be and/or feel should be learnable. In fact, even some concept classes of interest in low-dimensional Euclidean spaces are not learnable. For applications such as image analysis, machine vision, and system identification, the concepts might be subsets of some infinite-dimensional function space, and the concept classes generally will not have finite VC dimension. Hence, for many applications the original PAC model is too restrictive in the sense that not enough problems are learnable in this framework.

A natural direction to pursue is to consider extensions or modifications of the original framework that enlarge the set of learnable concept classes. Two general approaches are to relax the learning requirements and to increase the power of the learner-environment or learner-teacher interactions. A considerable amount of work has been done along these lines. For example, learnability with respect to a class of distributions (as opposed to the original distribution-free framework) has been studied (Benedek & Itai, 1988; Kulkarni, 1989, 1991; Natarajan, 1988, 1989). Notably, Benedek and Itai (1988) first studied learnability with respect to a fixed and known probability distribution, and characterized learnability in this case in terms of the metric entropy of the concept class. Others have considered particular instances of learnability with respect to a fixed distribution. Regarding the learner-environment interactions, in the original model the examples provided to the learner are obtained from some probability distribution over which the learner has no control. In this sense, the model assumes a purely passive learner. There has been quite a bit of work done on increasing the power of the learner's information-gathering mechanism. For example, Angluin (1986, 1988) has studied a variety of oracles and their effect on learning; Amsterdam (1988) considered a model that gives the learner some control over the choice of examples by allowing the learner to focus attention on some chosen region of the instance space; and Eisenberg and Rivest (1990) studied the effect on sample complexity of allowing membership queries in addition to random examples. Ben-David et al. (1990) have studied a model in which the information received by the learner at each step consists of an approximation of the distance from the learner's hypothesis to the true concept. They obtained bounds on the sample size needed for learning in this model also in terms of metric entropy.

In this article, we study the limits of what can be gained by allowing the most general set of binary-valued learner-environment interactions, which give the learner complete control over the information gathering. Our focus is on the information (or sample) complexity of learning as opposed to computational complexity. Our goal is to obtain bounds on how much oracles can aid in learning as far as information complexity is concerned. (Of course, oracles can also play a crucial role in reducing computational complexity as well, but we do not consider this issue here.) We consider completely "active" learning in that the learner is allowed to ask arbitrary yes/no (i.e., binary-valued) questions, and these questions need not be decided on beforehand. That is, the questions the learner asks can depend on previous answers and can also be generated randomly. Many of the oracles previously considered in the literature are simply particular types of yes/no questions (although those oracles that provide counterexamples are not). Both active learning with respect to

a fixed distribution and distribution-free active learning are considered. Since we are concerned with active learning, the probability distribution is not used to generate the examples, but is used only to measure the distance between concepts.

Definitions of passive and active learning are provided in section 2. In section 3, active learning with respect to a fixed distribution is considered. A simple information-theoretic argument shows that active learning does not enlarge the set of learnable concept classes, but as expected can reduce the sample complexity of learning. In section 4, distribution-free active learning is considered. In this case, active learning can take place only in the degenerate situation of a finite concept class. We also consider a form of distribution-free learning in which we assume that the learner knows the distribution being used, so that "distribution-free" refers only to the requirement that a bound can be obtained on the number of yes/no questions required independent of the distribution used to measure distance between concepts. However, even in this case active learning surprisingly does not enlarge the set of learnable concept classes, but does reduce the sample complexity as expected.

2. Definitions of passive and active learnability

The definitions below follow closely the notation of Blumer et al. (1986). Let X be a set that is assumed to be fixed and known. X is sometimes called the *instance space*. Typically, X is taken to be either $\mathbf{R}^n$ (especially $\mathbf{R}^2$) or the set of binary n-vectors. A *concept* is a subset of X, and a collection of concepts $C \subseteq 2^X$ will be called a *concept class*. An element $x \in X$ will be called a *sample*, and a pair $\langle x, a \rangle$ with $x \in X$ and $a \in \{0, 1\}$ will be called a *labeled sample*. Likewise, $\bar{x} = (x_1, \ldots, x_m) \in X^m$ is called an *m-sample*, and a *labeled m-sample* is an m-tuple $(\langle x_1, a_1 \rangle, \ldots, \langle x_m, a_m \rangle)$ where $a_i = a_j$ if $x_i = x_j$. For $\bar{x} = (x_1, \ldots, x_m) \in X^m$ and $c \in C$, the *labeled m-sample of c generated by* $\bar{x}$ is given by $sam_c(\bar{x}) = (\langle x_1, I_c(x_1) \rangle, \ldots, \langle x_m, I_c(x_m) \rangle)$, where $I_c(\cdot)$ is the indicator function for the set c. The *sample space* of C is denoted by S_C and consists of all labeled m-samples for all $c \in C$, all $\bar{x} \in X^m$, and all $m \geq 1$.

Let H be a collection of subsets of X. H is called the *hypothesis class*, and the elements of H are called *hypotheses*. Let F_{CH} be the set of all functions $f: S_C \to H$. Given a probability distribution P on X, the *error* of f with respect to P for a concept $c \in C$ and sample $\bar{x}$ is defined as $error_{f,c,P}(\bar{x}) = P(c \Delta h)$, where $h = f(sam_c(\bar{x}))$ and $c \Delta h$ denotes the symmetric difference of the sets c and h. As is often done in the literature, we will be considering the case $H = C$ throughout, so we will simply speak of learnability of C rather than learnability of (C, H), and will use the notation F_C rather than F_{CH}. Finally, in the definition of passive learnability to be given below, the samples used in forming a hypothesis will be drawn from X independently according to the same probability measure P. Hence, an m-sample will be drawn from X^m according to the product measure P^m. We can now state the following definition of passive learnability for a class of distributions.

Definition 1 (Passive Learnability for a Class of Distributions) *Let $\mathcal{P}$ be a fixed and known collection of probability measures. The concept class C is said to be* passively learnable with respect to $\mathcal{P}$ *if there exists a function $f \in F_C$ such that for every $\epsilon, \delta > 0$ there*

is a $0 < m \triangleq m(\epsilon, \delta) < \infty$ *such that for every probability measure* $P \in \mathsf{P}$ *and every* $c \in C$, *if* $\bar{x} \in X^m$ *is chosen at random according to* P^m, *then the probability that* $error_{f,c,P}(\bar{x}) < \epsilon$ *is greater than* $1 - \delta$.

If P is the set of all probability distributions over some fixed σ-algebra of X (which we will denote by P^*), then the above definition reduces to the version from Blumer et al. (1986) of Valiant's (1984) original definition (without restrictions on computability) for learnability for all distributions. If P consists of a single distribution, then the above definition reduces to that used by Benedek and Itai (1988).

By active learning we will mean that the learner is allowed to ask arbitrary yes/no questions. Again, we will define learnability only for the case $H = C$. For a fixed distribution, the only object unknown to the learner is the chosen concept. In this case, an arbitrary binary question provides information of the type $c \in C_0$ where C_0 is some subset of C. That is, all binary questions can be reduced to partitioning C into two pieces and asking to which of the two pieces c belongs. For distribution-free learning (or more generally, learning for a class of distributions), the distribution P is also unknown. In this case, every binary question can be reduced to the form "Is $(c, P) \in q$?" where q is an arbitrary subset of $C \times \mathsf{P}$, so that C and P can be simultaneously and dependently partitioned. Thus, the information the active learner obtains is of the form $(\langle q_1, a_1 \rangle, \ldots, \langle q_m, a_m \rangle)$ where $q_i \subseteq C \times \mathsf{P}$ and $a_i = 1$ if $(c, P) \in q_i$ and $a_i = 0$ otherwise. The q_i correspond to the binary valued (i.e., yes/no) questions and a_i denotes the answer to the question q_i when the true concept and probability measure are c and P, respectively. In general, q_i can be generated randomly or deterministically and can depend on all previous questions and answers $\langle q_1, a_1 \rangle, \ldots, \langle q_{i-1}, a_{i-1} \rangle$. The q_i are not allowed to depend explicitly on the true concept c and probability measure P, but can depend on them implicitly through answers to previous questions. Let $\bar{q} = (q_1, \ldots, q_m)$ denote a set of m questions generated in such a manner, and let $sam_{c,P}(\bar{q})$ denote the set of m question and answer pairs when the true concept and probability measure are c and P, respectively. Let $S_{C,\mathsf{P}}$ denote all sets of m question and answer pairs generated in such a manner for all $c \in C$, $P \in \mathsf{P}$, and $m \geq 1$. By an active learning algorithm we mean an algorithm A for selecting $q_1, \ldots, q_m$ together with a mapping $f: S_{C,\mathsf{P}} \rightarrow C$ for generating a hypothesis from $sam_{c,P}(\bar{q})$. In general, A and/or f may be probabilisitic, which results in probabilistic active learning algorithms. If both A and f are deterministic, we have a deterministic active learning algorithm. Note that if the distribution P is known and is computable, then with a probabilistic algorithm an active learner can simulate the information received by a passive learner by simply generating random examples and asking whether they are elements of the unknown concept.

Definition 2 (Active Learnability for a Class of Distributions) *Let* P *be a fixed and known collection of probability measures. C is said to be* actively learnable with respect to P *if there exists a function* $f: S_{C,\mathsf{P}} \rightarrow C$ *and an algorithm A for selecting* $\bar{q}$ *such that for every* $\epsilon, \delta > 0$ *there is a* $0 < m(\epsilon, \delta) < \infty$ *such that for every probability measure* $P \in \mathsf{P}$ *and every* $c \in C$, *if* $h = f(sam_{c,P}(\bar{q}))$ *then the probability (with respect to any randomness in A and f) that* $P(h \Delta c) < \epsilon$ *is greater than* $1 - \delta$.

3. Active learning for a fixed distribution

In this section, we consider active learning with respect to a fixed and known probability distribution. That is, P consists of a single distribution P that is known to the learner. Benedek and Itai (1988) obtained conditions for passive learnability in this case in terms of a quantity known as metric entropy.

Definition 3 (Metric Entropy) *Let (Y, ρ) be a metric space. Define $N(\epsilon) \equiv N(\epsilon, Y, \rho)$ to be the smallest integer n such that there exists $y_1, \ldots, y_n \in Y$ with $Y = \bigcup_{i=1}^{n} B_\epsilon(y_i)$, where $B_\epsilon(y_i)$ is the open ball of radius ϵ centered at y_i. If no such n exists, then $N(\epsilon, Y, \rho) = \infty$. The* metric entropy *of Y (often called the* ϵ-entropy*) is defined to be* $\log_2 N(\epsilon)$.

$N(\epsilon)$ represents the smallest number of balls of radius ϵ that are required to cover Y. For another interpretation, suppose we wish to approximate Y by a finite set of points so that every element of Y is within ϵ of at least one member of the finite set. Then $N(\epsilon)$ is the smallest number of points possible in such a finite approximation of Y. The notion of metric entropy for various metric spaces has been studied and used by a number of authors (e.g., see Dudley, 1978; Kolmogorov & Tihomirov, 1961; Tikhomirov, 1963).

In the present application, the measure of error $d_P(c_1, c_2) \equiv P(c_1 \Delta c_2)$ between two concepts with respect to a distribution P is a pseudo-metric. Note that $d_P(\cdot, \cdot)$ is generally only a pseudo-metric, since c_1 and c_2 may be unequal but may differ on a set of measure zero with respect to P. For convenience, if P is a distribution we will use the notation $N(\epsilon, C, P)$ (instead of $N(\epsilon, C, d_P)$), and we will speak of the metric entropy of C with respect to P, with the understanding that the metric being used is $d_P(\cdot, \cdot)$.

Benedek and Itai (1988) proved that a concept class C is passively learnable for a fixed distribution P iff C has finite metric entropy with respect to P, and they provided upper and lower bounds on the number of samples required. Specifically, they showed that any passive learning algorithm requires at least $\log_2(1 - \delta)N(2\epsilon, C, P)$ samples and that $(32/\epsilon)\ln(N(\epsilon/2)/\delta)$ samples is sufficient. The following result shows that the same condition of finite metric entropy is required in the case of active learning. In active learning, the learner wants to encode the concept class to an accuracy ϵ with a binary alphabet, so the situation is essentially an elementary problem in source coding from information theory (Gallager, 1968). However, the learner wants to minimize the length of the longest codeword rather than the mean codeword length.

Theorem 1 *A concept class C is actively learnable with respect to a distribution P iff $N(\epsilon, C, P) < \infty$ for all $\epsilon > 0$. Furthermore, $\lceil \log_2(1 - \delta)N(2\epsilon, C, P) \rceil$ queries are necessary, and $\lceil \log_2(1 - \delta)N(\epsilon, C, P) \rceil$ queries are sufficient. For deterministic learning algorithms, $\lceil \log_2 N(\epsilon, C, P) \rceil$ queries are both necessary and sufficient.*

Proof: First consider $\delta = 0$. $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions are sufficient, since one can construct an ϵ-approximation to C with $N(\epsilon, C, P)$ concepts, then ask $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions to identify one of these $N(\epsilon, C, P)$ concepts that is within ϵ of the true concept. $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions are necessary, since by definition every ϵ-approximation to C has at least $N(\epsilon, C, P)$ elements. Hence, with any fewer questions there is necessarily a concept in C that is not ϵ-close to any concept the learner might output.

The essential idea of the argument above is that the learner must be able to encode $N(\epsilon, C, P)$ distinct possibilities and to do so requires $\lceil \log_2 N(\epsilon, C, P) \rceil$ questions. Now, for $\delta > 0$, the learner is allowed to make a mistake with probability δ. In this case, it is sufficient that the learner be able to encode $(1 - \delta)N(\epsilon, C, P)$ possibilities, since the learner could first randomly select $(1 - \delta)N(\epsilon, C, P)$ concepts from an ϵ-approximation of $N(\epsilon, C, P)$ concepts (each with equal probability) and then ask questions to select one of the $(1 - \delta)N(\epsilon, C, P)$ concepts, if there is one, that is ϵ-close to the true concept. To show the lower bound, first note that we can find $N(2\epsilon) = N(2\epsilon, C, P)$ concepts $c_1, \ldots, c_{N(2\epsilon)}$ that are pairwise at least 2ϵ apart, since at least $N(2\epsilon)$ balls of radius 2ϵ are required to cover C. Then the balls $B_\epsilon(c_i)$ of radius ϵ centered at these c_i are disjoint. For each i, if c_i is the true concept, then the learning algorithm must output a hypothesis $h \in B_\epsilon(c_i)$ with probability greater than $1 - \delta$. Hence, if k queries are asked, then

$$
\begin{aligned}
(1 - \delta)N(2\epsilon, C, P) &\leq \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | c = c_i) \\
&= \sum_{i=1}^{N(2\epsilon)} \int \Pr(h \in B_\epsilon(c_i) | c = c_i, q_1, \ldots, q_k) dA(q_1, \ldots, q_k) \\
&= \sum_{i=1}^{N(2\epsilon)} \int \sum_{a_1, \ldots, a_k} \Pr(h \in B_\epsilon(c_i) | c = c_i, \\
&\qquad \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&= \int \sum_{a_1, \ldots, a_k} \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | c = c_i, \\
&\qquad \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&= \int \sum_{a_1, \ldots, a_k} \sum_{i=1}^{N(2\epsilon)} \Pr(h \in B_\epsilon(c_i) | \langle q_1, a_1 \rangle, \ldots, \langle q_k, a_k \rangle) dA(q_1, \ldots, q_k) \\
&\leq \int \sum_{a_1, \ldots, a_k} 1 \, dA(q_1, \ldots, q_k) \\
&= \int 2^k dA(q_1, \ldots, q_k) \\
&= 2^k
\end{aligned}
$$

The integral in the above chain of equalities and inequalities is with respect to the distribution modeling any randomness in the algorithm A used to generate the questions. (We assume that the necessary quantities are integrable with respect to this distribution.) The fourth equality (i.e., where conditioning on $c = c_i$ is dropped) follows from the fact that the hypothesis generated by the learner is independent of the true concept given the queries and answers, and the second inequality follows from the fact that the $B_\epsilon(c_i)$ are disjoint. Thus, since k is an integer, $k \geq \lceil \log_2(1 - \delta)N(2\epsilon, C, P) \rceil$.

Finally, if fewer than $N(\epsilon, C, P)$ possibilities are encoded, then some type of probabilistic algorithm must necessarily be used, since otherwise there would be some concept that the learner would always fail to learn to within ϵ. ■

Thus, compared with passive learning for a fixed distribution, active learning does not enlarge the set of learnable concept classes, but as expected, fewer queries are required in general. However, only a factor of $1/\epsilon$, some constants, and a factor of $1/\delta$ in the logarithm are gained by allowing active learning, which may or may not be significant depending on the behavior of $N(\epsilon, C, P)$ as a function of ϵ.

Note that in active learning very little is gained by allowing the learner to make mistakes with probability δ. That is, there is a very weak dependence on δ in the sample size bounds. In fact for any $\delta \leq 1/2$, we have $\log_2(1 - \delta)N(2\epsilon, C, P) = \log_2 N(2\epsilon, C, P) - \log_2 1/(1 - \delta) \geq \log_2 N(2\epsilon, C, P) - 1$, so that even allowing the learner to make mistakes half the time results in the lower bound differing from the upper bound and the bound for $\delta = 0$ essentially by only the term 2ϵ versus ϵ in the metric entropy. Also, note that theorem 1 is true for learnability with respect to an arbitrary metric and not just those induced by probability measures.

4. Distribution-free active learning

Distribution-free learning (active or passive) corresponds to the case where P is the set of all probability measures P^* over, say, the Borel σ-algebra. A fundamental result of Blumer et al. (1986) relates passive learnability for all distributions (i.e., distribution-free) to the Vapnik-Chervonenkis (VC) dimension of the concept class to be learned.

Definition 4 (Vapnik-Chervonenkis Dimension) *Let $C \subseteq 2^X$. For any infinite set $S \subseteq X$, let $\Pi_C(S) = \{S \cap c : c \in C\}$. S is said to be* shattered *by C if $\Pi_C(S) = 2^S$. The* Vapnik-Chervonenkis dimension *of C is defined to be the largest integer d for which there exists a set $S \subseteq X$ of cardinality d such that S is shattered by C. If no such largest integer exists, then the VC dimension of C is infinite.*

Blumer et al. (1986) proved that a concept class C (satisfying certain measurability conditions with which we will not concern ourselves) is learnable for all distributions iff C has finite VC dimension, and they provided upper and lower bounds on the number of samples required. Specifically, if C has VC dimension $d < \infty$, they showed that $\max(1/2\epsilon \log 1/\delta, d(1 - 2(\epsilon + \delta - \epsilon\delta)))$ samples are necessary and $\max(4/\epsilon \log 2/\sigma, 8d/\epsilon \log 8d/\epsilon)$ samples are sufficient for learnability, although since their work some refinements have been made in these bounds (e.g., see Ehrenfeucht et al., 1989).

The case of distribution-free active learnability is a little more subtle than active learnability for a fixed distribution. For both active and passive learning, the requirement that the learning be distribution-free imposes two difficulties. The first is that there must exist a uniform bound on the number of examples or queries over all distributions—i.e., a bound independent of the underlying distribution. The second is that the distribution is unknown to the learner, so that the learner does not know how to evaluate distances between concepts. Hence, since the metric is unknown, the learner cannot simply replace the concept class with a finite ϵ-approximation as in the case of a fixed and known distribution.

For passive learnability, the requirement that the concept class have finite VC dimension is necessary and sufficient to overcome both of these difficulties. However, for active learning the second difficulty is severe enough that no learning can take place as long as the concept class is infinite.

Lemma 1 *Let C be an infinite set of concepts. If $c_1, \ldots, c_n \in C$ is any finite set of concepts in C, then there exists $c_{n+1} \in C$ and a distribution P such that $d_P(c_{n+1}, c_i) \geq 1/2$ for $i = 1, \ldots, n$.*

Proof: Consider all sets of the form $b_1 \cap b_2 \cap \cdots \cap b_n$ where b_i is either c_i or $\bar{c}_i$. There are at most 2^n distinct sets $B_1, \ldots, B_{2^n}$ of this form. Note that the B_i are disjoint, their union is X, and each c_i for $i = 1, \ldots, n$ consists of a union of certain B_i. Since C is infinite, there is a set $c_{n+1} \in C$ such that for some nonempty B_k, $c_{n+1} \cap B_k$ is nonempty and $c_{n+1} \cap B_k \neq B_k$. Hence, there exist points $x_1, x_2 \in X$ with $x_1 \in c_{n+1} \cap B_k$ and $x_2 \in B_k \setminus c_{n+1}$. Let P be the probability measure that assigns probability 1/2 to x_1 and 1/2 to x_2. For each $i = 1, \ldots, n$, either $B_k \subseteq c_i$ or $B_k \cap c_i = \emptyset$. Thus, in either case, $c_{n+1} \Delta c_i$ contains exactly one of x_1 or x_2 so that $d_P(c_{n+1}, c_i) = 1/2$ for $i = 1, \ldots, n$. ■

Theorem 2 *C is actively learnable for all distributions iff C is finite.*

Proof: If C is finite it is clearly actively learnable, since the learner need only ask $\lceil \log_2 |C| \rceil$ questions where $|C|$ is the cardinality of C to decide which concept is the correct one.

If C is infinite we will show that C is not actively learnable by showing that, after finitely many questions, an adversary could give answers so that there are still infinitely many candidate concepts that are far apart under infinitely many remaining probability distributions. Since C is infinite, we can repeatedly apply the lemma above to obtain an infinite sequence of concepts $c_1, c_2, \ldots$ and an associated sequence of probability measures $P_1, P_2, \ldots$ such that under the distribution P_i, the concept c_i is a distance 1/2 away from all preceding concepts—i.e., for each i, $d_{P_i}(c_i, c_j) = 1/2$ for $j = 1, \ldots, i - 1$.

Now, any question that the active learner can ask is of the form "Is $(c, P) \in q$?" where q is a subset of $C \times \mathcal{P}$. Consider the pairs $(c_1, P_1), (c_2, P_2), \ldots$. Either q or $\bar{q}$ (or both) contain an infinite number of the pairs (c_i, P_i). Thus, an adversary could always give an answer such that an infinite number of pairs (c_i, P_i) remain as candidates for the true concept and probability measure. Similarly, after any finite number of questions, an infinite number of (c_i, P_i) pairs remain as candidates. Thus, by the property that $d_{P_i}(c_i, c_j) = 1/2$ for $j = 1, \ldots, i - 1$, it follows that for any $\epsilon < 1/2$ the active learner cannot learn the target concept. ■

Essentially, if the distribution is unknown, then the active learner has no idea about "where" to seek information about the concept. On the other hand, in passive learnability the examples are provided according to the underlying distribution, so that information is obtained in regions of importance. Hence, in the distribution-free case, random samples (from the distribution used to evaluate performance) are indispensable.

Suppose that we remove the second difficulty by assuming that the learner has knowledge of the underlying distribution. Then the learner knows the metric being used and so can form a finite approximation to the concept class. In this case, the distribution-free requirement plays a part only in forcing a uniform bound on the number of queries needed. Certainly, the active learner can learn any concept class that is learnable by a passive learner, since the active learner could simply ask queries according to the known distribution to simulate a passive learner. However, the following theorem shows that active learning, even with the side information as to the distribution being used, does not enlarge the set of learnable concept classes.

Theorem 3 *If the learner knows the underlying probability distribution, then C is actively learnable for all distributions iff C has finite VC dimension. Furthermore,* $\lceil \sup_P \log_2(1 - \delta)N(2\epsilon, C, P) \rceil$ *questions are necessary and* $\lceil \sup_P \log_2(1 - \delta)N(\epsilon, C, P) \rceil$ *questions are sufficient. For deterministic algorithms,* $\lceil \sup_P \log N(\epsilon, C, P) \rceil$ *questions are both necessary and sufficient.*

Proof: If the distribution is known to the learner, then the result of theorem 1 applies for each distribution. Learnability for all distributions then simply imposes the uniform (upper and lower) bounds requiring the supremum over all distributions for both general (i.e., probabilistic) active learning algorithms and for deterministic algorithms. For the first part of the theorem, we need the following result relating the VC dimension of a concept class to its metric entropy: the VC dimension of C is finite iff $\sup_P N(\epsilon, C, P) < \infty$ for all $\epsilon > 0$ (e.g., see Benedek & Itai, 1988; Kulkarni, 1989, 1991, and references therein). The first part of the theorem follows immediately from this result. ■

Thus, even with this extra "side" information, the set of learnable concept classes is not enlarged by allowing an active learner. However, as before, one would expect an improvement in the number of samples required. A direct comparison is not immediate, since the bounds for passive learnability involve the VC dimension, while the results above are in terms of the metric entropy. A comparison can be made using bounds relating the VC dimension of a concept class to its metric entropy with respect to various distributions, which provide upper and lower bounds to $\sup_P N(\epsilon, C, P)$. Upper bounds are more difficult to obtain, since these require a uniform bound on the metric entropy over all distributions. The lower bounds result from statements of the form that there exists a distribution P (typically a uniform distribution over some finite set of points) for which $N(\epsilon, C, P)$ is greater than some function of the VC dimension. However, most previous lower bounds are not particularly useful for small ϵ—i.e., the bounds remain finite as $\epsilon \to 0$. This is the best that can be obtained assuming only that C has VC dimension $d < \infty$, since C itself could be finite. The following result assumes that C is infinite but makes no assumption about the VC dimension of C.

Lemma 2 *Let C be a concept class with an infinite number of distinct concepts. Then for each* $\epsilon > 0$ *there is a probability distribution P such that* $N(\epsilon, C, P) > 1/(2\epsilon)$.

Proof: First, we show by induction that given n distinct concepts, $n - 1$ points $x_1, \ldots, x_{n-1}$ can be found such that the n concepts give rise to distinct subsets of $x_1, \ldots, x_{n-1}$. This is clearly true for $n = 2$. Suppose it is true for $n = k$. Then for $n = k + 1$ concepts $c_1, \ldots, c_{k+1}$, apply the induction hypothesis to $c_1, \ldots, c_k$ to get $x_1, \ldots, x_{k-1}$, which distinguish $c_1, \ldots, c_k$. c_{k+1} can agree with at most one of $c_1, \ldots, c_k$. Then another point x_k can be chosen to distinguish these two.

Now, let $\epsilon > 0$ and set $n = \lfloor 1/2\epsilon \rfloor$. Let $c_1, \ldots, c_n$ be n distinct concepts in C, and let $x_1, \ldots, x_{n-1}$ be $n - 1$ points that distinguish $c_1, \ldots, c_n$. Let P be the uniform distribution on $x_1, \ldots, x_{n-1}$. Since the c_i are distinguished by the x_i, $d_P(c_i, c_j) \geq 1/(n - 1) = 1/(\lfloor 1/2\epsilon \rfloor - 1) > 2\epsilon$. Hence, every concept is within ϵ to at most one of $c_1, \ldots, c_n$ so that $N(\epsilon, C, P) \geq n = \lfloor 1/2\epsilon \rfloor$. ■

The following theorem summarizes the result of the lemma and prevous upper and lower bounds obtained by others.

Theorem 4 *Let C be a concept class with infinitely many concepts and let $1 \leq d < \infty$ be the VC dimension of C. For $\epsilon \leq 1/4$,*

$$\sup_P \log_2 N(\epsilon, C, P) \geq \max\left(2d(1/2 - 2\epsilon)^2 \log_2 e, \log_2 \frac{1}{2\epsilon}\right)$$

and for $\epsilon \leq 1/(2d)$,

$$\sup_P \log_2 N(\epsilon, C, P) \leq d \log_2\left(\frac{2e}{\epsilon} \ln \frac{2e}{\epsilon}\right) + 1$$

Proof: The first term of the lower bound is from Kulkarni (1989, 1991), and the second term of the lower bound follows from lemma 2. The upper bound is from Haussler (1990), which is a refinement of a result from Pollard (1984) using techniques originally from Dudley (1978). A weaker upper bound was also given in Benedek and Itai (1988). ■

This theorem gives bounds on $\sup_P \log_2 N(\epsilon, C, P)$, which can be used with theorem 3 to obtain bounds on the number of questions needed in distribution-free active learning (with the side information of the distribution being used) directly in terms of ϵ, δ, and the VC dimension of C. As stated, the bounds are directly applicable to deterministic active learning algorithms or for active learning with $\delta = 0$. For probabilistic algorithms with $\delta > 0$, the quantity $\log_2 1/(1 - \delta)$ needs to be subtracted from both the lower and upper bounds of theorem 4. Specifically, for distribution-free active learning (with side information as to the distribution being used), $\max(2d(1/2 - 4\epsilon)^2 \log_2 e, \log_2 1/4\epsilon) - \log_2 1/(1 - \delta)$ samples are necessary for $\epsilon \leq 1/8$, and $d\log_2(2e/\epsilon \ln 2e/\epsilon) + 1 - \log_2 1/(1 - \delta)$ samples are sufficient for $\epsilon \leq 1/(2d)$. As with the bounds for fixed distribution active learning, note the very weak dependence of the bounds on δ. The primary difference between the bounds for active (with side information) versus passive distribution-free learning in the $1/\epsilon$ and $\log(1/\delta)$ behavior of the passive learning bounds.

5. Discussion

In this article we considered the effect on PAC learnability of allowing a rich set of learner-environment interactions. Previous work along these lines has provided the learner with access to various types of oracles. Many of the oracles considered in the literature answer queries that are special cases of yes/no questions (although those oracles that provide counter-examples are not of this type). As expected, the use of oracles can often aid in the learning process. To understand the limits of how much could be gained through oracles, we have considered an active learning model in which the learner chooses the information received by asking arbitrary yes/no questions about the unknown concept and/or probability distribution. Our focus was on the information complexity of learning, although we recognize that oracles can also play an important role in reducing the computational complexity of learning algorithms. In fact, our results indicate that sometimes (depending on the metric entropy) the improvement in sample complexity may not be too significant, so that the computational role may often be the primary benefit of using oracles.

For a fixed distribution, active learning does not enlarge the set of learnable concept classes, but it does have lower sample complexity than passive learning. For distribution-free active learning, the set of learnable concept classes is drastically reduced to the degenerate case of finite concept classes. Furthermore, even if the learner is told the distribution but is still required to learn uniformly over all distributions, a concept class is actively learnable iff it has finite VC dimension.

For completeness, we mention that results can also be obtained if the learner is provided with "noisy" answers to the queries. The effects of various types of noise in passive learning have been studied (Angluin & Laird, 1988; Kearns & Li, 1988; Sloan, 1988). For active learning, two natural noise models are random noise in which the answer to a query is incorrect with some probability $\eta < 1/2$ independent of other queries, and malicious noise in which an adversary gets to choose a certain number of queries to receive incorrect answers. For random noise, the problem is equivalent to communication through a binary symmetric channel. Specifically, the answer to each query can be considered as either a 0 ("no") or a 1 ("yes"). Getting noisy answers corresponds to communicating through a noisy channel that transmits the correct value with probability $1 - \eta$, but with probability η transmits a 0 as a 1 or a 1 as a 0. This channel behavior is precisely the definition of the standard example from information theory of a binary symmetric channel. Thus, standard results from information theory on the capacity and coding for such channels (Gallager, 1968) can be applied for this model of random noise. For malicious noise, some results on binary searching with these types of errors (Rivest et al., 1980) can be applied. For both noise models, the conditions for fixed distribution and distribution-free learnability are the same as the noise-free case, but with a larger sample complexity. However, the more interesting aspects of our results are the indications of the limitations of active learning, and the noise-free case makes stronger negative statements.

Finally, an open problem that may be interesting to pursue is to study the reduction in sample complexity of distribution-free learning if the learner has access to both random examples and arbitrary yes/no questions. This is similar to the problem considered in Eisenberg and Rivest (1990), but there the learner could only choose examples to be labeled rather than ask arbitrary questions. Our result for the case where the learner knows the

distribution being used provides a lower bound, but if the distribution is not known, then we expect that for certain concept classes much stronger lower bounds would hold. In particular, we conjecture that results analogous to those in Eisenberg and Rivest (1990) hold in the case of arbitrary binary-valued questions, so that, for example, asking yes/no questions could reduce the sample complexity to learn a dense-in-itself concept class by only a constant factor.

Acknowledgments

We would like to thank David Tse, Ofer Zeitouni, Avi Lele, and Ron Rivest for helpful discussions. This work was supported by the U.S. Army Research Office under Contract DAAL03-86-K-0171, by the Department of the Navy under Air Force Contract F19628-90-C-0002, by the National Science Foundation under contract ECS-8552419, and by the Air Force under contract AFOSR-91-0368. This work was partially done while S. Kulkarni was with the Laboratory for Information and Decision Systems, M.I.T.

References

Amsterdam, J. (1988). Extending the Valiant learning model. *Proceedings of the 5th International Conference on Machine Learning* (pp. 381–394). Philadelphia, PA: Morgan Kaufmann.

Angluin, D., & Laird, P. (1988). Learning from noisy examples. *Machine Learning*, *2*, 343–370.

Angluin, D. (1986). *Types of queries for concept learning* (Technical Report YALEU/DCS/TR-479). New Haven, CT: Yale University, Department of Computer Science.

Angluin, D. (1988). Queries and concept learning. *Machine Learning*, *2*, 319–342.

Ben-David, S., Itai A., & Kushilevitz, E. (1990). Learning by distances. *Proceedings of the Third Annual Workshop on Computational Learning Theory* (pp. 232–245). Rochester, NY: Morgan Kaufmann.

Benedek, G.M., & Itai, A. (1988). Learnability by fixed distributions. *Proceedings of First Workshop on Computational Learning Theory* (pp. 80–90). Cambridge, MA: Morgan Kaufmann.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1986). Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. *Proceedings of the 18th ACM Symposium on Theory of Computing* (pp. 273–282). Berkeley, CA.

Dudley, R.M. (1978). Central limit theorems for empirical measures. *Annals of Probability*, *6*(6), 899–929.

Ehrenfeucht, A., Haussler, D., Kearns, M., & Valiant L. (1989). A general lower bound on the number of examples needed for learning. *Information and Computation*, *82*(3), 247–251.

Eisenberg, B., & Rivest, R.L. (1990). On the sample complexity of pac-learning using random and chosen examples. *Third Workshop on Computational Learning Theory* (pp. 154–162). Rochester, NY: Morgan Kaufmann.

Gallager, R.G. (1968). *Information theory and reliable communication*. New York: Wiley & Sons.

Haussler, D. (1990). *Decision theoretic generalizations of the PAC model for neural net and other learning applications* (Technical Report UCSC-CRL-91-02). Santa Cruz, CA: University of California-Santa Cruz, Computer Research Laboratory.

Kearns, M., & Li, M. (1988). Learning in the presence of malicious errors. *Proceedings of the 20th ACM Symposium on Theory of Computing* (pp. 267–279). Chicago, IL.

Kolmogorov, A.N., & Tihomirov, V.M. (1961). ϵ-Entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations*, *17*, 277–364.

Kulkarni, S.R. (1989). On metric entropy, Vapnik-Chervonenkis dimension and learnability for a class of distributions (Report CICS-P-160). Cambridge, MA: M.I.T., Center for Intelligent Control Systems.

Kulkarni, S.R. (1991). *Problems of computational and information complexity in machine vision and learning*. Ph.D. thesis, Department of Electrical Engineering and Computer Science, M.I.T., Cambridge, MA.

Natarajan, B.K. (1988). Learning over classes of distributions. *Proceedings of the First Workshop on Computational Learning Theory* (pp. 408–409). Cambridge, MA: Morgan Kaufmann.

Natarajan, B.K. (1989). Probably-approximate learning over classes of distributions. Unpublished manuscript.

Pollard, D. (1984). *Convergence of stochastic processes.* New York, NY: Springer-Verlag.

Rivest, R.L., Meyer, A.R., Kleitman, D.J., Winklmann, K., & Spencer, J. (1980). Coping with errors in binary search procedures. *Journal of Computer and System Sciences, 20,* 396–404.

Sloan, R. (1988). Types of noise in data for concept learning. *Proceedings of the First Workshop on Computational Learning Theory* (pp. 91–96). Cambridge, MA: Morgan Kaufmann.

Tikhomirov, V.M. (1963). Kolmogorov's work on ϵ-entropy of functional classes and the superposition of functions. *Russian Mathematical Surveys, k8,* 51–75.

Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM, 27*(11), 1134–1142.

Vapnik, V.N., & Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and its Applications, 16*(2), 264–280.

Received January 24, 1991
Accepted January 27, 1992
Final Manuscript July 16, 1992