On Agnostic Learning with $\{0, *, 1\}$-valued and Real-valued Hypotheses

Philip M. Long

Genome Institute of Singapore
1 Research Link
IMA Building
National University of Singapore
Singapore 117604, Republic of Singapore

Abstract. We consider the problem of classification using a variant of the agnostic learning model in which the algorithm's hypothesis is evaluated by comparison with hypotheses that do not classify all possible instances. Such hypotheses are formalized as functions from the instance space X to $\{0, *, 1\}$, where $*$ is interpreted as "don't know". We provide a characterization of the sets of $\{0, *, 1\}$-valued functions that are learnable in this setting. Using a similar analysis, we improve on sufficient conditions for a class of real-valued functions to be agnostically learnable with a particular relative accuracy; in particular, we improve by a factor of two the scale at which scale-sensitive dimensions must be finite in order to imply learnability.

1 Introduction

In agnostic learning [13, 17], an algorithm tries to find a hypothesis that generalizes nearly as well as is possible using any hypothesis in some class that is known a priori; this class is sometimes called the *comparison class*. This framework can be applied for analyzing algorithms for two-class classification problems; in this case, one can view hypotheses as functions from some domain X to $\{0, 1\}$.

In this paper, we consider a modified framework in which the members of the comparison class do not classify all elements of the domain, and are regarded to be wrong an any domain elements that they do not classify. Formally, hypotheses in this framework map X to $\{0, *, 1\}$, where $*$ is regarded as "don't know". This offers a clean way to make formal use of the intuition that points are unlikely to fall in the unclassified region, since results in this framework are strong to the extent that this is true.

For example, it can be used to formalize the assumption that there is a halfspace that is likely to classify instances correctly with a certain margin; such a halfspace has small error, even if instances falling close to its separating hyperplane are regarded as being classified incorrectly (i.e. are mapped to $*$). This viewpoint is implicit in the manner in which the "margin percentile bounds" for generalization of support vector machines [3, 2, 11] are formulated. These results bound the probability that there is some halfspace that classifies a large fraction

of a random sample correctly with a large margin, but fails to generalize well. Such results are interesting when it is likely, for a collection of random examples, that some halfspace gets most of the examples correct with a large margin, and this is the case when some halfspace is likely to classify individual random examples correctly with a large margin. A similar line of reasoning suggests that this assumption is implicit in Ben-David and Simon's [6] choice of analysis for their computationally efficient algorithm; indeed, agnostic learning of $\{0, *, 1\}$-valued functions in the model studied here abstracts the optimization criterion studied in their paper (see also [7]).

In this paper, we show that a generalization of the VC-dimension to $\{0, *, 1\}$-valued functions introduced in [4] provides a characterization of learnability in this setting, in that a class of functions is learnable if an only if its generalized VC-dimension is finite.

Next, we turn to the problem of learning with real-valued hypotheses. Scale-sensitive notions of the dimension of a class of real-valued functions have been used to characterize the learnability of classes of real-valued functions in different settings [1, 5]: loosely, these results say that a class can be learned to any accuracy if and only if its dimension is finite at all scales. Previous work [4] considered the following question: at what scale does the dimension need to be finite for learning to a particular relative accuracy to be possible? This work left roughly a factor of two gap between the scales at which finite dimension is necessary and is sufficient. In this paper, we close this gap, improving by a factor of two the bound on the scale at which the dimension of a class of real-valued functions must be finite for it to be agnostically learnable.

The model of agnostic learning of $\{0, *, 1\}$-valued functions calls to mind the "sleeping experts" framework [12], but there are many differences, including the usual differences between batch and online learning settings. Blum, et al [8] studied a variant of the model of PAC learning with membership queries in which queries falling in a given region are answered with "don't know" and the distribution assigned zero weight to this "don't know" region.

2 Characterization of agnostic learnability with $\{0, *, 1\}$-valued hypothesis

2.1 Definitions

Say a set F of functions from X to $\{0, *, 1\}$ *shatters* elements $x_1, ..., x_d$ of X if

$$\{0,1\}^d \subseteq \{(f(x_1), ..., f(x_d)) : f \in F\}.$$

Define $\mathrm{VCdim}(F)$ [19, 4] to be the size of the largest set shattered by F.

An *example* is an element of $X \times \{0, 1\}$ and a *sample* is a finite sequence of examples. A *hypothesis* is a function from X to $\{0, *, 1\}$. Define $\ell : \{0, *, 1\} \times \{0, 1\} \to \{0, 1\}$ by $\ell(\hat{y}, y) = 1$ iff $\hat{y} \neq y$. For a hypothesis h, and a probability distribution P over $X \times \{0, 1\}$, define the *error* of h with respect to P, to be $\mathrm{er}_P(h) = \mathbf{Pr}_{(x,y)\sim P}(h(x) \neq y)$.

A *learning strategy* is a mapping from samples to hypotheses. A *prediction strategy* [15] takes as input a sample and an element of X, and outputs an element of $\{0,1\}$.

A set F of functions from X to $\{0,*,1\}$ is said to be *agnostically learnable* if there is a learning strategy such that, for all $\epsilon, \delta > 0$, there is a natural number m such that, for any probability distribution P over $X \times \{0,1\}$, if m examples are drawn independently at random according to P, and the resulting sample is passed to A which outputs a hypothesis h, then, with probability at least $1-\delta$, $\mathrm{er}_P(h) \leq (\inf_{f \in F} \mathrm{er}_P(f)) + \epsilon$.

2.2 Overview of some technical issues involved

The model of agnostic learning of $\{0,*,1\}$-valued functions falls within the general decision-theoretic framework proposed by Haussler [13]. In a special case of Haussler's framework, there is an instance space X, an action space A, an outcome space Y, a loss function $\ell : A \times Y \to \mathbf{R}^+$ and a comparison class F of functions mapping X to A. Given examples $(x_1, y_1), ..., (x_m, y_m)$ drawn independently at random according to a probability distribution P over $X \times Y$, a learning algorithm outputs a hypothesis h mapping X to A. Roughly, the goal is for $\mathbf{E}_{(x,y)\sim P}(\ell(h(x), y))$ to be close to $\inf_{f \in F} \mathbf{E}_{(x,y)\sim P}(\ell(f(x), y))$.

The model of this paper can be recovered by setting $A = \{0,*,1\}, Y = \{0,1\}$, and letting ℓ be the discrete loss, i.e. $\ell(\hat{y}, y) = 1$ if $\hat{y} \neq y$ and $\ell(\hat{y}, y) = 0$ if $\hat{y} = y$.

Unfortunately, some of the general analysis techniques [19, 18, 13] that have been applied in a wide range of concrete problems falling within this framework cannot be applied in the $\{0,*,1\}$ case. The by now standard analysis considers a class of loss functions defined as follows. For each f, define $\ell_f : X \times Y \to \mathbf{R}^+$ to give the loss incurred by f, i.e. $\ell_f(x,y) = \ell(f(x), y)$. Then $\ell_F = \{\ell_f : f \in F\}$. The usual analysis proceeds by showing that conditions on F imply that ℓ_F is somehow "limited". For example, if $A = \{0,1\}$, $Y = \{0,1\}$ and ℓ is the discrete loss, then $\mathrm{VCdim}(\ell_F) \leq \mathrm{VCdim}(F)$. In our setting, it appears that nothing useful of this type is true; the set F of all functions from $\mathbf{N}$ to $\{0,*\}$ has $\mathrm{VCdim}(F) = 0$, but $\mathrm{VCdim}(\ell_F) = \infty$.

Instead, we use an approach from [10, 4], in which given

$$(x_1, f(x_1)), ..., (x_m, f(x_m)),$$

and wanting to evaluate $h(x)$, the algorithm constructs a small cover of the restrictions of the functions in F to $x_1, ..., x_m, x$. In this context, loosely speaking, a cover of a set of functions is another set of functions for which each element of the set being covered is approximated well by some element of the set doing the covering. To analyze such an algorithm in this setting required a lemma about the existence of small covers. All bounds we know on covering numbers for learning applications proceed by first bounding packing numbers, and then appealing to a general bound on covering numbers in terms of packing numbers. (Roughly, a packing number of a set is the size of the largest pairwise distant subset.) It appears that this cannot work in this setting, because the relevant

notion of "approximation" (defined below) is not a metric. The main technical novelty in this paper is a proof of a covering lemma that does not rely on packing.

2.3 The covering lemma

For this subsection, fix a finite set Z. Say that a function g from Z to $\{0,1\}$ *k-covers* a function f from Z to $\{0,*,1\}$ if $|\{z \in Z : f(z) \neq * \text{ and } f(z) \neq g(z)\}| \leq k$.

Say that a set G of $\{0,1\}$-valued functions k-covers a set F of $\{0,*,1\}$-valued functions if each function in F is k-covered by some function in G.

For technical reasons, it will be useful for a moment to consider learning when the *examples* can be labelled with $*$. In this context, an $*$ can be interpreted as "doesn't matter". For a hypothesis h, a function f from Z to $\{0,*,1\}$, and a probability distribution D over Z, define the *error* of h with respect to f and D, to be

$$\text{er}_{f,D}(h) = \mathbf{Pr}_{z \sim D}((h(z) \neq f(z)) \wedge (f(z) \neq *))$$

We will make use of the following known result about this model.

Lemma 1 ([4]). *Choose a set F of functions from Z to $\{0,*,1\}$, and let d be the VC-dimension of F. There is a mapping A from finite sequences of elements of $Z \times \{0,*,1\}$ to hypotheses such that, for any probability distribution D over Z, for any $f \in F$, and for any positive integer t, if $z_1, ..., z_t$ are chosen independently at random according to D, and A is applied to $(z_1, f(z_1)), ..., (z_t, f(z_t))$, and h is the resulting hypothesis, then $\mathbf{E}(\text{er}_{f,D}(h)) \leq \frac{d}{t+1}$.*

Lemma 2. *Let $m = |Z|$. Choose a set F of functions from Z to $\{0,*,1\}$ and let $d = \text{VCdim}(F)$. Choose an integer k such that $1 \leq k \leq m$. There is a set G of $\{0,1\}$-valued functions and a subset F' of F such that (a) $|F'| \geq |F|/2$, (b) G k-covers F', and (c) $|G| \leq 3^{\lceil 2dm/k \rceil}$.*

Proof: Define A as in Lemma 1. Let D be the uniform distribution over Z. Let P be the uniform distribution over F. Choose t (its value will be set later). Suppose that $z_1, ..., z_t$ are chosen independently at random according to D, f is chosen independently according to P, and $(z_1, f(z_1)), ..., (z_t, f(z_t))$ are passed to A; let $A(z_1, ..., z_t; f)$ be A's hypothesis (viewed as a random variable). Then Lemma 1 says that

$$\forall f \in F, \mathbf{E}_{z_1,...,z_t \sim D^t}(\text{er}_{f,D}(A(z_1, ..., z_t; f))) < \frac{d}{t}.$$

Markov's inequality implies that

$$\forall f \in F, \mathbf{Pr}_{z_1,...,z_t \sim D^t}\left(\text{er}_{f,D}(A(z_1, ..., z_t; f)) > \frac{2d}{t}\right) \leq 1/2.$$

Thus,

$$\mathbf{Pr}_{f \sim P, z_1,...,z_t \sim D^t}\left(\text{er}_{f,D}(A(z_1, ..., z_t; f)) > \frac{2d}{t}\right) \leq 1/2.$$

Fubini's Theorem implies that

$$\mathbf{E}_{z_1,\ldots,z_t\sim D^t}\left(\mathbf{Pr}_{f\sim P}(\mathrm{er}_{f,D}(A(z_1,\ldots,z_t;f))>\frac{2d}{t})\right)\leq 1/2.$$

Thus,

$$\exists z_1,\ldots,z_t,\ \mathbf{Pr}_{f\sim P}\left(\mathrm{er}_{f,D}(A(z_1,\ldots,z_t;f))>\frac{2d}{t}\right)\leq 1/2. \tag{1}$$

Choose such a sequence $z_1,\ldots,z_t$. Let $G=\{A(z_1,\ldots,z_t;f):f\in F\}$, and let $F'=\left\{f\in F:\mathrm{er}_{f,D}(A(z_1,\ldots,z_t;f))\leq\frac{2d}{t}\right\}$. Note that G $\frac{2dm}{t}$-covers F', and, by (1), $|F'|\geq|F|/2$. Suppose $t=\lceil 2dm/k\rceil$; then $\frac{2dm}{t}\leq k$. There are only 3^t possible inputs to A with instances $x_1,\ldots,x_t$. Thus, $|H|\leq 3^t$, completing the proof. □

Theorem 1. *Let $m=|Z|$. Choose a set F of functions from Z to $\{0,*,1\}$ and let $d=\mathrm{VCdim}(F)$. Choose an integer k such that $1\leq k\leq m$. There is a set G of $\{0,1\}$-valued functions that k-covers F and for which $|G|\leq\lceil m\log_2 3\rceil 3^{\lceil 2dm/k\rceil}$.*

Proof: Construct a sequence $G_1,G_2,\ldots,G_{\lceil\log_2|F|\rceil}$ of sets of functions from X to $\{0,1\}$ by repeatedly applying Lemma 2 to k-cover at last half of the remaining functions in F, and then deleting the covered functions. Let $G=\cup G_i$. Then G k-covers F, and $|G|\leq\lceil\log_2|F|\rceil 3^{\lceil 2dm/k\rceil}\leq\lceil m\log_2 3\rceil 3^{\lceil 2dm/k\rceil}$.

2.4 Learning

Theorem 2. *A set F of functions from X to $\{0,*,1\}$ is learnable if and only if $\mathrm{VCdim}(F)$ is finite.*

The necessity follows from the corresponding result for the $\{0,1\}$ case [9]. The sufficiency is a direct consequence of the following theorem. The following proof closely follows that of Theorem 21 of [4]; the main difference is that it appeals to the new Theorem 1 of the present paper.

Theorem 3. *Choose a set F of functions from X to $\{0,*,1\}$ for which $\mathrm{VCdim}(F)$ is finite. Let $d=\mathrm{VCdim}(F)$.*

There is a prediction strategy A and constants c and m_0 such that, for any probability distribution P over $X\times\{0,*,1\}$, for any $m\geq m_0$, if $(x_1,y_1),\ldots,(x_m,y_m)$ are drawn independently at random according to P, and $(x_1,y_1),\ldots,(x_{m-1},y_{m-1})$ and x_m are given to A, which outputs $\hat{y}_m$, then

$$\mathbf{E}(\ell(\hat{y}_m,y_m))-\inf_{f\in F}\mathrm{er}_P(f)\leq c\left(\frac{d}{m}\right)^{1/3}.$$

Proof: Assume without loss of generality that m is even. Choose $\alpha>0$ (its value will be set later).

Choose a function Φ that maps from X^m to the set of finite subsets of $\{0,1\}^m$ such that, for any $(x_1,\ldots,x_m)\in X^m$, $\Phi(x_1,\ldots,x_m)$ is one of the smallest sets that

αm-covers $\{(f(x_1), ..., f(x_m)) : f \in F\}$ and $\Phi(x_1, ..., x_m)$ is invariant under permutations of its arguments. (When defining "αm-covers" above, we are viewing an element of $\{0, *, 1\}^m$ as a function from $\{1, ..., m\}$ to $\{0, *, 1\}$.)

Consider the prediction strategy A that chooses $\hat{y} = (\hat{y}_1, ..., \hat{y}_m)$ from among the elements of $\Phi(x_1, ..., x_m)$ in order to minimize $\sum_{i=1}^{m/2} \ell(\hat{y}_i, y_i)$ (the error on the first half of the sample only), and outputs $\hat{y}_m$.

For $a \in \{0,1\}^m$, $b \in \{0, *, 1\}^m$, define

$$\ell^{\text{first}}(a, b) = \frac{2}{m} \sum_{i=1}^{m/2} \ell(a_i, b_i),$$

$$\ell^{\text{last}}(a, b) = \frac{2}{m} \sum_{i=m/2+1}^{m} \ell(a_i, b_i),$$

and

$$\ell^{\text{all}}(a, b) = \frac{1}{m} \sum_{i=1}^{m} \ell(a_i, b_i).$$

Fix a distribution P on $X \times \{0, 1\}$, and suppose $(x_1, y_1), ..., (x_m, y_m)$ are chosen independently at random from P. Let $x = (x_1, ..., x_m)$ and $y = (y_1, ..., y_m)$. Choose $f^* \in F$ that satisfies $\text{er}_P(f^*) \leq \inf_{f \in F} \text{er}_P(f) + \alpha$. Since $\Phi(x)$ αm-covers $\{(f(x_1), ..., f(x_m)) : f \in F\}$

$$\exists t^* \in \Phi(x_1, ..., x_m), \ell^{\text{all}}(t^*, y) \leq \alpha + \ell^{\text{all}}((f^*(x_1), ..., f^*(x_m)), y). \tag{2}$$

(If $f^*(x_i) = *$, whatever the values of t_i and y_i, $\ell(t_i, y_i) \leq \ell(f^*(x_i), y_i)$.)

Applying the Hoeffding bound,

$$\mathbf{Pr}\left(\ell^{\text{all}}((f^*(x_1), ..., f^*(x_m)), y) > \text{er}_P(f^*) + \alpha\right) \leq e^{-2\alpha^2 m}. \tag{3}$$

Now, let U be the uniform distribution over $\{-1, 1\}^{m/2}$. Then, since Φ is invariant under permutations,

$$\begin{aligned}
&\mathbf{Pr}(\exists t \in \Phi(x) \ |\ell^{\text{first}}(t, y) - \ell^{\text{last}}(t, y)| > 2\alpha) \\
&\leq \sup_{(x,y)} \mathbf{Pr}_{u \in U} \left(\exists t \in \Phi(x), \left|\tfrac{2}{m} \textstyle\sum_{i=1}^{m/2} u_i \left(\ell(t_i, y_i) - \ell(t_{i+m/2}, y_{i+m/2})\right)\right| > 2\alpha\right)
\end{aligned}$$

For any fixed $t \in \Phi(x)$, Hoeffding's inequality implies

$$\mathbf{Pr}_{u \in U} \left(\left|\frac{2}{m} \sum_{i=1}^{m/2} u_i \left(\ell(t_i, y_i) - \ell(t_{i+m/2}, y_{i+m/2})\right)\right| > 2\alpha\right) \leq 2e^{-\alpha^2 m}.$$

So with probability at least $1 - |\Phi(x)| 2e^{-\alpha^2 m}$, for all t in $\Phi(x)$,

$$|\ell^{\text{first}}(t, y) - \ell^{\text{last}}(t, y)| \leq 2\alpha.$$

This implies

$$|\ell^{\text{first}}(t,y) - \ell^{\text{all}}(t,y)| \leq \alpha$$

and

$$|\ell^{\text{last}}(t,y) - \ell^{\text{all}}(t,y)| \leq \alpha.$$

So, combining with (3), with probability at least $1 - (1+2|\Phi(x)|)e^{-\alpha^2 m}$, the $\hat{y} \in \Phi(x)$ with minimal $\ell^{\text{first}}(\hat{y},y)$ satisfies

$$\begin{aligned}
\ell^{\text{all}}(\hat{y},y) &\leq \ell^{\text{first}}(\hat{y},y) + \alpha \\
&\leq \ell^{\text{first}}(t^*,y) + \alpha \\
&\leq \ell^{\text{all}}(t^*,y) + 2\alpha \\
&\leq \ell^{\text{all}}((f^*(x_1),...,f^*(x_m)),y) + 3\alpha \\
&\leq \text{er}_P(f^*) + 4\alpha
\end{aligned}$$

and hence

$$\begin{aligned}
\ell^{\text{last}}(\hat{y},y) &\leq \text{er}_P(f^*) + 5\alpha \\
&\leq \inf_{f\in F} \text{er}_P(f) + 6\alpha.
\end{aligned}$$

That is,

$$\mathbf{Pr}\left(\ell^{\text{last}}(\hat{y},y) > \inf_{f\in F} \text{er}_P(f) + 6\alpha\right) < (1+2|\Phi(x)|)e^{-\alpha^2 m}$$

which implies

$$\mathbf{E}(\ell^{\text{last}}(\hat{y},y)) - \inf_{f\in F} \text{er}_P(f) < 6\alpha + (1+2|\Phi(x)|)e^{-\alpha^2 m}.$$

Thus, since any of $(x_{m/2+1},y_{m/2+1}),...,(x_m,y_m)$ was equally likely to have been the last,

$$\mathbf{E}(\ell(\hat{y}_m,y_m)) - \inf_{f\in F} \text{er}_P(f) < 6\alpha + (1+2|\Phi(x)|)e^{-\alpha^2 m}.$$

Let $\alpha = \left(\frac{2c_1 d}{m}\right)^{1/3} + \sqrt{\frac{2\ln m}{m}}$. Theorem 1 implies that there is are constants c_1 and m_0 such that for all $m \geq m_0$,

$$\mathbf{E}(\ell(\hat{y}_m,y_m)) - \inf_{f\in F} \text{er}_P(f) < 6\alpha + \exp(c_1 d/\alpha - \alpha^2 m). \tag{4}$$

The following sequence of implications are immediate:

$$\begin{aligned}
&\alpha \geq \left(\frac{2c_1 d}{m}\right)^{1/3} \text{ and } \alpha \geq \sqrt{\frac{2\ln m}{m}} \\
&\alpha^2 m/2 \geq c_1 d/\alpha \text{ and } \alpha^2 m/2 \geq \ln\frac{1}{\alpha} \\
&\alpha^2 m \geq c_1 d/\alpha + \ln\frac{1}{\alpha} \\
&\exp(c_1 d/\alpha - \alpha^2 m) \leq \alpha.
\end{aligned}$$

Applying (4), we get

$$\mathbf{E}(\ell(\hat{y}_m, y_m)) - \inf_{f \in F} \mathrm{er}_P(f) < 7\alpha.$$

Substituting the value of α completes the proof. □

Armed with Theorem 3, straightforward application of known techniques [14] (almost identical to the last paragraph of the proof of Theorem 21 in [4]) gets us the rest of the way to prove Theorem 2.

Proof (for Theorem 2): Theorem 3, together with Fubini's theorem, implies that there is a learning algorithm whose hypothesis h satisfies $\mathbf{E}(\mathrm{er}_P(h) - \inf_{f \in F} \mathrm{er}_P(f)) = c(d/t)^{1/3}$, where this expectation is with respect to t random examples. Markov's inequality implies that $\mathbf{Pr}(\mathrm{er}_P(h) - \inf_{f \in F} \mathrm{er}_P(f) > 2c(d/t)^{1/3}) \leq 1/2$. If we run the algorithm repeatedly $\approx \log_2 2/\delta$ times using t examples each time, with probability $1 - \delta/2$, one of resulting hypotheses will satisfy $\mathrm{er}_P(h) - \inf_{f \in F} \mathrm{er}_P(f) \leq 2c(d/t)^{1/3}$. Hoeffding's inequality implies that $\mathrm{poly}(t, 1/\delta)$ additional examples are sufficient to test of all the returned hypotheses, and find one that satisfies $\mathrm{er}_P(h) - \inf_{f \in F} \mathrm{er}_P(f) \leq 4c(d/t)^{1/3}$ with probability $1 - \delta$. □

2.5 Rounding

One might hope that all the $*$'s in a class of $\{0, *, 1\}$ valued functions can be "rounded" to 0 or 1 without increasing its VC-dimension. This would lead to a better bound than Theorem 3, and perhaps a simpler proof of Theorem 2. Unfortunately, it is not true.

Proposition 1. *There is a set X, and a set F of functions from X to $\{0, *, 1\}$ such that for any set G of functions from X to $\{0, 1\}$ that 0-covers F,* $\mathrm{VCdim}(G) > \mathrm{VCdim}(F)$.

Proof: Define F as in Figure 1. By inspection, $\mathrm{VCdim}(F) = 1$. It is not possible

	x_1	x_2	x_3
f_1	0	0	0
f_2	0	$*$	1
f_3	1	0	$*$
f_4	$*$	1	0
f_5	1	1	1

Fig. 1. F from Proposition 1 in table form.

for a single function to 0-cover two elements of F, since each pair of functions in F differ on some domain element on which neither evaluates to $*$. Thus, any G that 0-covers F must have $|G| \geq 5$, and therefore, by the Sauer-Shelah Lemma, $\mathrm{VCdim}(G) > 1$. □

3 Real-valued hypotheses

For a function f from X to $[0,1]$, a real threshold r and a non-negative margin γ, define $\psi_{r,\gamma}(f) : X \to \{0, *, 1\}$ to indicate whether $f(x)$ is above or below r by a margin γ as follows

$$(\psi_{r,\gamma}(f))(x) = \begin{cases} 1 \text{ if } f(x) \geq r + \gamma \\ 0 \text{ if } f(x) \leq r - \gamma \\ * \text{ if } |f(x) - r| < \gamma. \end{cases}$$

For a class F of functions from X to $[0,1]$, let $\psi_{r,\gamma}(F) = \{\psi_{r,\gamma}(f) : f \in F\}$. Let $\mathrm{fatV}_F(\gamma) = \max_r \mathrm{VCdim}(\psi_{r,\gamma}(F))$. (This notion of dimension was proposed by Alon, et al [1].)

For this section, let us broaden the notion of an example to be an arbitrary element of $X \times [0,1]$ and redefine a learning and a prediction strategy accordingly. For a probability distribution P over $X \times [0,1]$ and a function h from X to $[0,1]$, let $\mathrm{er}_P(h) = \mathbf{E}_{(x,y)\sim P}(|h(x) - y|)$. For $\epsilon > 0$, we then say that a set F of functions from X to $[0,1]$ is *agnostically learnable to within* ϵ if there is a learning strategy A such that, for all $\delta > 0$, there is a natural number m such that, for any probability distribution P over $X \times [0,1]$, if m examples are drawn independently at random according to P, and the resulting sample is passed to A which outputs a hypothesis h, then, with probability at least $1-\delta$, $\mathrm{er}_P(h) \leq (\inf_{f\in F} \mathrm{er}_P(f)) + \epsilon$.

The following is the main result of this section.

Theorem 4. *For any set F of functions from X to $[0,1]$, if there is an $\alpha > 0$ such that $\mathrm{fatV}_F(\epsilon - \alpha)$ is finite, then F is agnostically learnable to within ϵ.*

This improves on the sufficient condition $(\exists \alpha > 0, \mathrm{fatV}_F(\epsilon/2 - \alpha) < \infty)$ from [4] by a factor of two on the scale at which the dimension of F is examined. The finiteness of $\mathrm{fatV}_F(\epsilon + \alpha)$ for some positive α has been shown to be necessary [4]. It implies a similar improvement on the sufficient condition stated in terms of Kearns and Schapire's [16] *fat-shattering function* [1, 4], closing a factor of two gap there as well.

Say that a set F of functions from X to $[0,1]$ is agnostically predictable to within $\epsilon > 0$ if there is a sample size m and a prediction strategy A such that, for any probability distribution P over $X \times [0,1]$, if $(x_1, y_1), ..., (x_m, y_m)$ are drawn independently at random according to P, and $(x_1, y_1), ..., (x_{m-1}, y_{m-1}), x_m$ are passed to A, which outputs $\hat{y}_m$, then $\mathbf{E}(|\hat{y}_m - y_m|) - \inf_{f\in F} \mathrm{er}_P(f) \leq \epsilon$.

We will make use of the following lemma, implicit in [14, 4].

Lemma 3. *For any X, and any set F of functions from X to $[0,1]$, if F is agnostically predictable to within $\epsilon > 0$, then F is agnostically learnable to within any $\epsilon' > \epsilon$.*

This enables us to prove Theorem 4 by analyzing a prediction strategy.

As in the previous section, we wanted a covering lemma whose proof doesn't go via packing; loosely speaking, here this is because one loses a factor of two converting between packing and covering.

3.1 Small covers

Choose a finite set Z. For functions f and g from Z to $[0,1]$, let $\ell_1(f,g) = \frac{1}{|Z|}\sum_{z\in Z}|f(z)-g(z)|$. Say that a set G of functions from Z to $[0,1]$ ϵ-covers a set F of functions from Z to $[0,1]$ if for every $f \in F$, there is a $g \in G$ for which $\ell_1(f,g) \leq \epsilon$. Let $\mathcal{N}(\epsilon, F)$ be the size of the smallest ϵ-cover of F.

For $\alpha > 0$ and $u \in \mathbf{R}$, let $Q_\alpha(u)$ denote the quantized version of u, with quantization width α. That is, define $Q_\alpha(u) = \alpha\lfloor u/\alpha \rfloor$. Let $Q_\alpha([0,1]) = \{Q_\alpha(u) : u \in [0,1]\}$. For a function f from Z to $\mathbf{R}$, define $Q_\alpha(f) : Z \to \mathbf{R}$ by $(Q_\alpha(f))(x) = Q_\alpha(f(x))$. Finally, for a set F of such functions, define $Q_\alpha(F) = \{Q_\alpha(f) : f \in F\}$.

Lemma 4. *For any set F of functions from Z to $[0,1]$, any $\epsilon > 0$, and $\alpha < \epsilon/2$, $\mathcal{N}(\epsilon, F) \leq \mathcal{N}(\epsilon - \alpha, Q_\alpha(F))$.*

For functions h and f from X to $[0,1]$, and a probability distribution D over Z, define the *error* of h with respect to f and D, to be $\mathrm{er}_{f,D}(h) = \mathbf{Pr}_{z\sim D}(|h(z) - f(z)|)$.

Lemma 5 ([4]). *Choose a set F of functions from Z to $[0,1]$. There is a mapping A from finite sequences of elements of $Z \times [0,1]$ to hypotheses such that, for any probability distribution D over Z, for any $f \in F$, and for any positive integer m, if $z_1, ..., z_t$ are chosen independently at random according to D, A is applied to $(z_1, f(z_1)), ..., (z_t, f(z_t))$, and h is the resulting hypothesis, then $\mathbf{E}(\mathrm{er}_{f,D}(h)) \leq \gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t+1}$.*

Lemma 6. *Let $m = |Z|$. Choose $0 < \gamma < \epsilon \leq 1$, $b \in \mathbf{N}$, and a set F of functions from Z to $Q_{1/b}([0,1])$. There is a set G of $\{0,1\}$-valued functions and a subset F' of F such that*

- $|F'| \geq |F|/2$,
- *G ϵ-covers F', and*
- $|G| \leq (b+1)^{\left\lceil \frac{2\mathrm{fatV}_F(\gamma)}{\epsilon-\gamma}\right\rceil}$.

Proof: Define A as in Lemma 5. Let D be the uniform distribution over X, and let P be the uniform distribution over F. Choose a positive integer t (its value will be set later). Suppose that $z_1, ..., z_t$ are chosen independently at random uniformly according to D, f is chosen independently according to P, $(z_1, f(z_1)), ..., (z_t, f(z_t))$ are passed to A; let $A(z_1, ..., z_t; f)$ be A's hypothesis. Then Lemma 1 says that

$$\forall f \in F, \mathbf{E}_{z_1,...,z_t\sim D^t}(\mathrm{er}_{f,D}(A(z_1, ..., z_t; f))) < \gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t}.$$

Arguing as in the proof of Lemma 2, we have

$$\exists z_1, ..., z_t,\ \mathbf{Pr}_{f\sim P}\left(\mathrm{er}_{f,D}(A(z_1, ..., z_t; f)) > \gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t}\right) \leq 1/2. \quad (5)$$

Choose such a sequence $z_1, ..., z_t$. Then if $G = \{A(z_1, ..., z_t; f) : f \in F\}$, and

$$F' = \left\{ f \in F : \mathrm{er}_{f,D}(A(z_1, ..., z_t; f)) \leq \gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t} \right\},$$

then G $(\gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t})$-covers F', and, by (5), $|F'| \geq |F|/2$. Suppose $t = \left\lceil \frac{2\mathrm{fatV}(\gamma)}{\epsilon - \gamma} \right\rceil$; then $\gamma + \frac{2\mathrm{fatV}_F(\gamma)}{t} \leq \epsilon$. There are only $(b+1)^t$ possible inputs to A with instances $z_1, ..., z_t$. Thus, $|H| \leq (b+1)^t$, completing the proof. □

Theorem 5. *Suppose $m = |Z|$. Choose a set F of functions from Z to $[0,1]$, and ϵ and $\alpha \in \mathbf{R}$ for which $0 < \alpha < \epsilon \leq 1$. Then $\mathcal{N}(\epsilon, F) \leq (m \log_2(3/\alpha + 2))(3/\alpha + 2)^{(6/\alpha+1)\mathrm{fatV}_F(\epsilon-\alpha)}$.*

Proof: Let $b = \lceil 3/\alpha \rceil$. Then $\mathrm{fatV}(\epsilon - 3/b) \leq \mathrm{fatV}(\epsilon - \alpha)$. Construct a sequence $G_1, G_2, ..., G_{\lceil \log_2 |Q_{1/b}(F)| \rceil}$ of sets of functions from Z to $\{0,1\}$ by repeatedly applying Lemma 6 to $(\epsilon - 1/b)$-cover at last half of the remaining functions in $Q_{1/b}(F)$, and then deleting the covered functions. Let $G = \cup G_i$. Then G $(\epsilon - 1/b)$-covers $Q_{1/b}(F)$, and

$$\begin{aligned} |G| &\leq (\log_2 |Q_{1/b}(F)|)(b+1)^{2b\mathrm{fatV}_{Q_{1/b}(F)}(\epsilon - 2/b)} \\ &\leq (m \log_2(b+1))(b+1)^{2b\mathrm{fatV}_{Q_{1/b}(F)}(\epsilon - 2/b)} \\ &\leq (m \log_2(b+1))(b+1)^{2b\mathrm{fatV}_F(\epsilon - 3/b)}, \end{aligned}$$

since, straight from the definitions, $\mathrm{fatV}_{Q_{1/b}(F)}(\epsilon - 2/b) \leq \mathrm{fatV}_F(\epsilon - 3/b)$. Applying Lemma 4 completes the proof. □

3.2 Learning

Like the proof of Theorem 3, the following proof closely follows that of Theorem 21 of [4], except that it appeals to the new Theorem 5.

Theorem 6. *Choose a set F of functions from X to $[0,1]$, and $0 < \epsilon \leq 1$. If there exists $\alpha > 0$ such that $\mathrm{fatV}_F(\epsilon - \alpha)$ is finite, then F is agnostically predictable to within ϵ.*

Proof: Let $d = \mathrm{fatV}_F(\epsilon - \alpha)$, $\kappa = \alpha/3$, and $\beta = \alpha/15$.

Choose a function Φ that maps from X^m to the set of finite subsets of $[0,1]^m$ such that, for any $(x_1, ..., x_m) \in X^m$, $\Phi(x_1, ..., x_m)$ is one of the smallest sets that $(\epsilon - 2\kappa)$-covers $\{(f(x_1), ..., f(x_m)) : f \in F\}$ and $\Phi(x_1, ..., x_m)$ is invariant under permutations of its arguments. (Recall once again that here we are viewing an element of $[0,1]^m$ as a function from $\{1, ..., m\}$ to $[0,1]$.)

Consider the prediction strategy A that, given input

$$(x_1, y_1), ..., (x_{m-1}, y_{m-1}), x_m,$$

chooses $\hat{y} = (\hat{y}_1, ..., \hat{y}_m)$ from among the elements of $\Phi(x_1, ..., x_m)$ in order to minimize $\sum_{i=1}^{m/2} |\hat{y}_i - y_i|$, and outputs $\hat{y}_m$.

For $a \in [0,1]^m$, $b \in [0,1]^m$, define

$$\ell^{\text{first}}(a,b) = \frac{2}{m} \sum_{i=1}^{m/2} |a_i - b_i|,$$

$$\ell^{\text{last}}(a,b) = \frac{2}{m} \sum_{i=m/2+1}^{m} |a_i - b_i|,$$

and

$$\ell^{\text{all}}(a,b) = \frac{1}{m} \sum_{i=1}^{m} |a_i - b_i|.$$

Choose any probability distribution P over $X \times [0,1]$, and an even positive integer m. Suppose $(x_1,y_1),...,(x_m,y_m)$ are drawn independently at random according to P, and $(x_1,y_1),...,(x_{m-1},y_{m-1})$ and x_m are given to A, which outputs $\hat{y}_m$. Let $x = (x_1,...,x_m)$ and $y = (y_1,...,y_m)$. Choose $f^* \in F$ that satisfies $\text{er}_P(f^*) \leq \inf_{f \in F} \text{er}_P(f) + \beta$. Since $\Phi(x)$ $\epsilon - 2\kappa$-covers $\{(f(x_1),...,f(x_m)) : f \in F\}$

$$\exists t^* \in \Phi(x_1,...,x_m), \ell^{\text{all}}(t^*,y) \leq \epsilon - 2\kappa + \ell^{\text{all}}((f^*(x_1),...,f^*(x_m)),y). \quad (6)$$

Applying the Hoeffding bound,

$$\mathbf{Pr}\left(\ell^{\text{all}}((f^*(x_1),...,f^*(x_m)),y) > \text{er}_P(f^*) + \beta\right) \leq e^{-2\beta^2 m}. \quad (7)$$

Now, let U be the uniform distribution over $\{0,1\}^{m/2}$. Then, since Φ is invariant under permutations,

$$\begin{aligned}&\mathbf{Pr}(\exists t \in \Phi(x)\ |\ell^{\text{first}}(t,y) - \ell^{\text{last}}(t,y)| > 2\beta) \\ &\leq \sup_{(x,y)} \mathbf{Pr}_{u \in U}\left(\exists t \in \Phi(x)\ \left|\tfrac{2}{m}\textstyle\sum_{i=1}^{m/2} u_i \left(\ell(t_i,y_i) - \ell(t_{i+m/2},y_{i+m/2})\right)\right| > 2\beta\right)\end{aligned}$$

For any fixed $t \in \Phi(x)$, Hoeffding's inequality implies

$$\mathbf{Pr}_{u \in U}\left(\left|\frac{2}{m}\sum_{i=1}^{m/2} u_i \left(\ell(t_i,y_i) - \ell(t_{i+m/2},y_{i+m/2})\right)\right| > 2\beta\right) \leq 2e^{-\beta^2 m}.$$

So with probability at least $1 - |\Phi(x)|2e^{-\beta^2 m}$, for all t in $\Phi(x)$,

$$|\ell^{\text{first}}(t,y) - \ell^{\text{last}}(t,y)| \leq 2\beta.$$

This implies

$$|\ell^{\text{first}}(t,y) - \ell^{\text{all}}(t,y)| \leq \beta$$

and

$$|\ell^{\text{last}}(t,y) - \ell^{\text{all}}(t,y)| \leq \beta.$$

So, combining with (7), with probability at least $1-(1+2|\Phi(x)|)e^{-\beta^2 m}$, the $\hat{y} \in \Phi(x)$ with minimal $\ell^{\text{first}}(\hat{y}, y)$ satisfies

$$\begin{aligned}
\ell^{\text{all}}(\hat{y}, y) &\leq \ell^{\text{first}}(\hat{y}, y) + \beta \\
&\leq \ell^{\text{first}}(t^*, y) + \beta \\
&\leq \ell^{\text{all}}(t^*, y) + 2\beta \\
&\leq \ell^{\text{all}}((f^*(x_1), ..., f^*(x_m)), y) + \epsilon - 2\kappa + 2\beta \\
&\leq \text{er}_P(f^*) + \epsilon - 2\kappa + 3\beta
\end{aligned}$$

and hence

$$\begin{aligned}
\ell^{\text{last}}(\hat{y}, y) &\leq \text{er}_P(f^*) + \epsilon - 2\kappa + 4\beta \\
&\leq \inf_{f \in F} \text{er}_P(f) + \epsilon - 2\kappa + 5\beta.
\end{aligned}$$

That is,

$$\mathbf{Pr}\left(\ell^{\text{last}}(\hat{y}, y) > \inf_{f \in F} \text{er}_P(f) + +\epsilon - 2\kappa + 5\beta\right) < (1+2|\Phi(x)|)e^{-\beta^2 m}$$

which implies

$$\mathbf{E}(\ell^{\text{last}}(\hat{y}, y)) - \inf_{f \in F} \text{er}_P(f) < \epsilon - 2\kappa + 5\beta + (1+2|\Phi(x)|)e^{-\beta^2 m}$$

and hence, since any of $(x_{m/2+1}, y_{m/2+1}), ..., (x_m, y_m)$ was equally likely to have been the last,

$$\mathbf{E}(|\hat{y}_m - y_m|) - \inf_{f \in F} \text{er}_P(f) < \epsilon - 2\kappa + 5\beta + (1+2|\Phi(x)|)e^{-\beta^2 m}.$$

Substituting $\kappa/5$ for β,

$$\mathbf{E}(|\hat{y}_m - y_m|) - \inf_{f \in F} \text{er}_P(f) < \epsilon - \kappa + (1+2|\Phi(x)|)e^{-\kappa^2 m/25}. \tag{8}$$

Recall that $d = \text{fatV}(\epsilon - \alpha) = \text{fatV}(\epsilon - 3\kappa)$, and $\Phi(x)$ is a minimum sized $\epsilon - 2\kappa$ cover of $\{(f(x_1), ..., f(x_m)) : f \in F\}$; Theorem 5 and (8) imply that

$$\begin{aligned}
&\mathbf{E}(|\hat{y}_m - y_m|) - \inf_{f \in F} \text{er}_P(f) \\
&\leq \epsilon - \kappa + (1 + 2(m \log_2(3/\kappa + 2))(3/\kappa + 2)^{(6/\kappa+1)d})e^{-\kappa^2 m/25}.
\end{aligned}$$

Thus, if m is large enough, $\mathbf{E}(|\hat{y}_m - y_m|) - \inf_{f \in F} \text{er}_P(f) \leq \epsilon$; this completes the proof. □

Acknowledgements

We'd like to express warm thanks to Peter Bartlett, Shai Ben-David, Nadav Eiron, David Haussler and Yishay Mansour for valuable conversations.

We acknowledge the support of National University of Singapore Academic Research Fund Grant RP252–000–070–107.

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the Association for Computing Machinery*, 44(4):616–631, 1997.
2. M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.
3. P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory*, 44(2):525–536, 1998.
4. P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.
5. P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.
6. S. Ben-David and H. U. Simon. Efficient learning of linear perceptrons. *Advances in Neural Information Processing Systems 14*, 2000.
7. Shai Ben-David, Nadav Eiron, and Hans U. Simon. The computational complexity of densest region detection. *Proceedings of the 2000 Conference on Computational Learning Theory*, 2000.
8. A. Blum, P. Chalasani, S. Goldman, and D. K. Slonim. Learning with unreliable boundary queries. *Proceedings of the 1995 Conference on Computational Learning Theory*, pages 98–107, 1995.
9. A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.
10. K.L. Buescher and P.R. Kumar. Learning stochastic functions by smooth simultaneous estimation. *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 272–279, 1992.
11. N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge University Press, 2000.
12. Y. Freund, R. E. Schapire, Y. Singer, and M.K. Warmuth. Using and combining predictors that specialize. *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing*, 1997.
13. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.
14. D. Haussler, M. Kearns, N. Littlestone, and M. K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95:129–161, 1991.
15. D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.
16. M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994.
17. M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17:115–141, 1994.
18. V. N. Vapnik. *Estimation of Dependencies based on Empirical Data*. Springer Verlag, 1982.
19. V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.