Note

Learnability with respect to fixed distributions

Gyora M. Benedek* and Alon Itai**
Computer Science Department, Technion, Haifa, Israel

Communicated by M.S. Paterson
Received February 1989
Revised September 1990

Abstract

Benedek, G.M., and A. Itai, Learnability with respect to fixed distributions (Note), Theoretical Computer Science 86 (1991) 377-389.

Valiant's protocol for learning is extended to the case where the distribution of the examples is known to the learner. Namely, the notion of a concept class C being learnable with respect to distribution D is defined and the learnable pairs (C, D) of concept classes C and distributions D are characterized. Another notion is the existence of a finite cover for C with respect to D. The main result is that C is learnable with respect to D if and only if C is finitely coverable with respect to D. The size of the cover is then related to the Vapnik-Chervonenkis dimension.

An additional property of the learning method is robustness, i.e., learning succeeds even if part of the input is erroneous. It is also shown that if D is discrete then every concept class is learnable with respect to D. The main concern of the paper is the number of examples sufficient to probabilistically identify (or approximate) a concept — not the time needed to compute it. Indeed, in some cases the function which associates a sample with a hypothesis is undecidable, and even if it is computable, the computation may be infeasible. The computational complexity of the algorithms used for learning are considered only for discrete distributions.

1. Introduction

In his seminal paper, Valiant [15] provided a complexity-theoretic basis for learning boolean formulae. He defined learnability by examples produced by an *arbitrary distribution.* In this paper we consider a theory for learnability for *particular*

* Current address: Rosh Intelligent Systems Ltd., P.O. Box 03552, Mevasseret Zion, 90805, Israel.

** Partially supported by the Technion VPR Fund number 120-0689 and by the Foundation for Research in Electronic, Computers and Communications (administered by the Israeli Academy of Sciences and Humanities).

distributions. We characterize those distributions and concept classes which are learnable. In the above paper, Valiant gives two reasons for requiring that learning be possible for *all* distributions:

(1) to prevent coding the answer by a clever choice of examples;

(2) the distribution of the examples is unknown.

However, in many cases the distribution of examples is known and that distribution is sufficient to prevent coding. Moreover, his definition excludes many natural concept classes, which intuitively should be learnable. For that reason, Kearns et al. [10] and Natarajan [11] also considered learning for particular distributions. However, each of these papers deals with one particular distribution. In order to get a clearer understanding of learnability, a comprehensive theory for learnability for particular distributions is required.

Our results are more in the line of Blumer et al. [5]. Whereas, Valiant and others [15, 13, 16, 17, 10] concentrated on learning boolean formulae, Blumer et al. [5, 6] considered arbitrary concepts. Using the Vapnik-Chervonenkis dimension [14], they give necessary and sufficient conditions for learnability. This dimension depends only on the structure of the concept class (i.e., it is independent of the distribution). They show that a concept class is learnable if and only if its dimension is finite.

Following the steps of Blumer et al., we consider learnability of arbitrary concepts; but whereas in their context the distribution is unknown to the learner, we consider learnability in the case where the distribution is known. We define the notion of "finitely coverable". This notion plays a role analogous to the Vapnik-Chervonenkis dimension, i.e., a concept class C is learnable with respect to a given distribution D if and only if it is finitely coverable with respect to D. (The definitions of "learnability" and "coverable" appear in Sections 2 and 4.) In Section 3 we prove that for any discrete distribution D all concept classes are learnable with respect to D. We give necessary and sufficient conditions for polynomial learnability in this case. In Section 5 the number of examples needed is related to the size of the cover, thus relating the size of the cover to the Vapnik-Chervonenkis dimension. Moreover, in Section 6, we show that the learning is robust (i.e., it succeeds even if part of the input is erroneous). Finally, in Section 7, we discuss an open problem regarding learnability for a set of distributions.

As in Blumer et al., we are mostly concerned with the number of examples sufficient to probabilistically identify a concept — not in the time needed to compute it (or an approximation). Indeed, in some cases the function which associates a sample with a hypothesis is undecidable, and even if it is computable the computation may be infeasible [13].

2. Learnability for distribution D

Following [5], let X be a set and D a distribution over X. A *concept class over* X is a nonempty set $C \subseteq 2^X$ of *concepts*. For $\boldsymbol{x} = (x_1, \ldots, x_l) \in X^l$ and $c \in C$, the *labeled l-sample of c* is given by $\mathrm{sam}_c(\mathbf{x}) = (\langle x_1, I_c(x_1)\rangle, \ldots, \langle x_l, I_c(x_l)\rangle)$, where $I_c(x_j)$ equals 1 if $x_j \in c$ and 0 otherwise. The *sample space of C*, denoted S_C, is the set of all labeled l-samples of c over all $c \in C$ and all $\mathbf{x} \in X^l$ for all $l \geqslant 1$.

Let C be a concept class over X and H an algebra of Borel sets over X. Then F_{CH} is the set of all functions $f: S_C \to H$. In the sequel we omit C and H when understood from the context.

Our model follows the functional model of learning as defined by Haussler et al. [9]. Consider two agents, T (teacher) and L (learner): T (who wants to teach L a *target* concept c) repeatedly picks at random, according to some distribution D, an element x from a set X and sends L the pair $\langle x, I_c(x)\rangle$. L, after receiving sufficiently many examples, applies a function $f \in F_{CH}$ to return the set $f((\langle x_1, I_c(x_1)\rangle, \ldots, \langle x_l, I_c(x_l)\rangle))$. (This function is not necessarily computable.)

As in [5], throughout the paper we assume that X is a fixed set, which is either finite or countable or E^r (Euclidean r-dimensional space) for some $r \geqslant 1$. In the latter case, we assume that each $c \in C$ and $h \in H$ is a Borel set.

Let Y_1, $Y_2 \subseteq X$ we say that Y_1 and Y_2 are *ε-close with respect to the distribution D* if $\Pr_D(Y_1 \oplus Y_2) < \varepsilon$ ($\oplus$ denotes the symmetric difference). Otherwise, Y_1 and Y_2 are *ε-far with respect to the distribution D*. Notice that $\Pr_D(Y_1 \oplus Y_2)$ is a pseudo-metric on the measurable sets of X. Thus, in particular, it obeys the triangle inequality,

$$\Pr_D(Y_1 \oplus Y_3) \leqslant \Pr_D(Y_1 \oplus Y_2) + \Pr_D(Y_2 \oplus Y_3).$$

Learnability for every distribution [5]. C is *learnable in terms of H* if there exists a function $f \in F_{CH}$ such that for every $\varepsilon, \delta > 0$ there is an $l > 0$ such that for every D and every target $c \in C$ if $\mathbf{x} \in X^l$ is selected at random by D then, with probability at least $1 - \delta, f(\mathrm{sam}_c(\mathbf{x}))$ is a set ε-close to c.

We now wish to extend the notion of learnability to sets which are learnable with respect to a particular distribution D, thus in particular D is known to the learner.

Learnability for a given distribution D. C is *learnable with respect to D in terms of H* if there exists a function $f \in F_{CH}$ such that for all $\varepsilon, \delta > 0$ there is an $l > 0$ such that for every target $c \in C$ and for $\mathbf{x} \in X^l$ selected at random by D, then with probability at least $1 - \delta, f(\mathrm{sam}_c(\mathbf{x}))$ is a set ε-close to c. In this case, we say that f *learns C with respect to D with accuracy ε and confidence δ.*

3. Discrete distributions

A distribution D over X is *discrete* if X contains a countable subset Y such that $\sum_{x \in Y} \Pr_D(\{x\}) = 1$. (We shall abuse the notation and write $\Pr_D(x)$ instead of $\Pr_D(\{x\})$.) The following generalization of the Coupon Collector problem [8] will help us prove that learning with respect to such distributions is easy.

Lemma 3.1. *Let $A_1, \ldots, A_r$ be events each with probability greater than or equal to η. Then in a sequence of $l = (1/\eta)\ln(r/\delta)$ independent trials*[1]*, the probability that every event occurred at least once is greater than $1-\delta$.*

We now state the following theorem.

Theorem 3.2. *Let X be a set, C some concept class over X and D a discrete distribution over X. Then C is learnable with respect to D.*

Proof. Since D is discrete, there are $x_i \in X$ for $i = 1, 2, 3, \ldots$ such that $\sum_{i=1}^{\infty} \Pr_D(x_i) = 1$. Without loss of generality, let the x_i's be ordered such that $\Pr_D(x_i) \geqslant \Pr_D(x_{i+1})$. For $\varepsilon > 0$ there is a k such that $\sum_{i=k+1}^{\infty} \Pr_D(x_i) < \varepsilon$.

Given $\delta > 0$ and $l \geqslant \ln(k/\delta)/\Pr_D(x_k)$ examples, then by Lemma 3.1, with probability greater than $1-\delta$, each element of $\{x_1, \ldots, x_k\}$ must appear in the examples at least once. Thus any subset of $\{x_i : i = 1, 2, \ldots\}$, consistent with the examples, is ε-close to the target concept. □

Example 3.3. Let $X = (0, 1)$ and C be all open sets in X. Blumer et al. [5] showed that C is not learnable for all distributions. However, for the distribution

$$\Pr(x) = \begin{cases} \dfrac{1}{2^n} & \text{if } x = \dfrac{1}{2n} \text{ for } n \geqslant 1, \\ 0 & \text{otherwise,} \end{cases}$$

C is learnable. Moreover, we may use the following algorithm.

Given $\varepsilon, \delta > 0$, let $N = \lfloor \log(1/\varepsilon) \rfloor + 1$ and $l = \lceil 2^N \ln(N/\delta) \rceil$. For a sample of size l, choose any set $h \in H$ consistent with the sample.

Note that $\sum_{i=N+1}^{\infty} \Pr(1/i) < \varepsilon$ and by Lemma 3.1, with probability at least $1-\delta$ all data points $1/(2i)$ for $i = 1, \ldots, N$ are included in the examples. Thus with probability at least $1-\delta$, h is ε-close to the target concept.

The following theorems show conditions for learning with a polynomial sample.

Theorem 3.4. *Let $X = \{x_i\}_{i=1}^{\infty}$ be an infinite countable set, D a distribution over X, $\beta > 0$ and $p_i = \Pr_D(x_i)$ a monotonically nonincreasing sequence. If $p_i = \mathrm{O}(i^{-(\beta+1)})$ then any concept class C over X is learnable with respect to D by a sample whose size is polynomial in δ^{-1} and ε^{-1} (the confidence and accuracy parameters).*

Proof. Define $R_k = \sum_{i=k+1}^{\infty} p_i$, for any k, and let b satisfy $p_i \leqslant b/i^{(\beta+1)}$. Then

$$R_k \leqslant b \sum_{i=k+1}^{\infty} i^{-(\beta+1)} \leqslant b \int_k^{\infty} \frac{\mathrm{d}x}{x^{\beta+1}} = \frac{-b}{\beta} \frac{1}{x^{\beta}} \bigg|_k^{\infty} = \frac{b}{\beta k^{\beta}} < \tfrac{1}{2}\varepsilon,$$

for

$$k = \left\lceil \left(\frac{b}{\frac{1}{2}\beta\varepsilon} \right)^{1/\beta} \right\rceil = \mathrm{O}\left(\frac{1}{\varepsilon} \right)^{1/\beta}.$$

[1] Throughout the paper ln denotes the natural logarithm and log the logarithm to the base 2.

Claim. *There exists $m \leqslant k$ such that $R_m \leqslant \varepsilon$ and $p_i \geqslant \varepsilon/(2k)$ for $i = 1, \ldots, m$.*

Proof. If $p_k \geqslant \varepsilon/(2k)$ we are done. Otherwise, let m be the (unique) index satisfying $R_m \leqslant \varepsilon < R_{m-1}$. Since $R_k \leqslant \frac{1}{2}\varepsilon$,

$$\sum_{i=m}^{k} p_i = R_{m-1} - R_k > \tfrac{1}{2}\varepsilon.$$

Since the p_i's are nonincreasing,

$$p_1 \geqslant p_2 \geqslant \cdots \geqslant p_m = \max\{p_m, \ldots, p_k\} > \tfrac{1}{2}\varepsilon/(k-m+1) > \tfrac{1}{2}\varepsilon/k. \qquad \square$$

To finish the proof of the theorem, for a sample of size

$$l = \lceil 2k\varepsilon^{-1} \ln(m\delta^{-1}) \rceil = \mathrm{O}(\varepsilon^{-(1+1/\beta)} \log \varepsilon^{-1} \log \delta^{-1}),$$

the learning function returns any hypothesis consistent with a sample.

Let A_i $(i = 1, \ldots, m)$ be the event that x_i is chosen (in a sample of size 1). By the claim, $\Pr_D(A_i) \geqslant \frac{1}{2}\varepsilon/k$. By Lemma 3.1, with probability greater than or equal to $1-\delta$, all the elements of $\{x_1, \ldots, x_m\}$ appeared in the sample, thus the hypothesis agrees with the target on a set of probability greater than $1-\varepsilon$. $\square$

Note that the only computation involved is that of finding the hypothesis from the sample. If this can be done in polynomial time then the entire learning process can be conducted in polynomial time.

Theorem 3.5. *Let D be a distribution over an infinite countable domain X, and β, $b > 0$. If for all $i \geqslant 2$, $\Pr_D(x_i) \geqslant b/(i \ln^{\beta+1} i)$, then there exists a concept class C over X such that any learning function that learns C with accuracy $\varepsilon \leqslant (b/\beta)^2$ and confidence $\delta = \frac{1}{2}$ requires at least $\lfloor \exp(\varepsilon^{-1/(2\beta)}) \rfloor$ many examples.*

Proof. Let C consist of all finite subsets of X. A learning function can learn C with confidence $\delta = \frac{1}{2}$ only if the probability of the points of the sample is $> 1-\varepsilon$.

After seeing $l > 1$ points the probability of the points not seen is greater than or equal to

$$R_l = \sum_{i=l+1}^{\infty} \Pr_D(x_i) \geqslant \sum_{i=l+1}^{\infty} \frac{b}{i \ln^{\beta+1} i} > \int_{l+1}^{\infty} \frac{b}{x \ln^{\beta+1} x}\,\mathrm{d}x = \int_{\ln(l+1)}^{\infty} \frac{b}{y^{\beta+1}}\,\mathrm{d}y$$

$$= -\frac{b}{\beta y^{\beta}}\bigg|_{\ln(l+1)}^{\infty} = \frac{b}{\beta \ln^{\beta}(l+1)}.$$

If $l < \lfloor \exp(\varepsilon^{-1/(2\beta)}) \rfloor$ then $l \leqslant \lfloor \exp(\varepsilon^{-1/(2\beta)}) \rfloor - 1 \leqslant \exp(\varepsilon^{-1/(2\beta)}) - 1$ and $\ln^{\beta}(l+1) \leqslant \varepsilon^{-1/2}$. Thus $R_l > (b/\beta)\sqrt{\varepsilon} \geqslant \varepsilon$. $\square$

Note that even larger lower bounds are implied when the probabilities converge more slowly (e.g., $p_i = \theta(1/i \ln i (\ln \ln i)^{\beta+1})$ implies a double exponential lower bound). Theorem 3.5 shows that when the probabilities converge slowly, we cannot always learn with a polynomial number of examples. The fast convergence is a sufficient, not necessary condition. In some cases (such as when the concept class consists of two disjoint concepts) a concept class can be learned with a polynomial number of examples even if the probabilities converge very slowly (see Section 5 for a characterization.)

4. Finite covers

The following definition is analogous to the Vapnik-Chervonenkis dimension [14, 5] in the sense that it characterizes learnability.

Definition (*Finite cover*). Let $\varepsilon > 0$, a set $H_\varepsilon \subseteq 2^X$ is an *ε-cover* of C with respect to D if for every $c \in C$ there is an $h \in H_\varepsilon$ ε-close to c. *C is finitely coverable with respect to D* if for every $\varepsilon > 0$ there is a finite ε-cover of C (the size of the cover may depend on ε). In the sequel we omit D when understood from the context.

The cardinality of a smallest ε-cover of C with respect to D is denoted by $n_D = n_D(C, \varepsilon)$.

Example 4.1. Let X be the closed segment $[0, 1]$ and D be the uniform distribution over X. For every $i \geq 1$ let O_i consist of all open segments of length 2^{-i}. The concept class C_n consists of concepts of the form $c = \bigcup_{i=1}^n o_i$ where $o_i \in O_i$ and $o_1, \ldots, o_n$ are pairwise disjoint. Finally, the concept class $C = \bigcup_{n=1}^\infty C_n$.

Claim. *C is finitely coverable with respect to D.*

Proof. Let

$$s_i(\varepsilon) = \left[\frac{(i-1)\varepsilon}{2\lceil \log \varepsilon^{-1}\rceil}, \frac{i\varepsilon}{2\lceil \log \varepsilon^{-1}\rceil}\right],$$

and H_ε consist of all possible unions of the $s_i(\varepsilon)$'s.

Every concept $c \in C_n$ has length $\sum_{i=1}^n 2^{-i} = 1 - 2^{-n}$ and this is also its probability. If $n > \log \varepsilon^{-1}$, then $\Pr(c) > 1 - \varepsilon$, thus c is ε-close to (and thus ε-covered by) the segment $[0, 1]$, which belongs to H_ε since it is the union of all the $s_i(\varepsilon)$'s.

For $n \leq \log \varepsilon^{-1}$, then there exist $o_1, \ldots, o_n$ such that $o_i \in O_i$ and $c = \bigcup_{i=1}^n o_i$ and each o_i is (ε/n)-close to a union of $s_j(\varepsilon)$'s. □

The following lemma shows that we may assume that C is covered by concepts.

Lemma 4.2. *C is finitely coverable with respect to D if and only if for every $\varepsilon > 0$ there is a finite subset C_ε of C which is an ε-cover of C with respect to D.*

Proof. If C is finitely coverable, let $\varepsilon > 0$ and let $h_1, \ldots, h_n$ be a minimum $\frac{1}{2}\varepsilon$-cover of C. Then for every h_i there is a concept $c_{j_i} \in C$ such that c_{j_i} is $\frac{1}{2}\varepsilon$-close to h_j. It is easy to show that $c_{j_1}, \ldots, c_{j_n}$ is an ε-cover of C. The other direction is trivial. □

Lemma 4.3. *Let C be a concept class, D a distribution and $\varepsilon > 0$. If there exist pairwise ε-far concepts, $c_1, \ldots, c_n \in C$, then every $\frac{1}{2}\varepsilon$-cover of C has at least n elements.*

Proof. Let $H_{\frac{1}{2}\varepsilon}$ be an $\frac{1}{2}\varepsilon$-cover and for $i = 1, \ldots, n$ let $h_i \in H_{\frac{1}{2}\varepsilon}$ be $\frac{1}{2}\varepsilon$-close to c_i. Because of the triangle inequality, no h_i can be close to more than one c_j. Therefore, $h_1, \ldots, h_n$ are distinct, and the cardinality of $H_{\frac{1}{2}\varepsilon}$ is greater than or equal to n. □

Lemma 4.4. *Let C be a concept class, D a distribution and $\varepsilon > 0$. If every ε-cover of C has at least n_D elements then there exist pairwise ε-far concepts $c_1, \ldots, c_{n_D} \in C$.*

Proof. We show by induction that for every $i \leqslant n_D$ there exists a set S_i of i ε-far concepts.

Basis: $i = 0$ — trivial.

Induction step: Let $S_{i-1} = \{c_1, \ldots, c_{i-1}\}$ consist of pairwise ε-far concepts. Since $i - 1 < n_D$, S_{i-1} is not an ε-cover, and there exists a concept c_i, ε-far from every $c_j \in S_{i-1}$. Define $S_i = S_{i-1} \cup \{c_i\}$. □

Lemma 4.5. *C is not finitely coverable if and only if for some $\varepsilon > 0$, C contains an infinite sequence of pairwise ε-far concepts.*

Proof. Assume C is not finitely coverable, we construct an infinite sequence $\{c_1, c_2, \ldots\}$ of pairwise ε-far concepts where c_i is constructed from $\{c_1, c_2, \ldots, c_{i-1}\}$ as in the previous lemma. The other direction follows from Lemma 4.3. □

Let C be a concept class which is finitely coverable with respect to distribution D. Then the following is a learning function for C with respect to D.

The best-agreement-learning-function

Input: l examples, $(\langle x_1, I_c(x_1)\rangle, \ldots, \langle x_l, I_c(x_l)\rangle)$.

(1) Let E_l be the maximum integer such that $54E_l \ln(E_l n_D(C, 1/(2E_l))) \leqslant l$.

(2) Let $B = \{b_1, \ldots, b_N\}$ be a minimum $(1/(2E_l))$-cover of C, so we have $N = n_D(C, 1/(2E_l))$.

Output: Any b_i such that the cardinality of $\{x_j : 1 \leqslant j \leqslant l, I_c(x_j) \neq I_{b_i}(x_j)\}$ is minimum (among the N elements of the cover B).

To show that this learning function indeed learns, we need the following technical lemma.

Lemma 4.6. *Let A be an event of probability at most p and B an event of probability at least q, for some $0 < p \leqslant q \leqslant 1$. Consider a sequence of l independent Bernoulli trials.*

(i) *The probability that A occurred $\lceil l(2p+q)/3 \rceil$ times or more is at most* $\exp(-l(q-p)^2/27p)$.

(ii) *The probability that B occurred $\lfloor l(p+q)/2 \rfloor$ times or less is at most* $\exp(-l(q-p)^2/8q)$.

Proof. Follows from the Chernoff inequalities (e.g., [2, Proposition 2.4]). For all n, p, β with $0 \leq p \leq 1, 0 \leq \beta \leq 1$

$$\sum_{k=\lceil(1+\beta)np\rceil}^{n} \binom{n}{k} p^k (1-p)^{n-k} \leq \exp(-\beta^2 np/3),$$

$$\sum_{k=0}^{\lfloor(1-\beta)np\rfloor} \binom{n}{k} p^k (1-p)^{n-k} \leq \exp(-\beta^2 np/2). \qquad \Box$$

Theorem 4.7. *Given a set X, a finitely coverable concept class $C \subseteq 2^X$, and a distribution D over X, the best-agreement-learning-function learns C with respect to D.*

Proof. Since E_l is a monotonically non-decreasing unbounded sequence it suffices to show that on l examples, with probability $1-(1/E_l)$, the output of the best-agreement-learning-function is $(1/E_l)$-close to c.

Let $B = \{b_1, \ldots, b_N\}$ be the $(1/(2E_l))$-cover of C found while evaluating the best-agreement-learning-function. Without loss of generality, let $b_N \in B$ be $(1/(2E_l))$-close to the target concept c, and $b_1, \ldots, b_m$ the $(1/E_l)$-far elements of B. Clearly, $m \leq N-1$.

Since b_N is $(1/(2E_l))$-close to c, the expected number of examples inconsistent with b_N (i.e., belonging to $c \oplus b_N$) is less than or equal to $l/(2E_l)$, while for $1 \leq i \leq m$ the expected number of examples belonging to $c \oplus b_i$ is greater than or equal to l/E_l. We will show that the returned value is indeed as indicated by the expectations, namely, with probability at least $1-(1/E_l)$, $c \oplus b_N$ contains less examples than any $c \oplus b_i$ $(i=1, \ldots, m)$. Thus with probability at least $1-(1/E_l)$ the algorithm does not choose any of these b_i's.

Let α be the event that at least $\lceil \frac{2}{3} l/E_l \rceil$ examples belong to $b_N \oplus c$ and for $i = 1, \ldots, m$, let β_i be the event that at most $\lfloor \frac{3}{4} l/E_l \rfloor$ examples belong to $b_i \oplus c$. If any b_i $(i \leq m)$ was chosen then either α or some β_i must have occurred.

Let A be the event that an example $x \in c \oplus b_N$, B_i the event that $x \in c \oplus b_i$. Then define $p = \Pr_D(A) \leq 1/2E_l$ and $q_i = \Pr_D(B_i) \geq 1/E_l$. By Lemma 4.6(i),

$$\begin{aligned}\Pr(\alpha) &\leq \exp\left(-\frac{(q_i - p)^2 l}{27p}\right) \leq \exp\left(-\frac{(1/(2E_l))^2 l}{(27/(2E_l))}\right) \\ &= \exp\left(-\frac{1}{54}\frac{l}{E_l}\right) \leq \exp\left(-\frac{1}{54E_l} \cdot 54E_l \ln(NE_l)\right) \leq \frac{1}{NE_l}.\end{aligned}$$

Since $(q_i - p)/q_i \geq \frac{1}{2}$, by Lemma 4.6(ii) for $i \leq m$,

$$\Pr(\beta_i) \leq \exp\left(-\frac{(q_i - p)^2 l}{8q_i}\right) \leq \exp\left(-\frac{(q_i - p)}{16} l\right) \leq \exp\left(-\frac{l}{32E_l}\right) < \frac{1}{NE_l}.$$

Thus, the probability that for some $i \leq m$, b_i is chosen is less than $(m+1)/NE_l \leq 1/E_l$. $\Box$

Notes. (1) If $m < N-1$, then there are some elements $b_{m+1}, \ldots, b_{N-1}$ in B such that $1/(2E_l) \leqslant \Pr(c \oplus b_i) < 1/E_l$. The function may prefer one of them over b_N.

(2) The naive algorithm, which returns some concept consistent with the examples, does not necessarily learn. For example, let $X = [0, 1]$, D be the uniform distribution and C consist of the set $[0, 1]$ and all finite sets. If the target concept is $[0, 1]$, then for any sample $\mathrm{sam}_{[0,1]}(\boldsymbol{x}) = (\langle x_1, 1\rangle, \ldots, \langle x_l, 1\rangle)$, the finite set $\{x_1, \ldots, x_l\}$ is consistent with the sample but not ε-close to $[0, 1]$ (for any $\varepsilon < 1$). Note, however, that $\{\emptyset, [0, 1]\}$ is an ε-cover (for all $\varepsilon > 0$), and thus C is learnable with respect to D.

Let C be a concept class, D a distribution, $\varepsilon, \delta > 0$ and $f \in F$. Then the *sample size required by* f *to learn* C *to accuracy* ε *and confidence* δ, denoted by $l_C^f(\varepsilon, \delta)$, is the minimal number l such that for every $c \in C$, if $\boldsymbol{x} \in X^l$ is selected at random by D then, with probability $1-\delta, f(\mathrm{sam}_c(\boldsymbol{x}))$ is a set ε-close to c.

Lemma 4.8. *Given a set* X, *a distribution* D *over* X, *a concept class* $C \subseteq 2^X$ *and* δ, $\varepsilon > 0$. *If there exists a set* $C_{2\varepsilon} \subseteq X$ *of* n *pairwise* 2ε*-far concepts, then for every* $f \in F$, $l_C^f(\varepsilon, \delta) \geqslant \log((1-\delta)n)$.

Proof. Let $f \in F$ learn C with respect to D with accuracy ε and confidence δ using sample size l. For $\boldsymbol{x} = (x_1, \ldots, x_l)$ and $\boldsymbol{L} = (L_1, \ldots, L_l) \in \{0, 1\}^l$, define $I(\boldsymbol{x}, \boldsymbol{L}) = (\langle x_1, L_1\rangle, \ldots, \langle x_l, L_l\rangle)$. For $c \in C$ and $\varepsilon > 0$ let

$$g_f(c, \boldsymbol{x}, \boldsymbol{L}, \varepsilon) = \begin{cases} 1 & \text{if } \Pr_D(f(I(\boldsymbol{x}, \boldsymbol{L})) \oplus c) < \varepsilon, \\ 0 & \text{otherwise.} \end{cases}$$

Let $\int_x g_f(c, \boldsymbol{x}, I_c(\boldsymbol{x}), \varepsilon)\, \mathrm{d}P_D$ be the expectation over $\boldsymbol{x}$ of the random variable g_f with respect to the l-fold distribution of D. Consider the sum

$$S = \sum_{c \in C_{2\varepsilon}} \int_x g_f(c, \boldsymbol{x}, I_c(\boldsymbol{x}), \varepsilon)\, \mathrm{d}P_D.$$

Since f learns C to accuracy ε and confidence δ using sample size l, $\int_x g_f(c, \boldsymbol{x}, I_c(\boldsymbol{x}), \varepsilon)\, \mathrm{d}P_D > 1-\delta$ for each $c \in C$, and we obtain

$$S > (1-\delta)n. \tag{1}$$

Rearranging the sum yields

$$S = \sum_{c \in C_{2\varepsilon}} \int_x g_f(c, \boldsymbol{x}, I_c(\boldsymbol{x}), \varepsilon)\, \mathrm{d}P_D \leqslant \sum_{c \in C_{2\varepsilon}} \int_x \sum_{\boldsymbol{L} \in \{0,1\}^l} g_f(c, \boldsymbol{x}, \boldsymbol{L}, \varepsilon)\, \mathrm{d}P_D$$
$$= \int_x \sum_{\boldsymbol{L} \subset \{0,1\}^l} \sum_{c \subset C_{2\varepsilon}} g_f(c, \boldsymbol{x}, \boldsymbol{L}, \varepsilon)\, \mathrm{d}P_D.$$

Since the $c \in C_{2\varepsilon}$ are 2ε-far, for every $\boldsymbol{x}$ and $\boldsymbol{L}$ there exists at most one $c \in C_{2\varepsilon}$ such that $g_f(c, \boldsymbol{x}, \boldsymbol{L}, \varepsilon) = 1$. Thus,

$$S \leqslant \int_x \left(\sum_{\boldsymbol{L} \in \{0,1\}^l} 1 \right) \mathrm{d}P_D = \int_x 2^l\, \mathrm{d}P_D = 2^l. \tag{2}$$

Combining (1) and (2) yields $l > \log((1-\delta)n)$. □

Theorem 4.7, Lemmas 4.5 and 4.8 yield our main result.

Theorem 4.9. *C is finitely coverable with respect to D if and only if C is learnable with respect to D.*

Note that the definition of learnability does not imply computability. However, if there is an effective procedure Q that, given D and ε, outputs a finite ε-cover for C and C is recursive, in the sense that there is a recursive function ϕ that given $c \in C$ and $x \in X$ determines whether $x \in c$ (it suffices that ϕ be defined only on members of the cover), then there exists a computable learning function. Furthermore, if Q requires polynomial time, then the size of the ε-cover is polynomial in ε^{-1}. If, in addition, ϕ is also polynomial time computable, then the time complexity of the learning algorithm is polynomial (in δ^{-1}, ε^{-1}, and the cumulative size of the examples).

By Lemma 4.2 if C is learnable (with respect to D) then there is a function that learns C and whose values are concepts (members of C).

5. Finite dimension and the size of the cover

We now present an interesting connection between the size of the cover and the Vapnik-Chervonenkis dimension of the concept class. First we quote a result presented in [5] and then relate it to the present work.

Let $T = \{(x_1, \ldots, x_n)\}$ be a subset of X. A concept class $C \subseteq 2^X$ *shatters* T if for every subset T' of T there is a concept $c \in C$ such that $T \cap c = T'$. Also, $\dim(C) = d$ if there is a set of d elements of X shattered by C and there is no set of $d+1$ elements shattered by C. If no such d exists, C has *infinite dimension.* The main result of [5] is that C is learnable for every distribution if and only if $\dim(C)$ is finite, namely

Theorem 5.1. *If $d = \dim(C) \geq 2$, $0 < \varepsilon \leq \frac{1}{2}$, $0 < \delta < 1$ then there is a function that learns C for every well behaved distribution*[2] *using $O((1/\varepsilon)\ln(1/\delta) + (d/\varepsilon)\ln(1/\varepsilon))$ examples.*

Recall that $n_D(\varepsilon)$ is the size of a minimum ε-cover of C with respect to D. Using Theorem 5.1 we are able to relate the size of an ε-cover to the dimension.

Theorem 5.2. *Let C be a concept class of finite dimension $d \geq 2$. Then the following three relations hold:*

(1) *There is a distribution D such that $n_D(\frac{1}{4}) \geq \lfloor \log d \rfloor$.*

(2) *If $\varepsilon < 1/(2d)$ then there is a distribution D such that $n_D(\varepsilon) \geq 2^d$.*

[2] The notion of well behaved distributions is discussed in [3, 5].

(3) *There exists a constant κ such that for every $0<\varepsilon\leq\frac{1}{2}$ and every well behaved distribution D, $n_D(\varepsilon)<\varepsilon^{-d\kappa/\varepsilon}$.*

Proof. Let $T=\{x_1,\ldots,x_d\}\subseteq X$ be shattered by C and let D be the uniform distribution over $\{x_1,\ldots,x_d\}$. That is,

$$\Pr(x)=\begin{cases}\dfrac{1}{d} & x\in\{x_1,\ldots,x_d\},\\ 0 & \text{otherwise}.\end{cases}$$

For $i=1,\ldots,\lfloor\log d\rfloor$ let $T_i=\{x_j\text{: the }i\text{th bit of }j\text{ is }1\}$. Since $T_i\subseteq T$ and T is shattered by C, there exist concepts $c_i\in C$ such that $T\cap c_i=T_i$. It is easy to see that the c_i's are $\frac{1}{2}$-far from one another, thus proving, by Lemma 4.3, the first inequality.

On the other hand, every two distinct subsets of T are at least $(1/d)$-far with respect to D. Since T is shattered by C there are at least 2^d concepts $(1/d)$-far from one another, which by Lemma 4.3 proves the second inequality.

For (3), by Theorem 5.1, for every distribution D and $\varepsilon>0$, C is learnable with accuracy $\frac{1}{2}\varepsilon$ and confidence $\frac{1}{2}$ using

$$l(\tfrac{1}{2}\varepsilon,\tfrac{1}{2})\leq\frac{\kappa' d}{\varepsilon}\log\frac{1}{\varepsilon}$$

examples. By Lemma 4.4 there is a set C_ε of $n_D(\varepsilon)$ pairwise ε-far concepts. By Lemma 4.8,

$$\log((1-\tfrac{1}{2})n_D(\varepsilon))\leq l(\tfrac{1}{2}\varepsilon,\tfrac{1}{2}).$$

Therefore, $n_D(\varepsilon)<\varepsilon^{-d\kappa/\varepsilon}$, where $\kappa=\kappa'+1$. □

6. Learning with errors

The output of the best-agreement-learning-function, presented in the previous section, is not necessarily consistent with the labeled sample presented as input. This fact suggests that the algorithm is robust, i.e., learning is possible even when some of the input contains some errors.

More precisely, let $0\leq\zeta<\frac{1}{2}$ be the probability that the label of a certain example is wrong. We assume that the errors are independent. In particular, the same data point, if repeated, may have different labelings. Let $c\in C$ be the target concept and $x\in c$ be some data point that, during the learning process, was randomly chosen several times. In an error free labeled sample the label of x must always be $I_c(x)$ (in this case 1). When independent errors are present the label of x can be 1 for some occurrences and 0 for others, thus the sample may become inconsistent. This can happen, for example, if the communication channel between the teacher and learner induces some random errors or if the "teacher" is a human expert making human errors or random measurement errors occur.

Even in this case the best-agreement-learning-function can learn C if it is provided with more examples.

Theorem 6.1. *Let C be a finitely coverable concept class, D a distribution and $0 \leqslant \zeta < \frac{1}{2}$ be the probability of error in the examples. Then the best-agreement-learning-function learns C with $l = \lceil 54 \ln(N/\delta)/(\varepsilon^2(1-2\zeta)^2) \rceil$ examples.*

Proof. The proof is similar to that of Theorem 4.7. Let $b_1, \ldots, b_m, \ldots, b_N$ be as defined there and p the probability that a single example x of c is inconsistent with b_N. There are two possibilities for this to happen:

(i) When the label of the example is correct and $I_c(x) \neq I_{b_N}(x)$.

(ii) $I_c(x) = I_{b_N}(x)$ but the label is incorrect.

Since b_N is $\frac{1}{2}\varepsilon$-close to c,

$$p \leqslant \tfrac{1}{2}\varepsilon(1-\zeta) + \zeta(1-\tfrac{1}{2}\varepsilon),$$

where the two terms correspond to cases (i) and (ii), respectively.

Similarly let q be the probability that a single example x of c is inconsistent with some b_i for $1 \leqslant i \leqslant m$. Since b_i is ε-far from c, the above considerations yield

$$q \geqslant \varepsilon(1-\zeta) + (1-\varepsilon)\zeta.$$

Note that $p \leqslant \frac{1}{2}\varepsilon(1-2\zeta) + \zeta$ and $q \geqslant \varepsilon(1-2\zeta) + \zeta$.

Let α be the event that at least $\lceil(\frac{2}{3}\varepsilon(1-2\zeta)+\zeta)l\rceil$ examples belong to $b_N \oplus c$ and for $i = 1, \ldots, m$ let β_i be the event that at most $\lfloor(\frac{3}{4}\varepsilon(1-2\zeta)+\zeta)l\rfloor$ examples belong to $b_i \oplus c$.

As in Theorem 4.7, we obtain by Lemma 4.6 (and the observation $\frac{1}{2}\varepsilon(1-2\zeta)+\zeta<\frac{1}{2}$) that $\Pr(\alpha) \leqslant \delta/N$ and $\Pr(\beta_i) < \delta/N$. Therefore, the probability that some b_i is chosen is less than $(m+1)\delta/N \leqslant \delta$. □

Using similar techniques Angluin and Laird [1] have independently constructed an algorithm that learns (for every distribution) finite concept classes from examples containing errors.

7. Conclusions

In this paper we have extended the notion of learnability and have shown how concept classes which were not learnable by previous definitions became (robustly) learnable. We found a general theorem that enables us to decide whether a concept class is learnable. We show that in the limit, learnability for all distributions, our condition is equivalent to the finite dimension condition presented in [5].

Open problem. Can learnability be generalized (possibly robustly) for a set $\mathcal{D}$ of distributions? We discussed two extreme cases: $\mathcal{D}$ is a singleton and $\mathcal{D}$ consists of all distributions. Two other cases are quite obvious: if $\mathcal{D}$ is finite this is analogous to the case where $\mathcal{D}$ is a singleton; if $\mathcal{D}$ is the set of all discrete distributions, this is analogous to the case where $\mathcal{D}$ consists of all distributions. We conjecture that a concept class C is learnable with respect to a set of distributions $\mathcal{D}$ if and only if there is a function $n_{\mathcal{C}}(\cdot)$ such that for every $\varepsilon > 0$ and every distribution $D \in \mathcal{D}$, C is finitely coverable by at most $n_{\mathcal{C}}(\varepsilon)$ sets. The difficulty arises from the fact that for different distributions there may be different covers and not all concepts consistent with the examples are close to one another. Natarajan [12] gave two conditions, one of which is sufficient and the other necessary.

The notion of learnability can be also extended by considering non-uniform learnability, i.e., learning when the number of examples needed depends on the target concept. Further results in this direction appear in [4].

Acknowledgment

It is a pleasure to thank Johann Makowsky and Ronny Roth for their comments and useful suggestions.

References

[1] D. Angluin and P.D. Laird, Learning from noisy examples, *Machine Learning* **2** (1988) 243-270.

[2] D. Angluin and L.G. Valiant, Fast probabilistic algorithms for Hamiltonian circuits and matchings, *J. Comput. System Sci.* **18** (1979) 155-193.

[3] S. Ben-David and G.M. Benedek, Measurability constraints of *PAC* learnability, Preprint, 1989.

[4] G.M. Benedek and A. Itai, Nonuniform learnability, in: *15th ICALP* (1988) 82-92.

[5] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik-Chervonenkis dimension, *J. Assoc. Comput. Mach.* **36** (1989) 929-965.

[6] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Occam's razor, *Inform. Process. Lett.* **24** (1987) 377-380.

[7] A. Ehrenfeucht, D. Haussler, M. Kearns and L. Valiant, A general lower bound on the number of examples needed for learning, in: *COLT 1988; Inform. and Comput.*, to appear.

[8] W. Feller, *An Introduction to Probability Theory and its Applications* (Wiley, New York, 1950).

[9] D. Haussler, M. Kearns, N. Littlestone and M. Warmuth, Equivalence of models for polynomial learnability, in: *COLT '88* (1988) 42-55.

[10] M. Kearns, Ming Li, L. Pitt and L.G. Valiant, On the learnability of boolean formulae, in: *Proc. 19th Ann. STOC* (ACM, New York, 1987) 285-295.

[11] B.K. Natarajan, On learning boolean functions, in: *Proc. 19th Ann. STOC* (ACM, New York, 1987) 296-304.

[12] B.K. Natarajan, Probably-approximate learning over classes of distributions, unpublished manuscript.

[13] L. Pitt and L.G. Valiant, Computational limitations on learning from examples, *J. ACM* **35**(4) (1988) 965-984.

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

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

[16] L.G. Valiant, Learning disjunctions of conjunctions, in: *Proc. 9th IJCAI*, Vol. 1, Los Angeles, CA (1985) 560-566.

[17] L.G. Valiant, Deductive learning, *Philos. Trans. Roy. London Ser. A* **312**, 441-446.