September 1989 LIDS-P-1910

On Metric Entropy, Vapnik-Chervonenkis Dimension, and Learnability for a Class of Distributions[1]

Sanjeev R. Kulkarni[2]

July 18, 1989

Abstract

In [23], Valiant proposed a formal framework for distribution-free concept learning which has generated a great deal of interest. A fundamental result regarding this framework was proved by Blumer et al. [6] characterizing those concept classes which are learnable in terms of their Vapnik-Chervonenkis (VC) dimension. More recently, Benedek and Itai [4] studied learnability with respect to a fixed probability distribution (a variant of the original distribution-free framework) and proved an analogous result characterizing learnability in this case. They also stated a conjecture regarding learnability for a class of distributions.

In this report, we first point out that the condition for learnability obtained in [4] is equivalent to the notion of finite metric entropy (which has been studied in other contexts). Some relationships, in addition to those shown in [4], between the VC dimension of a concept class and its metric entropy with respect to various distributions are then discussed. Finally, we prove some partial results regarding learnability for a class of distributions.

[1] Supported by the U.S. Army Research Office contract DAAL03-86-K-0171 (Center for Intelligent Control Systems) and by the Department of the Navy for SDIO.

[2] Center for Intelligent Control Systems, M.I.T., 77 Massachusetts Ave., 35-423, Cambridge, MA, 02139 and M.I.T. Lincoln Laboratory, 244 Wood St., Lexington, MA 02173.

1 Introduction

In [23], Valiant proposed a precise framework to capture the notion of what we mean by 'learning from examples'. The essential idea consists of approximating an unknown 'concept' from a finite number of positive and negative 'examples' of the concept. For example, the concept might be some unknown geometric figure in the plane, and the positive and negative examples are points inside and outside the figure, respectively. The goal is to approximate the figure from a finite number of such points. The examples are assumed to be drawn according to some probability distribution, and the same distribution is used to evaluate how well a concept is learned. However, no assumptions are made about which particular distribution is used. That is, the learning is required to take place for every distribution.

Valiant's seminal paper [23] has spawned a large amount of work analyzing and extending the formal learning framework which was originally proposed. A fundamental paper was written by Blumer et al. [6] which gave a characterization of learnability for the distribution-free framework in terms of a combinatorial parameter which measures the 'size' of a concept class. Benedek and Itai [4] studied a variation of Valiant's learning framework in which the examples are assumed to be drawn from a fixed and known distribution. They gave a characterization of learnability in this case in terms of a different measure of the size of a concept class.

In Section 2, we give some definitions, a precise description of the learning framework, and some previous results from [6] and [4]. The definitions and notation used are essentially those from [6], which are a slight variation from those originally given in [23]. The major result of [6] states that a concept class is learnable for every distribution iff it has finite Vapnik-Chervonenkis (VC) dimension. An analogous result of [4] characterizes learnability for a fixed distribution. We point out that their characterization is identical to that of finite metric entropy, which has been studied in other contexts. The results characterizing learnability suggest that there may be relationships between the VC dimension of a concept class and its metric entropy with respect to various distributions. Some such relationships, in addition to those investigated in [4], are discussed in Section 3. We state an earlier result from [8] and prove a new result, both of which offer some improvements on different results of [4]. In Section 4 we consider learnability for a class of distributions, which is a natural extension of learnability for a fixed distribution. Benedek and Itai [4] posed the characterization of learnability in this case as an open problem. They conjectured that a concept class is learnable with respect to a class of distributions iff the metric entropy of the concept class with respect to each distribution is uniformly bounded over the class of distributions. We prove some partial results for this problem. Although the results we prove are far from verifying the conjecture in general, they are consistent with it. Furthermore, they provide some indication of conditions when power is gained by requiring learnability only for a class of distributions rather than for all distributions. Finally, in Section 5 we briefly summarize and mention some related work that has been done on Valiant's learning framework.

2 Definitions and Previous Results Characterizing Learnability

In this section, we describe the formal model of learning introduced by Valiant [23] (learnability for all distributions) and a variant (learnability for a fixed distribution), and we state previous results characterizing learnability in these cases. The result of Blumer et al. [6] characterizes learnability for all distributions in terms of a quantity known as the VC dimension. The result of Benedek and Itai [4] characterizes learnability for a fixed distribution in terms of a quantity which we point out is essentially metric entropy.

Informally, Valiant's learning framework can be described as follows. The *learner* wishes to learn a concept unknown to him. The *teacher* provides the learner with random positive and negative examples of the concept drawn according to some probability distribution. From a finite set of examples, the learner outputs a hypothesis which is his current estimate of the concept. The error of the estimate is taken as the probability that the hypothesis will incorrectly classify the next randomly chosen example. The learner cannot be expected to exactly identify the concept since only a finite number of examples are seen. Also, since the examples are randomly chosen, there is some chance that the hypothesis will be very far off (due to poor examples). Hence, the learner is only required to closely approximate the concept with sufficiently high probability from some finite number of examples. Furthermore, the number of examples required for a given accuracy and confidence should be independent of the distribution from which the examples are drawn. Below, we will describe this framework precisely, following closely the notation of [6].

Let X be a set which 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* will refer to 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, $\overline{x} = (x_1, \ldots, x_m) \in X^m$ 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 $\overline{x} = (x_1, \ldots, x_m) \in X^m$ and $c \in C$, the *labeled m-sample of c generated by* $\overline{x}$ is given by $sam_c(\overline{x}) = (\langle x_1, I_c(x_1) \rangle, \ldots, \langle x_m, I_c(x_m) \rangle)$ 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 $\overline{x} \in X^m$, 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 $f : S_C \rightarrow H$. A function $f \in F_{CH}$ is called *consistent* if it always produces a hypothesis which agrees with the samples, i.e. whenever $h = f(\langle x_1, a_1 \rangle, \ldots, \langle x_m, a_m \rangle)$ we have $I_h(x_i) = a_i$ for $i = 1, \ldots, m$. Given a probability distribution P on X, the *error* of f with respect to P for a concept $c \in C$ and sample $\overline{x}$ is defined as $error_{f,c,P}(\overline{x}) = P(c \Delta h)$ where $h = f(sam_c(\overline{x}))$ and $c \Delta h$ denotes the symmetric difference of the sets c and h. Finally, in the definition of 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 learnability for every distribution, which is the version from Blumer et al. [6] of Valiant's [23] original definition (without restrictions on computability – see below).

Definition 1 (Learnability for Every Distribution) *The pair (C, H) is* learnable *if there exists a function $f \in F_{CH}$ such that for every $\epsilon, \delta > 0$ there is a $0 < m < \infty$ such that for every probability measure P and every $c \in C$, if $\overline{x} \in X^m$ is chosen at random according to P^m then the probability that $error_{f,c,P}(\overline{x}) < \epsilon$ is greater than $1 - \delta$.*

Several comments concerning this definition are in order. First, learnability depends on both the concept class C and the hypothesis class H, which is why we defined learnability in terms of the pair (C, H). However, in the literature the case $H \supseteq C$ is often considered, in which case, for convenience, we may speak of learnability of C in place of (C, C). Second, the sample size m is clearly a function of ϵ and δ but a fixed $m = m(\epsilon, \delta)$ must work uniformly for every distribution P and concept $c \in C$. Because of this, the term *distribution-free learning* is often used to describe this learning framework. Finally, ϵ can be thought of as an accuracy parameter while δ can be thought of as a confidence parameter. The definition requires that the learning algorithm f output a hypothesis that with high probability (greater than $1 - \delta$) is approximately correct (to within ϵ). Angluin and Laird [2] used the term *probably approximately correct* (PAC) learning to describe this definition.

A somewhat more general and useful definition of learnability was actually used by Valiant in [23] and later by others. This definition incorporates both a notion of the size or complexity of concepts and the central idea that the learning algorithm (i.e., the function which produces a hypothesis from labeled samples) should have polynomial complexity in the various parameters. Other variations of this definition, such as seeing positive examples only, or having the choice of positive or negative examples, have also been considered. Some equivalences among the various definitions learnability were shown in [10]. In this report, we will not consider these variations. Also, we will be considering the case that $H \supseteq C$ throughout, so that we will simply speak of learnability of C rather than learnability of (C, H).

A fundamental result of Blumer et al. [6] relates learnability for every distribution to the Vapnik-Chervonenkis (VC) dimension of the concept class to be learned. The notion of VC dimension was introduced in [25] and has been studied and used in [8, 26, 11]. Many interesting concept classes have been shown to have finite VC dimension.

Definition 2 (Vapnik-Chervonenkis Dimension) *Let $C \subseteq 2^X$. For any finite 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.*

A concept class C will be called *trivial* if C contains only one concept or two disjoint concepts. In [6], a definition was also given for what they called a *well-behaved* concept class, which involves the measurability of certain sets used in the proof of their theorem. We will not concern ourselves with the definition here. The following theorem is stated exactly from [6] and was their main result.

Theorem 1 *For any nontrivial, well-behaved concept class C, the following are equivalent:*

(i) The VC dimension of C is finite.

(ii) C is learnable.

(iii) If d is the VC dimension of C then

(a) for sample size greater than $\max(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{8d}{\epsilon})$, any consistent function $f \in F_{CH}$ is a learning algorithm for C, and

(b) for $\epsilon < \frac{1}{2}$ and sample size less than $\max(\frac{1}{2\epsilon}\log\frac{1}{\delta}, d(1-2(\epsilon+\delta-\epsilon\delta)))$, no function $f \in F_{CH}$ where $C \subseteq H$ is a learning algorithm for C.

A definition of learnability similar to that of Definition 1 can be given for the case of a single, fixed, and known probability measure.

Definition 3 (Learnability for a Fixed Distribution) *Let P be a fixed and known probability measure. The pair (C, H) is said to be* learnable with respect to P *if there exists a function $f \in F_{CH}$ such that for every $\epsilon, \delta > 0$ there is a $0 < m < \infty$ such that for every $c \in C$, if $\overline{x} \in X^m$ is chosen at random according to P^m then the probability that $error_{f,c,P}(\overline{x}) < \epsilon$ is greater than $1 - \delta$.*

Conditions for learnability in this case were studied by Benedek and Itai [4]. They introduced the notion of what they called a 'finite cover' for a concept class with respect to a distribution and were able to show that finite coverability characterizes learnability for a fixed distribution. It turns out that their definition of finite coverability is identical to the notion of metric entropy, which has been studied in other literature. Specifically, the measure of error between two concepts with respect to a distribution is a semi-metric. The notion of finite coverability is identical to the notion of finite metric entropy with respect to the semi-metric induced by the distribution P.

We define metric entropy below, but first show that P induces a semi-metric on the concept class. Define $d_P(c_1, c_2) = P(c_1 \Delta c_2)$ for $c_1, c_2 \subseteq X$ and measurable with respect to P. For $c_1, c_2 \in C$,

$d_P(c_1, c_2)$ just represents the error between c_1 and c_2 that has been used throughout. In the following proposition we prove that $d_P(\cdot, \cdot)$ defines a semi-metric on the set of all subsets of X measurable with respect to P, and hence defines a semi-metric on the concept class C.

Proposition 1 *For any probability measure P, $d_P(c_1, c_2) = P(c_1 \Delta c_2)$ is a semi-metric on the σ-algebra $\mathcal{S}$ of subsets of X measurable with respect to P. I.e., for all $c_1, c_2, c_3 \in \mathcal{S}$*

(i) $d_P(c_1, c_2) \geq 0$

(ii) $d_P(c_1, c_2) = d_P(c_2, c_1)$

(iii) $d_P(c_1, c_3) \leq d_P(c_1, c_2) + d_P(c_2, c_3)$

Proof: (i) is true since P is a probability measure, and (ii) is true since $c_1 \Delta c_2 = c_2 \Delta c_1$. (iii) follows from subadditivity and the fact that $c_1 \Delta c_3 \subseteq (c_1 \Delta c_2) \cup (c_2 \Delta c_3)$. ∎

Note that $d_P(\cdot, \cdot)$ is only a semi-metric since it does not usually satisfy the requirement of a metric that $d_P(c_1, c_2) = 0$ iff $c_1 = c_2$. That is, c_1 and c_2 may be unequal but may differ on a set of measure zero with respect to P, so that $d_P(c_1, c_2) = 0$.

We now define metric entropy.

Definition 4 (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 $Y = \cup_{i=1}^{n} B_\epsilon(y_i)$ 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 which 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 [8, 9, 12, 16, 17, 22]).

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 [4] proved that a concept class C is learnable for a fixed distribution P iff C has finite metric entropy with respect to P. We state their results formally in the following theorem, which we have written in a form analagous to Theorem 1.

Theorem 2 *Let C be a concept class and P be a fixed and known probability measure. The following are equivalent:*

(i) The metric entropy of C with respect to P is finite for all $\epsilon > 0$.

(ii) C is learnable with respect to P.

(iii) If $N(\epsilon) = N(\epsilon, C, P)$ is the size of a minimal ϵ-approximation of C with respect to P and $C^{(\epsilon/2)} = \{y_1, \ldots, y_{N(\epsilon/2)}\}$ is an $\frac{\epsilon}{2}$-approximation to C then

(a) for sample size greater than $(32/\epsilon)\ln(N(\epsilon/2)/\delta)$ any function $f : S_C \rightarrow C^{(\epsilon/2)}$ which minimizes the number of disagreements on the samples is a learning algorithm for C, and

(b) for sample size less than $\log_2[(1-\delta)N(2\epsilon)]$ no function $f \in F_{CH}$ is a learning algorithm for C.

Note that in condition (iii)(a), only functions whose range was a finite $\frac{\epsilon}{2}$-approximation to C were considered. As noted in [4], a function that simply returns some concept consistent with the samples does not necessarily learn. In fact, they claim that they found examples where for every finite sample there are concepts ϵ-far from the target concept (even with $\epsilon = 1$) that are still consistent with the samples. The following is a simple example which substantiates their claim. Let $X = [0,1]$, P be the uniform distribution on X, and C be the concept class containing all finite sets of points and the entire unit interval. That is,

$$C = \{\{x_1,\ldots,x_r\} : 1 \leq r < \infty \text{ and } x_i \in [0,1],\ i = 1,\ldots,r\} \cup \{[0,1]\}$$

If the target concept is $[0,1]$ then for every finite sample there are many concepts which are consistent with the sample but are ϵ-far (with $\epsilon = 1$) from $[0,1]$. Namely, any finite set of points which contains the points of the sample is a concept with this property.

3 Relationships Between Metric Entropy and the Vapnik-Chervonenkis Dimension

In the previous section we stated a result from [6] which showed that the VC dimension of a concept class characterizes learnability for every distribution. A similar result from [4] was stated which showed that the metric entropy of a concept class characterizes learnability for a fixed distribution. These two results naturally suggest that there may be some relationships between the VC dimension of a concept class and its metric entropy with respect to various distributions. This is indeed the case. In this section, we discuss some relationships explored by [4], prove a further result, and state an earlier result from [8].

The following theorem was shown in [4], and is stated as it appeared there.

Theorem 3 *Let C be a concept class of finite dimension $d \geq 1$ and let $N(\epsilon, C, P)$ be the size of a minimum ϵ-cover of C with respect to probability measure P. Then the following relations hold:*

(i) There is a distribution P such that $\lfloor \log_2 d \rfloor \leq N(\frac{1}{4}, C, P)$.

(ii) If $\epsilon < \frac{1}{2d}$ then there is a distribution P such that $2^d \leq N(\epsilon, C, P)$.

(iii) If $\epsilon < \frac{1}{2}$ then $N(\epsilon, C, P) < 1.002\,(16d/\epsilon)^{16d/\epsilon}$ for every probability measure P.

The proofs of these relations are straightforward and were given in [4]. However, some comments on each of these relations are in order.

First, a statement more general than (ii) can be made which does not depend on the VC dimension of C. Specifically, let $x_1,\ldots,x_n \in X$ be distinct points and let $c_1,\ldots,c_k \in C$ be concepts whose intersection with $\{x_1,\ldots,x_n\}$ gives rise to distinct subsets, i.e. $c_i \cap \{x_1,\ldots,x_n\} \neq c_j \cap \{x_1,\ldots,x_n\}$ for $i \neq j$. Note that we must necessarily have $k \leq 2^n$. If we take P to be the uniform distribution on $\{x_1,\ldots,x_n\}$ then we obtain $N(\epsilon, C, P) \geq k$ for $\epsilon < \frac{1}{2n}$. This reduces to (ii) if C has VC dimension d and $c_1,\ldots,c_{2^d}$ are concepts which shatter the set of points $\{x_1,\ldots,x_d\}$. However, our statement is more general since, regardless of the VC dimension of C, it may be possible to find n concepts which give rise to n distinct subsets of $\{x_1,\ldots,x_n\}$ so that $N(\epsilon, C, P) \geq n$ for $\epsilon < \frac{1}{2n}$.

The result in (iii) was obtained somewhat indirectly in [4] by using upper and lower bounds for the number of samples required for learning (from [6] and [4] respectively). The following result along the lines of (iii) was shown in [8]. Note that the bound does not appear exactly as in [8] since the definition of VC dimension used in [8] corresponds to $d+1$.

Proposition 2 *If C is a concept class with VC dimension d, then there is a constant $K = K(d)$ such that for $0 < \epsilon \leq \frac{1}{2}$ we have $N(\epsilon, C, P) \leq K(d)\epsilon^{-(d+1)}|\ln \epsilon|^{d+1}$ for every probability measure P.*

For a fixed concept class (and hence fixed d), this bound provides a much tighter bound on $N(\epsilon, C, P)$ as a function of ϵ than the bound of (iii) (namely, polynomial versus exponential in $\frac{1}{\epsilon}$).

Now, regarding relation (i), we note that if the VC dimension of C is infinite then we can find a sequence of distributions P_n for $n = 1, 2, \ldots$ such that $\lim_{n\to\infty} N(\frac{1}{4}, C, P_n) = \infty$. Relation (i) is proved by considering the uniform distribution on a finite set of d points shattered by C. If the VC dimension of C is infinite, our comment follows by taking P_n to be the uniform distribution over n points shattered by C and using (the proof of) relation (i) for each $n = 1, 2, \ldots$. In general, for a concept class of infinite VC dimension, we may not necessarily be able to find a particular distribution P for which $N(\epsilon, C, P) = \infty$, but will only be able to approach infinite metric entropy by a sequence of distributions. However, in some cases we can achieve infinite metric entropy as shown by the following example. Let $X = [0, 1]$ and let C be the set of all Borel sets. Then taking P to be the uniform distribution, we have $N(\frac{1}{4}, C, P) = \infty$ since the infinite collection of sets corresponding to the Haar basis functions (i.e., $c_n = \{x \in [0,1] :$ *the* nth *digit in the binary expansion of x is* $1\}$) are pairwise a distance $\frac{1}{2}$ apart with respect to P.

Finally, we prove a result which has a larger range of applicability than (ii) and gives a stronger dependence on d than (i) for $\epsilon < \frac{1}{4}$. Although the bound of (ii) is exponential in d, it is valid only for $\epsilon < \frac{1}{2d}$, so that the range of applicability goes to zero as $d \longrightarrow \infty$. On the other hand, (i) is valid for a fixed ϵ independent of d (namely $\epsilon = \frac{1}{4}$) but gives only logarithmic dependence on d. The following bound gives exponential dependence on d for a fixed range of applicability ($\epsilon < \frac{1}{4}$).

Proposition 3 *If C is a concept class of finite dimension $d \geq 1$ then there is a probability measure P such that*

$$e^{2(\frac{1}{2}-2\epsilon)^2 d} \leq N(\epsilon, C, P)$$

for all $\epsilon \leq \frac{1}{4}$.

Proof: Let $\{x_1, \ldots, x_d\}$ be a set of d points that is shattered by C, and let p be the uniform distribution on $\{x_1, \ldots, x_d\}$, i.e. $P(x_i) = \frac{1}{d}$ for $i = 1, \ldots, d$. For this distribution, the only relevant property of a concept c is the set of x_i which are contained in c. Hence, we can represent c by a d bit binary string with a one in position i indicating that $x_i \in c$, and we can identify the concept class C with the set of all d bit binary strings.

If we can find n concepts that are pairwise more than 2ϵ apart, then $N(\epsilon, C, P) \geq n$ since each of the non-overlapping ϵ balls around these n concepts must contain a member of an ϵ-cover (see Lemma 3 of [4]). Given two concepts c_1, c_2 represented as binary strings, $d_P(c_1, c_2) = \frac{k}{d}$ where k is the number of bits on which c_1 and c_2 differ, and so $d_P(c_1, c_2) \leq 2\epsilon$ iff c_2 differs from c_1 on $k \leq 2\epsilon d$ bits. The number of binary strings that differ on k bits from a given string is $\binom{d}{k}$. Therefore, the number of concepts that are a distance less than or equal to 2ϵ from a given concept is $\sum_{0\leq k\leq 2\epsilon d}\binom{d}{k}$. Since the total number of concepts is 2^d, we can find at least

$$2^d / \sum_{0\leq k\leq 2\epsilon d}\binom{d}{k}$$

concepts that are more than 2ϵ apart, so that

$$N(\epsilon, C, P) \geq 2^d / \sum_{0\leq k\leq 2\epsilon d}\binom{d}{k}$$

Now, Dudley [8] states the Chernoff-Okamoto inequality

$$\sum_{0\leq k\leq m}\binom{n}{k}p^k(1-p)^{n-k}\leq e^{-(np-m)^2/[2np(1-p)]}$$

for $p\leq\frac{1}{2}$ and $m\leq np$, which can be obtained from a more general inequality (for sums of bounded random variables) of Hoeffding [13]. Taking $n=d$, $p=\frac{1}{2}$, and $m=2\epsilon d$ we obtain

$$\sum_{0\leq k\leq 2\epsilon d}\binom{d}{k}\leq 2^d e^{-2(\frac{1}{2}-2\epsilon)^2 d}$$

for $\epsilon\leq\frac{1}{4}$. Using this in our earlier bound on $N(\epsilon,C,P)$, we get

$$N(\epsilon,C,P)\geq e^{2(\frac{1}{2}-2\epsilon)^2 d}$$

for $\epsilon\leq\frac{1}{4}$ which is the desired inequality. ▮

4 Partial Results on Learnability for a Class of Distributions

In this section we prove some partial results regarding learnability for a class of distributions. The definition of learnability in this case is completely analagous to the definitions given earlier, but for completeness we state it formally.

Definition 5 (Learnability for a Class of Distributions) *Let $\mathcal{P}$ be a fixed and known collection of probability measures. The pair (C,H) is said to be* learnable with respect to $\mathcal{P}$ *if there exists a function $f\in F_{CH}$ such that for every $\epsilon,\delta>0$ there is a $0<m<\infty$ such that for every probability measure $P\in\mathcal{P}$ and every $c\in C$, if $\overline{x}\in X^m$ is chosen at random according to P^m then the probability that $error_{f,c,P}(\overline{x})<\epsilon$ is greater than $1-\delta$.*

Benedek and Itai [4] posed the problem of characterizing learnability for a class of distributions as an open problem, and they made the following conjecture.

Conjecture 1 *A concept class C is learnable with respect to a class of distributions $\mathcal{P}$ iff for every $\epsilon>0$,*

$$N(\epsilon,C,\mathcal{P})\equiv\sup_{P\in\mathcal{P}}N(\epsilon,C,P)<\infty$$

The notation defined in the statement of the conjecture will be used throughout. Namely, if $\mathcal{P}$ is any class of distributions, then $N(\epsilon,C,\mathcal{P})$ is defined by $N(\epsilon,C,\mathcal{P})=\sup_{P\in\mathcal{P}}N(\epsilon,C,P)$.

For a single distribution, the conjecture reduces immediately to the known result of [4] (stated in Section 2). For every distribution, the results of Section 3 imply that the condition $\sup_{all\ P}N(\epsilon,C,P)<\infty\ \forall\epsilon>0$ is equivalent to the condition that C have finite VC dimension. Hence, the conjecture in this case reduces to the known result of [6] (stated in Section 2. As pointed out in [4], the case where $\mathcal{P}$ is finite is similar to the case of a single distribution, and the case where $\mathcal{P}$ contains all discrete distributions is similar to the case of all distributions. The result for all discrete distributions follows again from Section 3 since $\sup_{discrete\ P}N(\epsilon,C,P)<\infty\ \forall\epsilon>0$ iff the VC dimension of C is finite.

We now prove some results for more general classes of distributions. Although our results are far from verifying the conjecture completely, the partial results we obtain are consistent with it.

One natural extension to considering a single distribution P_0 is to consider the class of all distributions sufficiently close to P_0. One measure of proximity of distributions is the total variation defined as follows. First, we assume that we are working with some fixed σ-algebra $\mathcal{S}$ of X. Let $\mathcal{P}^*$ denote the set of all probability measures defined on $\mathcal{S}$. For $P_1, P_2 \in \mathcal{P}^*$, the *total variation* between P_1 and P_2 is defined as

$$||P_1 - P_2|| = \sup_{A \in S} |P_1(A) - P_2(A)|$$

For a given distribution P_0 and $0 \leq \lambda \leq 1$ define

$$\mathcal{P}_v(P_0, \lambda) = \{P \in \mathcal{P}^* : ||P - P_0|| \leq \lambda\}$$

$\mathcal{P}_v(P_0, \lambda)$ represents all probability measures which are within λ of P_0 in total variation. For $\lambda = 0$, $\mathcal{P}_v(P_0, 0)$ contains only the distribution P_0, and for $\lambda = 1$, $\mathcal{P}_v(P_0, 1)$ contains all distributions.

Another possibility for generating a class of distributions from P_0 utilizes the property that a convex combination of two probability measures is also a probability measure. Specifically, if P_1 and P_2 are probability measures then $\lambda P_1 + (1 - \lambda)P_2$ is also a probability measure for $0 \leq \lambda \leq 1$. One interpretation of this convex combination is that with probability λ a point is drawn according to P_1, and with probability $1 - \lambda$ the point is drawn according to P_2. Given a distribution P_0 and $0 \leq \lambda \leq 1$, define

$$\mathcal{P}_l(P_0, \lambda) = \{(1 - \eta)P_0 + \eta P : \eta \leq \lambda, P \in \mathcal{P}^*\}$$

The distributions in $\mathcal{P}_l(P_0, \lambda)$ can be thought of as those obtained by using P_0 with probability greater than or equal to $1 - \lambda$ and using an arbitrary distribution otherwise. Note that, as with $\mathcal{P}_v(P_0, \lambda)$, we have $\mathcal{P}_l(P_0, 0) = \{P_0\}$ and $\mathcal{P}_l(P_0, 1) = \mathcal{P}^*$.

Both $\mathcal{P}_l(P_0, \lambda)$ and $\mathcal{P}_v(P_0, \lambda)$ can be thought as 'spheres' of distributions centered at P_0, i.e. all distributions sufficiently 'close' to P_0 in an appropriate sense. The following proposition verifies the conjecture for $\mathcal{P}_l(P_0, \lambda)$ and $\mathcal{P}_v(P_0, \lambda)$ and shows that a concept class is learnable for $\mathcal{P}_l(P_0, \lambda)$ or $\mathcal{P}_v(P_0, \lambda)$ with $\lambda > 0$ iff it is learnable for all distributions.

Proposition 4 *Let C be a concept class, P_0 a fixed distribution, and $0 < \lambda \leq 1$. Then the following are equivalent:*

(i) $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) < \infty$ for all $\epsilon > 0$

(ii) C has finite VC dimension

(iii) C is learnable for $\mathcal{P}_l(P_0, \lambda)$

Furthermore, $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}_v(P_0, \lambda)$ so that the above are equivalent for $\mathcal{P}_v(P_0, \lambda)$ as well.

Proof: (ii) $\Rightarrow$ (iii) This follows from the results of [6] (what we have called Theorem 1). Namely, (ii) implies learnability for all distributions which implies learnability for $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}^*$.

(iii) $\Rightarrow$ (i) If $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) = \infty$ for some $\epsilon > 0$, then for every $M < \infty$ there exists $P_M \in \mathcal{P}_l(P_0, \lambda)$ such that $N(\epsilon, C, P_M) > M$. But then by the results of [4] (what we have called Theorem 2), more than $\log_2 N(\epsilon, C, P_M) \geq \log_2(1 - \delta)M$ samples are required to learn for P_M. Since M is arbitrary, letting $M \to \infty$ contradicts the fact that C is learnable for $\mathcal{P}_l(P_0, \lambda)$. Thus, $N(\epsilon, C, \mathcal{P}_l(P_0, \lambda)) < \infty$ for all $\epsilon > 0$.

(i) $\Rightarrow$ (ii) For every $P \in \mathcal{P}^*$, let $Q = (1 - \lambda)P_0 + \lambda P \in \mathcal{P}_l(P_0, \lambda)$. If $c_1, c_2 \subseteq X$ are any measurable sets, then

$$\begin{aligned} d_Q(c_1, c_2) &= Q(c_1 \Delta c_2) = (1 - \lambda)P_0(c_1 \Delta c_2) + \lambda P(c_1 \Delta c_2) \\ &\geq \lambda P(c_1 \Delta c_2) = \lambda d_P(c_1 \Delta c_2) \end{aligned}$$

Therefore, $N(\lambda\epsilon, C, Q) \geq N(\epsilon, C, P)$ and so

$$\begin{aligned} N(\epsilon, C, \mathcal{P}^*) &= \sup_{P \in \mathcal{P}^*} N(\epsilon, C, P) \leq \sup_{P \in \mathcal{P}^*} N(\lambda\epsilon, C, (1-\lambda)P_0 + \lambda P) \\ &= \sup_{Q \in \mathcal{P}_l(P_0,\lambda)} N(\lambda\epsilon, C, Q) < \infty \end{aligned}$$

Hence, from the results of Section 3, C has finite VC dimension.

Finally, to show $\mathcal{P}_l(P_0, \lambda) \subseteq \mathcal{P}_v(P_0, \lambda)$, let $Q \in \mathcal{P}_l(P_0, \lambda)$. Then $Q = (1-\eta)P_0 + \eta P$ for some $P \in \mathcal{P}^*$ and $\eta \leq \lambda$. For every $A \in \mathcal{S}$, we have

$$\begin{aligned} |Q(A) - P_0(A)| &= |(1-\eta)P_0(A) + \eta P(A) - P_0(A)| \\ &= \eta |P(A) - P_0(A)| \leq \eta \leq \lambda \end{aligned}$$

Therefore, $||Q - P_0|| \leq \lambda$ so that $Q \in \mathcal{P}_v(P_0, \lambda)$. ∎

The following result shows that learnability of a concept class is retained under finite unions of distribution classes. That is, if a concept class C is learnable for a finite number of sets of distributions $\mathcal{P}_1, \ldots, \mathcal{P}_n$ then it is learnable with respect to their union $\mathcal{P} = \cup_{i=1}^n \mathcal{P}_i$. This is to be expected if the conjecture is true since $N(\epsilon, C, \mathcal{P}) = \max_i N(\epsilon, C, \mathcal{P}_i) < \infty$ iff $N(\epsilon, C, \mathcal{P}_i) < \infty$ for $i = 1, \ldots, n$.

Proposition 5 *Let C be a concept class, and let $\mathcal{P}_1, \ldots, \mathcal{P}_n$ be n sets of distributions. If C is learnable with respect to $\mathcal{P}_i$ for $i = 1, \ldots, n$ then C is learnable with respect $\cup_{i=1}^n \mathcal{P}_i$.*

Proof: Let f_i be an algorithm which learns C with respect to $\mathcal{P}_i$, and let $m_i(\epsilon, \delta)$ be the number of samples required by f_i to learn with accuracy ϵ and confidence δ. Define an algorithm f as follows. Ask for

$$m(\epsilon, \delta) = \max_{1 \leq i \leq n} m_i(\frac{\epsilon}{2}, \frac{\delta}{2}) + \frac{32}{\epsilon} \ln \frac{n}{\delta/2}$$

samples. Using the first $\max_i m_i(\frac{\epsilon}{2}, \frac{\delta}{2})$ samples, form hypotheses $h_1, \ldots, h_n$ using algorithms $f_1, \ldots, f_n$ respectively. Then, using the last $\frac{32}{\epsilon} \ln \frac{n}{\delta/2}$ samples, let f output the hypothesis h_i which is inconsistent with the fewest number of this second group of samples. We claim that f is a learning algorithm for C with respect to $\cup_{i=1}^n \mathcal{P}_i$.

Let $P \in \cup_{i=1}^n P_i$, and let $c \in C$. Then $P \in \mathcal{P}_k$ for some k. Since the f_i are learning algorithms with respect to the $\mathcal{P}_i$, at least one h_i (namely h_k) is within $\frac{\epsilon}{2}$ of c with probability (with respect to product measures of P) greater than $1 - \frac{\delta}{2}$. Given that h_i is within $\frac{\epsilon}{2}$ of c for some i, the proof of Lemma 4 from [4] shows that the most consistent hypothesis (on the second group of samples) is within ϵ of c with probability greater than $1 - \frac{\delta}{2}$. Therefore, if A denotes the event that at least one h_i is within $\frac{\epsilon}{2}$ of c then

$$\begin{aligned} Pr\{d_P(f(sam_c(\overline{x})), c) < \epsilon\} &= Pr\{d_P(f(sam_c(\overline{x})), c) < \epsilon \mid A\} \cdot Pr\{A\} \\ &\geq (1 - \frac{\delta}{2})(1 - \frac{\delta}{2}) > 1 - \delta \end{aligned}$$

Thus, f is a learning algorithm for C with respect to $\cup_{i=1}^n \mathcal{P}_i$ using $m(\epsilon, \delta)$ samples. ∎

Note that the above result is not true in general for an infinite number of classes of distributions since the sample complexity of the corresponding algorithms may be unbounded (i.e. we may have $\sup_i N(\epsilon, C, \mathcal{P}_i) = \infty$). However, even if $N(\epsilon, C, \mathcal{P}_i)$ is uniformly bounded the proof above does not go through since the application of Lemma 4 from [4] requires finitely many hypotheses. This is essentially the difficulty encountered in attempting to prove the conjecture directly.

For a finite number of distributions $P_1, \ldots, P_n$, define their *convex hull*, denoted by $conv(P_1, \ldots, \mathcal{P}_n)$, as the set of distributions that can be written as a convex combination of $P_1, \ldots, \mathcal{P}_n$. That is,

$$conv(P_1, \ldots, P_n) = \{\lambda_1 P_1 + \cdots + \lambda_n P_n : 0 \leq \lambda_i \leq 1 \textit{ and } \lambda_1 + \cdots + \lambda_n = 1\}$$

We now prove the following proposition.

Proposition 6 *Let C be a concept class and let $P_1, \ldots, P_n$ be probability measures. The following are equivalent:*

(i) C is learnable with respect to P_i for each $i = 1, \ldots, n$.

(ii) $N(\epsilon, C, conv(P_1, \ldots, P_n)) < \infty$ for all $\epsilon > 0$.

(iii) C is learnable with respect to $conv(P_1, \ldots, P_n)$.

Proof: (iii) $\Rightarrow$ (i) This is immediate.

(i) $\Rightarrow$ (ii) Since C is learnable with respect to P_i for each i, by Theorem 2 we have $N(\epsilon, C, P_i) < \infty$ for all $\epsilon > 0$ and $i = 1, \ldots, n$. Let $N_i(\epsilon) = N(\epsilon, C, P_i)$ and let $c_{i,1}, \ldots, c_{i,N_i(\epsilon/2)}$ be an $\frac{\epsilon}{2}$-approximation of C with respect to d_{P_i}. For each $i = 1, \ldots, n$, let $C_{i,j} = \{c \in C : d_{P_i}(c, c_{i,j}) \leq \frac{\epsilon}{2}\}$ for $j = 1, \ldots, N_i(\frac{\epsilon}{2})$. We have $C = \cup_{j=1}^{N_i(\epsilon/2)} C_{i,j}$ for all $i = 1, \ldots, n$. Let

$$C_{k_1,\ldots,k_n} = \bigcap_{i=1}^{n} C_{i,k_i}$$

for $1 \leq k_i \leq N_i(\frac{\epsilon}{2})$, $i = 1, \ldots, n$. Clearly,

$$C = \bigcup_{all\ (k_1,\ldots,k_n)} C_{k_1,\ldots,k_n}$$

Also, by construction the 'diameter' of each $C_{k_1,\ldots,k_n}$ with respect to d_{P_i} is less than or equal to ϵ for all $i = 1, \ldots, n$, i.e. for each $i = 1, \ldots, n$ we have

$$\sup_{c_1,c_2 \in C_{k_1,\ldots,k_n}} d_{P_i}(c_1, c_2) \leq \epsilon$$

Hence, if we define a metric $\rho(\cdot, \cdot)$ by

$$\rho(c_1, c_2) = \max_{1 \leq i \leq n} d_{P_i}(c_1, c_2)$$

then $N(\epsilon, C, \rho) \leq \prod_{i=1}^{n} N_i(\frac{\epsilon}{2}) < \infty$ since we can form an ϵ-approximation of C with respect to ρ by simply taking any point from each $C_{k_1,\ldots,k_n}$ that is nonempty.

Now, if $Q \in conv(P_1, \ldots, P_n)$ then $Q = \sum_{i=1}^{n} \lambda_i P_i$ for some $0 \leq \lambda_i \leq 1$ with $\sum_{i=1}^{n} \lambda_i = 1$. For any measurable $c_1, c_2 \subseteq X$, we have

$$\begin{aligned} d_Q(c_1, c_2) &= \sum_{i=1}^{n} \lambda_i d_{P_i}(c_1, c_2) \\ &\leq \left(\sum_{i=1}^{n} \lambda_i\right) \max_{1 \leq i \leq n} d_{P_i}(c_1, c_2) = \rho(c_1, c_2) \end{aligned}$$

so that $N(\epsilon, C, Q) \leq N(\epsilon, C, \rho)$. Thus,

$$N(\epsilon, C, conv(P_1, \ldots, P_n)) = \sup_{Q \in conv(P_1, \ldots, P_n)} N(\epsilon, C, Q) \leq \prod_{i=1}^{n} N_i(\frac{\epsilon}{2}) < \infty$$

(ii) $\Rightarrow$ (iii) If $N(\epsilon, C, conv(P_1, \ldots, P_n)) < \infty$ for all $\epsilon > 0$, then, in particular, $N(\epsilon, C, P_i) < \infty$ for $i = 1, \ldots, n$ and $\epsilon > 0$. Therefore, we can employ the construction used above in proving that (i) implies (ii) to get a finite $\frac{\epsilon}{2}$-approximation of C uniformly for all $Q \in conv(P_1, \ldots, P_n)$. As shown above, such an approximation can be found with less than or equal to $\prod_{i=1}^{n} N_i(\frac{\epsilon}{4})$ elements where $N_i(\frac{\epsilon}{4}) = N(\frac{\epsilon}{4}, C, P_i)$. Thus, using the proof of Lemma 4 from [4], the algorithm which takes $\frac{32}{\epsilon} (\ln \frac{1}{\delta} + \sum_{i=1}^{n} \ln N_i(\frac{\epsilon}{4}))$ samples and outputs an element of the $\frac{\epsilon}{2}$-approximation with the smallest number of inconsistent samples is a learning algorithm for C with respect to $conv(P_1, \ldots, P_n)$. ∎

The above proposition verifies the conjecture for classes of distributions which are 'convex polyhedra with finitely many sides' in the space of all distributions. In fact, combined with the previous proposition, the conjecture is verified for all finite unions of such polyhedra.

5 Summary

It was first pointed out that the condition for learnability with respect to a fixed distribution obtained in [4] is identical to the notion of finite metric entropy. Metric entropy has been studied elsewhere, and perhaps results from that literature may have applications to concept learning. In considering relationships between the VC dimension of a concept class and its metric entropy, we extended a result of [4] and stated an earlier result from [8]. Finally, we proved some partial results concerning learnability with respect to a class of distributions. These results are consistent with a conjecture in [4]. Specifically, it was shown that the conjecture holds for any 'sphere' of distributions and for any set of distributions which is a finite union of 'convex polyhedra with finitely many sides'. In addition to verifying the conjecture in these cases, the results indicate some limitations of attempting to enlarge the set of learnable concept classes by requiring learnability only for a class of distributions as opposed to all distributions.

In closing, we briefly mention some other work that has been done on Valiant's learning framework. (Note that this is not intended to be a complete survey.) A considerable amount of work has been done on studying specific learnable concept classes taking into consideration issues of computational difficulty. In fact, much of [23] focused on certain special classes of Boolean functions (see also [15, 20, 24]). Several papers have dealt with the interesting issue of noise in the samples [2, 14, 21, 24]. A result concerning noisy samples was also given in [4] for the case of learnability with respect to a fixed distribution. Another interesting idea involves the introduction of a measure of the complexity of concepts, and allowing the number of samples to depend on this complexity. This has been studied in [6, 7, 5, 10, 18]. We stated our definitions of learnability in terms of both the concept class C and the hypothesis class H, but assumed throughout that $H \supseteq C$. Considerations in the more general case have been discussed in [1, 3, 18]. The use of more powerful oracles (i.e., protocols which allow the learner to get information other than just random samples) have been considered in [1, 23] Finally, [19] has considered learnability of continuous valued functions (as opposed to the usual binary valued functions).

Acknowledgments

I would like to thank Prof. Sanjoy Mitter for introducing me to this subject and Avi Lele, Prof. Sanjoy Mitter, and Prof. Jack Koplowitz for many valuable discussions.

References

[1] Amsterdam, J. Extending the Valiant learning model. *Proc. 5th Int. Conf. on Machine Learning*, pp. 381-394, 1988.

[2] Angluin, D. and P. Laird. Learning from noisy examples. *Machine Learning* 2, pp.343-370, 1988.

[3] Baum, E.B. Complete representations for learning from examples. From *Complexity in Information Theory*, Y.S. Abu-Mostafa, edt., pp. 77-98, Springer-Verlag, 1988.

[4] Benedek, G.M. and A. Itai. Learnability by fixed distributions. *Proc. of First Workshop on Computational Learning Theory*, pp. 80-90, 1988.

[5] Benedek, G.M. and A. Itai. Nonuniform learnability. *ICALP*, pp. 82-92, 1988.

[6] Blumer, A., A. Ehrenfeucht, D. Haussler, M. Warmuth. Classifying learnable geometric concepts with the Vapnik-Chervonenkis dimension. *Proc. 18th ACM Symp. on Theory of Comp.*, Berkeley, CA, pp. 273-282, 1986.

[7] Blumer, A., A. Ehrenfeucht, D. Haussler, M. Warmuth. Occam's razor. *Info. Proc. Let.* 24, pp. 377-380, 1987. pp. 273-282, 1986.

[8] Dudley, R.M. Central limit theorems for empirical measures. *Ann. Probability* 6(6), pp. 899-929, 1978.

[9] Dudley, R.M. Metric entropy of some classes of sets with differentiable boundaries. *J. Approx. Theory* 10(3), pp. 227-236, 1974.

[10] Haussler, D., M. Kearns, N. Littlestone, M.K. Warmuth. Equivalence of models for polynomial learnability. *Proc. First Workshop on Computational Learning Theory*, pp. 42-55, 1988.

[11] Haussler, D. and E. Welzl. ϵ-Nets and simplex range queries. *Discrete and Comput. Geom.* 2, pp. 127-151, 1987.

[12] Hawkes, J. Hausdorff measure, entropy, and the independence of small sets. *Proc. London Math. Soc.* (3)28, pp. 700-724, 1974.

[13] Hoeffding, W. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* 58, pp. 13-30, 1963.

[14] Kearns, M. and M. Li. Learning in the presence of malicious errors. *Proc. 20th ACM Symp. on Theory of Comp.*, Chicago, Illinois, pp. 267-279, 1988.

[15] Kearns, M., M. Li, L. Pitt, L. Valiant. On the learnability of Boolean formulae. *Proc. 19th ACM Symp. on Theory of Comp.*, New York, New York, pp. 285-295, 1987.

[16] Kolmogorov, A.N. and V.M. Tihomirov, ϵ-Entropy and ϵ-capacity of sets in functional spaces. *Amer. Math. Soc. Transl.* 17, pp. 277-364, 1961.

[17] Koplowitz, J. On the entropy and reconstruction of digitized planar curves. IEEE Int. Symp. on Info. Theory, Brighton, England, June, 1985.

[18] Linial, N., Y. Mansour, R.L. Rivest. Results on learnability and the Vapnik-Chervonenkis dimension. *Proc. First Workshop on Computational Learning Theory*, pp. 56-68, 1988.

[19] Natarajan, B.K. and P.T. Tadepalli. Two new frameworks for learning. *Proc. 5th Int. Conf. on Machine Learning*, pp. 402-415, 1988.

[20] Rivest, R.L. Learning decision lists. *Machine Learning* 2(3), pp. 229-246, 1987.

[21] Sloan, R. Types of noise in data for concept learning. *Proc. First Workshop on Computational Learning Theory*, pp. 91-96, 1988.

[22] Tikhomirov, V.M. Kolmogorov's work on ϵ-entropy of functional classes and the superposition of functions. *Russian Math. Surveys*, vol. k8, pp. 51-75, 1963.

[23] Valiant, L.G. A theory of the learnable. *Comm. ACM* 27(11), pp. 1134-1142, 1984.

[24] Valiant, L.G. Learning disjunctions of conjunctions. *Proc. IJCAI*, pp. 560-566, 1985.

[25] Vapnik, V.N. and A.Ya. Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. *Theory of Prob. and its Appl.* 16(2), pp. 264-280, 1971.

[26] Wenocur, R.S. and R.M. Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Math.* 33, pp. 313-318, 1981.