A Metric Entropy Bound is Not Sufficient for Learnability

R. M. Dudley[1], S. Kulkarni[2], T. Richardson[3], and O. Zeitouni[4]

Submitted: Oct. 16, 1992
Revised: Sept. 17, 1993

Abstract

We prove by means of a counterexample that it is not sufficient, for PAC learning under a class of distributions, to have a uniform bound on the metric entropy of the class of concepts to be learned. This settles a conjecture of Benedek and Itai.

Key Words: learning, estimation, PAC, metric entropy, class of distributions

1 Introduction

Let $(\mathcal{X}, \mathcal{B})$ be a measurable space. Let $\mathcal{P}$ be a class of probability measures on $(\mathcal{X}, \mathcal{B})$. Let $\mathcal{C}$ (the "concept class" in the language of learning theory, as introduced in [6]) be a subset of $\mathcal{B}$. Suppose one is given a sequence of i.i.d., $\mathcal{X}$ valued random variables $X_1, \ldots, X_n$ distributed according to P^n, where $P \in \mathcal{P}$. In addition, for some unknown $c \in \mathcal{C}$, one is given data $(X_1, I_c(X_1)), \ldots, (X_n, I_c(X_n))$ which we henceforth denote by $\mathcal{D}_n(c)$. The problem of learning consists roughly of the question "given $\mathcal{C}, \mathcal{P}$, how large should n be for approximating c with high accuracy and low probability of error based on the data $\mathcal{D}_n(c)$?" In mathematical terms,

[1]Dept. of Mathematics, M.I.T., Cambridge, MA 02139. Research partially supported by National Science Foundation grants.

[2]Dept. of Electrical Engineering, Princeton University, Princeton, NJ 08544. The work of this author was supported in part by the Army Research Office under grant DAAL03-92-G-0320 and by the National Science Foundation under grant IRI-92-09577.

[3]AT&T Bell Labs, 600 Mountain Av., Murray Hill, NJ 07974.

[4]Dept. of Electrical Engineering, Technion, Haifa 32000, Israel. The work of this author was done while visiting the Center for Intelligent Control Systems at M.I.T. under support of the U.S. Army Research Office grant DAAL03-92-G-0115.

assume that $(\mathcal{X}, \mathcal{B})$ is a Borel space, and define on $\mathcal{B}$ the pseudo metric $d_P(c_1, c_2) = P(c_1 \triangle c_2)$. Let $\mathcal{T}$ be the algebra of all four subsets of $\{0,1\}$. A learning rule is a map $T^n : (\mathcal{X} \times \{0,1\})^n \to \mathcal{C}$ such that, for any $c \in \mathcal{C}$, any $P \in \mathcal{P}$, and any $\epsilon > 0$,

$$\{(X_1, \dots, X_n, i_1, ..., i_n) : \; d_P(c, T^n((X_1, i_1), \dots, (X_n, i_n))) > \epsilon\} \; \in \; \mathcal{B}^n \; \otimes \mathcal{T}^n . \tag{1}$$

It follows that for any $c, d \in \mathcal{C}$,

$$\{(X_1, \dots, X_n) : \; d_P(d, T^n(\mathcal{D}_n(c))) > \epsilon\} \in \mathcal{B}^n . \tag{2}$$

We say that the concept class $\mathcal{C}$ is PAC learnable under the class of probability measures $\mathcal{P}$ (in short: $\mathcal{C}$ is PAC learnable under $\mathcal{P}$) if, for every $\epsilon > 0$, $\delta > 0$, there exist an integer $n = n(\mathcal{P}, \mathcal{C}, \epsilon, \delta)$ and a learning rule T^n such that, for any $P \in \mathcal{P}$ and $c \in \mathcal{C}$,

$$P^n(\{(X_1, \dots, X_n) : \; d_P(c, T^n(\mathcal{D}_n(c))) > \epsilon\}) < \delta \, . \tag{3}$$

The notion of learnability in the form (3) has recently received much attention (e.g., see [1, 4, 6]), and in the learning literature is referred to as Probably Approximately Correct (PAC) learning, for reasons obvious from its definition. Intuitively, in PAC learning one attempts to achieve a good prediction on future samples, after seeing some finite number of samples, uniformly in $P \in \mathcal{P}$ and $c \in \mathcal{C}$.

Sufficient and necessary conditions for PAC learnability are by now well known for some cases. Let $B(c, \epsilon) = \{\tilde{c} \in \mathcal{B} : \; d_P(c, \tilde{c}) < \epsilon\}$, and define the *$\epsilon$-covering number* of $\mathcal{C}$ with respect to P by

$$N(\epsilon, \mathcal{C}, P) = \inf\{N : \; \exists c_1, \dots, c_N \in \mathcal{B} \; \text{ such that } \; \mathcal{C} \subset \cup_{i=1}^N B(c_i, \epsilon)\} \, .$$

The balls $B(c_i, \epsilon)$ above are said to form an *ϵ-cover* of $\mathcal{C}$, and $\log N(\epsilon, \mathcal{C}, P)$ is often referred to as the *metric entropy* of $\mathcal{C}$ with respect to P. A necessary and sufficient condition for PAC learnability of $\mathcal{C}$ in the special case where $\mathcal{P}$ is a singleton, namely $\mathcal{P} \equiv \{P\}$, is that $N(\epsilon, \mathcal{C}, P) < \infty$ for all $\epsilon > 0$ (see [2] and, in greater generality, [7], pp. 149–151). Moreover, if

$\mathcal{P} = M_1(\mathcal{X})$, the space of Borel probability measures on $\mathcal{X}$, then (under suitable measurability conditions) a well known necessary and sufficient condition for PAC learnability of $\mathcal{C}$ under $\mathcal{P}$ is that the VC dimension of $\mathcal{C}$ be finite, which turns out to be equivalent to the condition that, for all $\epsilon > 0$, $\sup_{P \in M_1(\mathcal{X})} N(\epsilon, \mathcal{C}, P) < \infty$ (see [1, 3, 4, 7, 8, 9] for proofs and additional background on the VC dimension and metric entropy). The similarity between these two extreme cases led Benedek and Itai to conjecture in [2] that the condition

$$\forall \epsilon > 0, \ \sup_{P \in \mathcal{P}} N(\epsilon, \mathcal{C}, P) < \infty \tag{4}$$

is necessary and sufficient for the PAC learnability of $\mathcal{C}$ under $\mathcal{P}$. While necessity is fairly obvious, the sufficiency part is less so because of the difficulty in simultaneously approximately determining $c \in \mathcal{C}$ and $P \in \mathcal{P}$. (We mention that if (4) is replaced by the stronger condition that there exists a fixed finite ϵ-cover of $\mathcal{C}$ under all $P \in \mathcal{P}$, then the sufficiency is just a standard extension of the single measure case. Some cases where (4) is sufficient are described in [5].) It is the purpose of this note to show, by a counterexample, that (4) is not sufficient in general for learnability. The question of finding a necessary and sufficient condition for PAC learnability of $\mathcal{C}$ under $\mathcal{P}$ remains open.

2 A Counterexample

Let $\Omega = \mathcal{X} = \{0,1\}^\infty$, let X^i denote the coordinate map of $X \in \mathcal{X}$, and let $\mathcal{B}$ be the Borel σ-field over $\mathcal{X}$. Let $(p_1, p_2, \ldots) \in [0,1]^\infty$ be defined by $p_i = 1/\log_2(i+1) \leq 1$, and note that for every finite n, $\sum_{i=1}^\infty p_i^n = \infty$. Identifying $p_i = P(X^i = 1)$, the vector $p_1, p_2, \ldots$ induces a product measure P_I on the product space $\mathcal{X}$. For any measure P on $\mathcal{X}$, P^n denotes the product measure on $\mathcal{X}^n$ obtained from P.

Let σ denote a permutation (possibly infinite) of the integers, i.e. $\sigma : N \to N$ is one to one and onto, and define P_σ as the measure on $\mathcal{X}$ induced by $(p_{\sigma^{-1}(1)}, p_{\sigma^{-1}(2)}, \ldots)$. The ensemble of all permutations is denoted Σ. Thus, $P_\sigma(X^{\sigma(i)} = 1) = p_i$ and, if σ is the identity map, then

P_σ equals the P_I defined above.

Now let $\mathcal{P} \equiv \{P_\sigma,\, \sigma \in \Sigma\}$, let $c_i \equiv \{X \in \mathcal{X} :\ X^i = 1\}$, and let $\mathcal{C} \equiv \{c_i,\, i \in N\}$. It is easy to check that for any $P \in \mathcal{P}$, $N(\epsilon, \mathcal{C}, P) < \infty$. Since any c_i with $p_{\sigma^{-1}(i)} < \epsilon$ satisfies $d_{P_\sigma}(c_i, \emptyset) < \epsilon$, we have that for any $P \in \mathcal{P}$,

$$N(\epsilon, \mathcal{C}, P) < 2^{1/\epsilon}\,.$$

It follows that $\sup_{P \in \mathcal{P}} N(\epsilon, \mathcal{C}, P) < \infty$. We now claim

Theorem 1 *$\mathcal{C}$ is not PAC learnable under $\mathcal{P}$.*

Proof: We use a random coding argument. Suppose that the theorem's assertion is false. Then, for each $\epsilon > 0, \delta > 0$, it is possible to find an $n = n(\epsilon, \delta)$ and a learning rule T^n which satisfy (3) for all $c \in \mathcal{C}$ and $P \in \mathcal{P}$. In particular, for any finite k, it satisfies (3) for $c \in \mathcal{C}^k$ and $P \in \mathcal{P}^k$, where $\mathcal{C}^k = \{c_i,\, i = 1, \ldots, k\}$, $\Sigma^k = \{\sigma :\ \sigma(i) = i\ \forall i > k\}$, and $\mathcal{P}^k = \{P_\sigma,\, \sigma \in \Sigma^k\}$, i.e. $\mathcal{P}^k$ are all possible permutations of the vector $(p_1, p_2, \ldots)$ which involve only the first k coordinates. Let the error event be defined as

$$\mathrm{er}_\sigma^c = \{(X_1, \ldots, X_n) :\ d_{P_\sigma}(c, T^n(\mathcal{D}_n(c))) > \epsilon\}\,.$$

(It follows from (2) that er_σ^c is a measurable event.) Then, for each $c \in \mathcal{C}^k$ and $P_\sigma \in \mathcal{P}^k$,

$$P_\sigma^n(\mathrm{er}_\sigma^c) < \delta\,.$$

In particular, if Q is any probability measure on the finite set $\{(\sigma, c) :\ \sigma \in \Sigma^k, c \in \mathcal{C}^k\}$, then

$$E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) < \delta\,. \tag{5}$$

Now choose Q such that $Q|_\Sigma$ is uniform over Σ^k while $c = c_{\sigma(1)}$ (i.e., $Q(\sigma, c) = 1/k!$ if $\sigma \in \Sigma^k$ and $c = c_{\sigma(1)}$, and $Q(\sigma, c) = 0$ otherwise). This Q forces the true concept to involve the coordinate of maximal probability (where in fact the probability is 1) in P_σ. Note that by our choice of Q, if $\epsilon < 1 - 1/\log_2(3) = \min_{j>1} d_{P_I}(c_1, c_j)$, then, when (σ, c) are distributed according to Q,

$$d_{P_\sigma}(c, \tilde{c}) < \epsilon \Rightarrow c = \tilde{c} = c_{\sigma(1)} \;\; Q \;\text{ a.s.}\;\; .$$

Thus, in this set-up, Q a.s.,

$$\mathrm{er}_\sigma^c = \{(X_1, \ldots, X_n) : \; c \neq T^n(\mathcal{D}_n(c))\}\,.$$

Using the notation σX to denote the element of $\mathcal{X}$ with coordinates $(\sigma X)^i = X^{\sigma^{-1}(i)}$ and $\sigma \mathcal{D}_n$ to denote the corresponding permutation on $\mathcal{D}_n(c)$ when $c = c_{\sigma(1)}$, i.e.,

$$\begin{aligned} \sigma \mathcal{D}_n &= ((\sigma X_1, I_{c_{\sigma(1)}}(\sigma X_1)), \ldots, (\sigma X_n, I_{c_{\sigma(1)}}(\sigma X_n))) \\ &= ((\sigma X_1, I_{c_1}(X_1)), \ldots, (\sigma X_n, I_{c_1}(X_n))), \end{aligned} \tag{6}$$

we have

$$\begin{aligned} E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) &= E_Q(P_\sigma^n(c \neq T^n(\mathcal{D}_n(c)))) \\ &= E_Q(P_\sigma^n(c_{\sigma(1)} \neq T^n(\mathcal{D}_n(c_{\sigma(1)})))) \\ &= E_Q(P_I^n(c_{\sigma(1)} \neq T^n(\sigma \mathcal{D}_n))) \\ &= E_{P_I^n} E_Q(1_{c_{\sigma(1)} \neq T^n(\sigma \mathcal{D}_n)})\,. \end{aligned} \tag{7}$$

For given vectors $\vec{x} = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $\vec{X} = (X_1, \ldots, X_n) \in \mathcal{X}^n$, denote by $S(\vec{X}, \vec{x})$ the set of permutations $\sigma \in \Sigma^k$ such that $\sigma \vec{X} = \vec{x}$. (Note that for many pairs $(\vec{X}, \vec{x})$, $S(\vec{X}, \vec{x})$ is empty.) It follows from the definition that, for $\sigma \in S(\vec{X}, \vec{x})$,

$$\sigma \mathcal{D}_n = ((x_1, I_{c(1)}(X_1), \ldots, (x_n, I_{c(1)}(X_n))).$$

By the construction of Q, the distribution of σ conditioned on $S(\vec{X}, \vec{x})$ is uniform there. Let now

$$J^{\vec{X}} = \{i \leq k : \; X_j^i = 1 \;\; \forall j = 1, \ldots, n\}$$

and

$$J^{\vec{x}} = \{i \le k : x_j^i = 1 \ \ \forall j = 1, \dots, n\}.$$

$S(\vec{X}, \vec{x})$ is non-empty only if $|J^{\vec{x}}| = |J^{\vec{X}}|$. When $\vec{X}$ has distribution P_I^n, we have $1 \in J^{\vec{X}}$ almost surely, so $|J^{\vec{X}}| \ge 1$. Let $\sigma_c \in \Sigma^k$ be a fixed permutation such that $\sigma_c(i) \in J^{\vec{x}}$ if $i \in J^{\vec{X}}$. Decompose each permutation $\sigma \in S(\vec{X}, \vec{x})$ into $\sigma = \sigma_c \circ \sigma_b \circ \sigma_a$, with $\sigma_a : J^{\vec{X}} \to J^{\vec{X}}$, and σ_a equals the identity on $\{1, \dots, k\} \setminus J^{\vec{X}}$ while $\sigma_b : \{1, \dots, k\} \setminus J^{\vec{X}} \to \{1, \dots, k\} \setminus J^{\vec{X}}$ and σ_b equals the identity on $J^{\vec{X}}$. This is always possible because all permutations in $S(\vec{X}, \vec{x})$ must satisfy $\sigma\vec{X} = \vec{x}$. Note that whenever $S(\vec{X}, \vec{x})$ is non-empty then $|\sigma_A| = |J^{\vec{X}}|!$, where

$$\sigma_A \triangleq \{\sigma_a : \sigma \in S(\vec{X}, \vec{x})\}, \quad \sigma_B \triangleq \{\sigma_b : \sigma \in S(\vec{X}, \vec{x})\}.$$

Using now (7),

$$\begin{aligned} E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) &= E_{P_I^n}\left(\sum_{\vec{x}} E_Q(1_{T^n(\sigma\mathcal{D}_n) \ne c_{\sigma(1)}} | \sigma \in S(\vec{X}, \vec{x})) Q(S(\vec{X}, \vec{x}))\right) \\ &= E_{P_I^n}\left(\sum_{\vec{x}} Q(S(\vec{X}, \vec{x})) \frac{\sum_{\sigma_b \in \sigma_B} \sum_{\sigma_a \in \sigma_A} 1_{T^n(\sigma\mathcal{D}_n) \ne c_{\sigma(1)}}}{\sum_{\sigma_b \in \sigma_B} \sum_{\sigma_a \in \sigma_A} 1}\right), \end{aligned} \tag{8}$$

where in the last equality we have used the uniformity of the conditional distribution over $S(\vec{X}, \vec{x})$, and the sum over $\vec{x}$ is taken over all *different* vectors in $\mathcal{X}^n$. By (6), $\sigma\mathcal{D}_n$ is constant for $\sigma \in S(\vec{X}, \vec{x})$, so

$$T^n(\sigma\mathcal{D}_n) = c_T$$

for some $c_T = c_T(\vec{X}, \vec{x}) \in \mathcal{C}$ not depending on $\sigma \in S(\vec{X}, \vec{x})$. Here $c_T(\cdot, \cdot)$ is measurable by (2). Thus, since the number of permutations $\sigma \in \sigma_A$ for which $T^n(\sigma\mathcal{D}_n) = c_T$ is at most equal to the number of permutations in σ_A which have a prescribed index in $J^{\vec{X}}$ unchanged,

$$\sum_{\sigma_a \in \sigma_A} 1_{T^n(\sigma\mathcal{D}_n) \ne c_{\sigma(1)}} \ge (|J^{\vec{X}}| - 1)(|J^{\vec{X}}| - 1)!$$

whereas

$$\sum_{\sigma_a \in \sigma_A} 1 = |J^{\vec{X}}|!.$$

It follows that, for any $\eta > 1$,

$$E_Q(P_\sigma^n(\mathrm{er}_\sigma^c)) \geq E_{P_I^n} \frac{(|J^{\vec{X}}|-1)(|J^{\vec{X}}|-1)!}{|J^{\vec{X}}|!} = (1 - E_{P_I^n} \frac{1}{|J^{\vec{X}}|}) \geq (1 - \frac{1}{\eta} - P_I^n(|J^{\vec{X}}| \leq \eta)) \,.$$

It remains therefore only to show that $|J^{\vec{X}}|$ may, with high probability, be made arbitrarily large by choosing a k large enough. But this is obvious because, by the Borel-Cantelli lemma, using $\vec{X}^i \triangleq (X_1^i, \ldots, X_n^i)$,

$$P_I^n(\vec{X}^i = (1, \ldots, 1) \text{ infinitely often}) = 1$$

since $\sum_{i=1}^\infty P_I^n(\vec{X}^i = (1, \ldots, 1)) \geq \sum_{i=1}^\infty p_i^n = \infty$. Thus, for any η, one may find a k large enough such that $P_I^n(|J^{\vec{X}}| \leq \eta)$ is arbitrarily small. $\square$

Remark: Note that we have actually shown that, for any fixed n and any $\epsilon < 1 - 1/\log_2(3)$, one may construct a $\mathcal{P}$ and a $\mathcal{C}$ such that the probability of error is arbitrarily close to 1. By defining $p_i, \; i \geq 2$ to be smaller, we could also take any $\epsilon < 1$.

References

[1] A. Blumer, A. Ehrenfeucht, D. Haussler, M. Warmuth, "Learnability and the Vapnik-Chervonenkis dimension," *J. ACM*, Vol. 36, No. 4, pp. 929-965, 1989.

[2] G.M. Benedek and A. Itai, "Learnability with respect to a fixed distribution," *Theoretical Computer Science*, Vol. 86, pp. 377-389, 1991.

[3] R.M. Dudley, "A Course on empirical processes", *École d'été de probabilités de St.-Flour*, 1982, *Lecture Notes in Math.* Vol. 1097, 1984, Springer, New York, 1-142.

[4] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Information and Computation*, Vol 20, pp. 78–150, 1992.

[5] S.R. Kulkarni, "Problems of computational and information complexity in machine vision and learning," Ph.D. thesis, Dept. of Electrical Engineering and Computer Science, M.I.T., June, 1991.

[6] L.G. Valiant, "A theory of the learnable," *Comm. ACM*, Vol. 27, No. 11, pp. 1134-1142, 1984.

[7] V.N. Vapnik, *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

[8] V.N. Vapnik and A.Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Prob. and its Appl.*, Vol. 16, No. 2, pp. 264-280, 1971.

[9] V.N. Vapnik and A.Ya. Chervonenkis, "Necessary and sufficient conditions for the uniform convergence of means to their expectations," *Theory of Prob. and its Appl.*, Vol. 26, No. 3, pp. 532-553, 1981.