

The sample complexity of learning

INTRODUCTION

Earlier, we proved that if H is finite, then it is potentially learnable. The proof depends critically on the finiteness of H and cannot be extended to provide results for infinite H . However, there are many situations where the hypothesis space is infinite, and it is desirable to extend the theory to cover this case.

The key to extending results on potential learnability to infinite spaces is the observation that what matters is not the cardinality of H , but rather what may be described as its ‘expressive power’. We shall formalise this notion in terms of the *Vapnik-Chervonenkis dimension* of H , a notion originally defined by Vapnik and Chervonenkis (1971), and introduced into learnability theory by Blumer *et al.* (1986, 1989).

We consider the *real perceptron*. This is the simplest type of artificial neural network. It has n inputs and a single active node. The arcs carrying the inputs have real-valued weights $\alpha_1, \alpha_2, \dots, \alpha_n$ and there is a real threshold value θ at the active node. The weighted sum of the inputs is applied to the active node and this node outputs 1 if and only if the weighted sum is at least the threshold value θ .

More precisely, for the real perceptron P_n on n inputs, the set of states is \mathbf{R}^{n+1} . For a state $\omega = (\alpha_1, \alpha_2, \dots, \alpha_n, \theta)$, the corresponding function $h_\omega \in H$, from $X = \mathbf{R}^n$ to $\{0, 1\}$, is given by

$$h_\omega(y) = \begin{cases} 1, & \text{if } \sum_{i=1}^n \alpha_i y_i \geq \theta; \\ 0, & \text{otherwise.} \end{cases}$$

THE GROWTH FUNCTION

Suppose that H is a hypothesis space defined on the example space X , and let $\mathbf{x} = (x_1, x_2, \dots, x_m)$ be a sample of length m of examples from X . We define $\Pi_H(\mathbf{x})$, the *number of classifications of \mathbf{x} by H* , to be the number of distinct vectors of the form

$$(h(x_1), h(x_2), \dots, h(x_m)),$$

2 The sample complexity of learning

as h runs through all hypotheses of H . Note that for any sample \mathbf{x} of length m , $\Pi_H(\mathbf{x}) \leq 2^m$. We define the *growth function* Π_H by

$$\Pi_H(m) = \max \{ \Pi_H(\mathbf{x}) : \mathbf{x} \in X^m \}.$$

In general, it is difficult to find an exact formula for the growth function of a hypothesis space. But we shall define a numerical parameter of a hypothesis space which is easier to estimate than the growth function, and which can be used to provide upper bounds for the growth function.

THE VC DIMENSION

We say that a sample \mathbf{x} of length m is *shattered* by H , or that H *shatters* \mathbf{x} , if $\Pi_H(\mathbf{x}) = 2^m$; that is, if H gives all possible classifications of \mathbf{x} . The examples in \mathbf{x} must be distinct for this to happen. When they are distinct, \mathbf{x} is shattered by H if and only if for any subset S of $E_{\mathbf{x}} = \{x_1, x_2, \dots, x_m\}$, there is some hypothesis h in H such that for $1 \leq i \leq m$,

$$h(x_i) = 1 \iff x_i \in S.$$

The Vapnik-Chervonenkis (VC-) dimension of H is the maximum length of a sample shattered by H ; if there is no such maximum, we say that the VC dimension of H is infinite. Thus,

$$\text{VCdim}(H) = \max \{ m : \Pi_H(m) = 2^m \},$$

where we take the maximum to be infinite if the set is unbounded.

The following simple result on *finite* hypothesis spaces is often useful.

Proposition If H is a finite hypothesis space, then

$$\text{VCdim}(H) \leq \lg |H|.$$

THE VC DIMENSION OF THE REAL PERCEPTRON

Theorem For any positive integer n , let P_n be the real perceptron with n inputs. Then

$$\text{VCdim}(P_n) = n + 1.$$

In state

$$\omega = (\alpha_1, \alpha_2, \dots, \alpha_n, \theta),$$

the function h_ω computed by the perceptron is the $\{0, 1\}$ -function such that

$$h_\omega(y) = 1 \iff \alpha_1 y_1 + \alpha_2 y_2 + \dots + \alpha_n y_n \geq \theta.$$

3 The sample complexity of learning

Thus the set of positive examples of h_ω is the *closed half-space*

$$l_\omega^+ = \left\{ y \in \mathbf{R}^n : \sum_{i=1}^n \alpha_i y_i \geq \theta \right\},$$

bounded by the *hyperplane*

$$l_\omega = \left\{ y \in \mathbf{R}^n : \sum_{i=1}^n \alpha_i y_i = \theta \right\}.$$

The set of negative examples of h_ω is then the *open half-space*

$$l_\omega^- = \left\{ y \in \mathbf{R}^n : \sum_{i=1}^n \alpha_i y_i < \theta \right\}.$$

$C \subseteq \mathbf{R}^n$ is *convex* if, given any two points x, y of S , the line segment between x and y lies entirely in C . More formally, C is convex if given any x, y in C and any real number λ with $0 \leq \lambda \leq 1$, the point $\lambda x + (1 - \lambda)y$ belongs to C .

The intersection of any number of convex sets is again convex and therefore for any non-empty set S of points of \mathbf{R}^n , there is a smallest convex set containing S . This set, denoted by $\text{conv}(S)$, is called the *convex hull* of S ; $\text{conv}(S)$ is the intersection of all convex sets containing S . For example, suppose that S is any finite set of points in the plane \mathbf{R}^2 . Then $\text{conv}(S)$ is the smallest closed region which is bounded by a polygon and which contains S .

Radon's Theorem states: let n be any positive integer, and let E be any set of $n + 2$ points in \mathbf{R}^n . Then there is a non-empty subset S of E such that

$$\text{conv}(S) \cap \text{conv}(E \setminus S) \neq \emptyset.$$

We can now give the proof that the VC-dim of the perceptron is $n + 1$.

Let $\mathbf{x} = (x_1, x_2, \dots, x_{n+2})$ be any sample of length $n + 2$. Suppose that the set $E_{\mathbf{x}}$ of examples in \mathbf{x} consists of $n + 2$ distinct points in \mathbf{R}^n . By Radon's Theorem, there is a non-empty subset S of $E_{\mathbf{x}}$ such that

$$\text{conv}(S) \cap \text{conv}(E_{\mathbf{x}} \setminus S) \neq \emptyset.$$

Suppose that there is h_ω in P_n such that S is the set of positive examples of h_ω in $E_{\mathbf{x}}$. Then we have

$$S \subseteq l_\omega^+, \quad E_{\mathbf{x}} \setminus S \subseteq l_\omega^-.$$

Since open and closed half-spaces are convex subsets of \mathbf{R}^n , we also have

$$\text{conv}(S) \subseteq l_\omega^+, \quad \text{conv}(E_{\mathbf{x}} \setminus S) \subseteq l_\omega^-.$$

Therefore

$$\text{conv}(S) \cap \text{conv}(E_{\mathbf{x}} \setminus S) \subseteq l_\omega^+ \cap l_\omega^- = \emptyset.$$

4 The sample complexity of learning

We deduce that no such h_ω exists and therefore that \mathbf{x} is not shattered by P_n . Thus no sample of length $n + 2$ is shattered by P_n and $\text{VCdim}(P_n) \leq n + 1$.

It remains to prove the reverse inequality. Let o denote the origin of \mathbf{R}^n and, for $1 \leq i \leq n$, let e_i be the point with a 1 in the i th coordinate and all other coordinates 0. We P_n shatters the sample

$$\mathbf{x} = (o, e_1, e_2, \dots, e_n)$$

of length $n + 1$.

Suppose $S \subseteq E_{\mathbf{x}} = \{o, e_1, \dots, e_n\}$. For $i = 1, 2, \dots, n$, let

$$\alpha_i = \begin{cases} 1, & \text{if } e_i \in S; \\ -1, & \text{if } e_i \notin S; \end{cases}$$

and let

$$\theta = \begin{cases} -1/2, & \text{if } o \in S; \\ 1/2, & \text{if } o \notin S. \end{cases}$$

Then it is straightforward to verify that if ω is the state

$$\omega = (\alpha_1, \alpha_2, \dots, \alpha_n, \theta)$$

of P_n then the set of positive examples of h_ω in $E_{\mathbf{x}}$ is precisely S . Therefore \mathbf{x} is shattered by P_n and, consequently, $\text{VCdim}(P_n) \geq n + 1$.

SAUER'S LEMMA RELATING GROWTH FUNCTION TO VC-DIMENSION

Assume that H has finite VC dimension. We have the following theorem, due to Sauer (1972) and Shelah (1972) independently. We shall refer to it as *Sauer's Lemma*.

Theorem (Sauer's Lemma) Let $d \geq 0$ and $m \geq 1$ be given integers and let H be a hypothesis space with $\text{VCdim}(H) = d$. Then

$$\Pi_H(m) \leq 1 + \binom{m}{1} + \binom{m}{2} + \dots + \binom{m}{d},$$

where the binomial numbers are defined by

$$\binom{m}{i} = \frac{m(m-1)\dots(m-i+1)}{1 \cdot 2 \dots i}.$$

The definition of the binomial numbers means that $\binom{a}{b}$ is zero whenever $b > a \geq 1$. Thus for $m \leq d$, the result asserts only that

$$\Pi_H(m) \leq 1 + \binom{m}{1} + \dots + \binom{m}{m} + 0 + 0 + \dots + 0 = 2^m,$$

which is trivial; we already know that Π_H takes these values in this range. However, when m is greater than d , the sum

$$\Phi(d, m) = 1 + \binom{m}{1} + \binom{m}{2} + \dots + \binom{m}{d}$$

is strictly less than 2^m : indeed, it follows from the explicit formula for the binomial numbers that it is a polynomial function of m with degree d . The following result makes this clear.

Proposition For all $m \geq d \geq 1$,

$$\Phi(d, m) < \left(\frac{em}{d}\right)^d,$$

where e is the base of natural logarithms.

In conjunction with Sauer's Lemma, this last result implies that when $\text{VCdim}(H) = d$, we have

$$\Pi_H(m) < \left(\frac{em}{d}\right)^d$$

for $m \geq d$. We shall see that this result is very significant, because it gives an explicit polynomial bound for Π_H as a function of m .

VC DIMENSION AND POTENTIAL LEARNABILITY

Our first result is that finite VC dimension is necessary for potential learnability.

Theorem If a hypothesis space has infinite VC dimension then it is not potentially learnable.

The converse of the preceding theorem is also true: finite VC dimension is sufficient for potential learnability. This result can be traced back to the statistical researches of Vapnik and Chervonenkis (1971).

Suppose that the hypothesis space H is defined on the example space X , and let t be any target concept in H , μ any probability distribution on X and ϵ any real number with $0 < \epsilon < 1$. The objects t, μ, ϵ are to be thought of as fixed, but arbitrary, in what follows. Define

$$Q_m^\epsilon = \{\mathbf{x} \in X^m \mid H[\mathbf{x}, t] \cap B_\epsilon \neq \emptyset\}.$$

The probability of choosing a training sample for which there is a consistent, but ϵ -bad, hypothesis is

$$\mu^m \{ \mathbf{s} \in S(m, t) \mid H[\mathbf{s}] \cap B_\epsilon \neq \emptyset \},$$

which is, by definition, $\mu^m(Q_m^\epsilon)$. Thus, in order to show that H is potentially learnable, it suffices to find an upper bound $f(m, \epsilon)$ for $\mu^m(Q_m^\epsilon)$ which is independent of both t and μ and which tends to 0 as m tends to infinity. For if there is such a bound then, given any δ between 0 and 1, we can use the fact that $f(m, \epsilon)$ tends to 0 to find m_0 such that for all $m \geq m_0$, $f(m, \epsilon) < \delta$. The value of m_0 depends on δ and ϵ but is independent of t and μ . So we have the $m_0(\delta, \epsilon)$ required in the definition of potential learnability.

Note that the m_0 thus obtained is also an upper bound for the sample complexity of any consistent learning algorithm for H . The hard part of the proof is to find the upper bound $f(m, \epsilon)$.

Theorem Suppose that H is a hypothesis space defined on an example space X , and that t , μ , and ϵ are arbitrary, but fixed. Then

$$\mu^m \{ \mathbf{s} \in S(m, t) \mid H[\mathbf{s}] \cap B_\epsilon \neq \emptyset \} < 2 \Pi_H(2m) 2^{-\epsilon m/2}$$

for all positive integers $m \geq 8/\epsilon$.

The right-hand side is the bound $f(m, \epsilon)$ for $\mu^m(Q_m^\epsilon)$, as postulated above. We have to show that it tends to zero as $m \rightarrow \infty$. If H has finite VC dimension then, by Sauer's Lemma, $\Pi_H(2m)$ is bounded by a polynomial function of m , and therefore $f(m, \epsilon)$ is eventually dominated by the negative exponential term. Thus the right-hand side tends to 0 as m tends to infinity and, by the above discussion, this establishes potential learnability for spaces of finite VC dimension.

SAMPLE COMPLEXITY OF CONSISTENT ALGORITHMS

We have seen that if a hypothesis space H has finite VC dimension, then H is potentially learnable. In other words, given a confidence parameter δ and an accuracy parameter ϵ ($0 < \delta, \epsilon < 1$), there is a sample length $m_0 = m_0(H, \delta, \epsilon)$ such that

$$m \geq m_0 \implies \mu^m \{ \mathbf{s} \in S(m, t) \mid H[\mathbf{s}] \cap B_\epsilon = \emptyset \} > 1 - \delta,$$

for any probability distribution μ on X and any target concept $t \in H$. It follows that any consistent learning algorithm L for H is pac and, further, that any $m_0(H, \delta, \epsilon)$ for which the above condition holds is an upper bound on the sample complexity $m_L(H, \delta, \epsilon)$.

Recall that in Chapter 4 we showed that if H is a finite hypothesis space and L is a consistent learning algorithm for H , then L is pac and

$$m_L(H, \delta, \epsilon) \leq \left\lceil \frac{1}{\epsilon} \ln \left(\frac{|H|}{\delta} \right) \right\rceil.$$

The upper bound for $m_L(H, \delta, \epsilon)$ which we now present depends on the VC dimension of H , rather than the cardinality of H .

Theorem Suppose that hypothesis space H has VC dimension $d \geq 1$. Then any consistent learning algorithm L for H is pac, with sample complexity

$$m_L(H, \delta, \epsilon) \leq \left\lceil \frac{4}{\epsilon} \left(d \lg \left(\frac{12}{\epsilon} \right) + \lg \left(\frac{2}{\delta} \right) \right) \right\rceil.$$

LOWER BOUNDS ON SAMPLE COMPLEXITY

We now present a result of Ehrenfeucht *et al.* (1989) which provides a lower bound on the sample complexity of *any* pac learning algorithm for a hypothesis space of finite VC dimension.

Theorem For any hypothesis space H of VC dimension $d \geq 1$, and for any pac learning algorithm L for H ,

$$m_L(H, \delta, \epsilon) > \frac{d-1}{32\epsilon},$$

for $\delta \leq 1/100$ and $\epsilon \leq 1/8$.

Corollary If a hypothesis space H has infinite VC dimension then there is no pac learning algorithm for H .

These results support the claim that the VC dimension is a good measure of the ‘expressive power’ of a hypothesis space H : the greater the VC dimension of H , the greater must be the sample complexity for pac learning H .

Another useful result concerning lower bounds is the following, due to Blumer *et al.* (1989). This bound involves ϵ and δ , but is independent of the VC dimension of the hypothesis space. It applies to *non-trivial* hypothesis spaces. By this we simply mean hypothesis spaces which consist of more than two hypotheses.

Theorem Suppose that L is any pac learning algorithm for the non-trivial hypothesis space H . Then

$$m_L(H, \delta, \epsilon) > \frac{(1-\epsilon)}{\epsilon} \ln \left(\frac{1}{\delta} \right),$$

for any $0 < \delta, \epsilon < 1$.

COMPARISON OF SAMPLE COMPLEXITY BOUNDS

Many of the preceding results can be generalised to deal with the case in which the concept space and the hypothesis space are different.

Theorem Let C be a concept space and H a hypothesis space, and suppose that H has finite VC dimension at least 1. If L is any consistent learning algorithm for (C, H) , then L is pac and the sample complexity of L satisfies

$$m_L(C, \delta, \epsilon) \leq \left\lceil \frac{4}{\epsilon} \left(\text{VCdim}(H) \lg \left(\frac{12}{\epsilon} \right) + \lg \left(\frac{2}{\delta} \right) \right) \right\rceil,$$

for any δ and ϵ .

Theorem Let C be a concept space and H a hypothesis space, such that C has VC dimension at least 1. Suppose that L is any pac learning algorithm for (C, H) . Then the sample complexity of L satisfies

$$m_L(C, \delta, \epsilon) > \max \left(\frac{\text{VCdim}(C) - 1}{32\epsilon}, \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right),$$

for all $\epsilon \leq 1/8$ and $\delta \leq 1/100$.

The significant factors in the bounds are the VC dimensions of the concept and the hypothesis spaces, and the parameters ϵ and δ . To simplify matters, and to suppress the less important constant factors in these expressions, we can use the O -notation and the Ω -notation. We write $f = O(g)$ when there is some constant C such that for all relevant values of x $f(x) \leq Cg(x)$. Similarly, we write $f = \Omega(g)$ when there is some positive constant K such that $f(x) \geq Kg(x)$.

Using these notations we can re-state the sample complexity bounds, remembering that the functions involved depend on the VC dimension of C or H and the accuracy and confidence parameters.

- If L is pac then C must have finite VC dimension, and

$$m_L(C, \delta, \epsilon) = \Omega \left(\frac{\text{VCdim}(C)}{\epsilon} + \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right).$$

- If H has finite VC dimension and L is consistent then L is pac, and

$$m_L(C, \delta, \epsilon) = O \left(\frac{\text{VCdim}(H)}{\epsilon} \ln \left(\frac{1}{\epsilon} \right) + \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right).$$

- If H is finite and L is consistent then L is pac and

$$m_L(C, \delta, \epsilon) = O \left(\frac{1}{\epsilon} \ln |H| + \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right).$$

In the case when $C = H$, the VC dimension d is finite, and L is consistent, we have the ‘lower’ and ‘upper’ bounds

$$m_L(H, \delta, \epsilon) = \Omega \left(\frac{d}{\epsilon} + \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right);$$

$$m_L(H, \delta, \epsilon) = O \left(\frac{d}{\epsilon} \ln \left(\frac{1}{\epsilon} \right) + \frac{1}{\epsilon} \ln \left(\frac{1}{\delta} \right) \right).$$

In general, the factor $\ln(1/\epsilon)$ which distinguishes the upper bound from the lower bound is unavoidable. Results of Haussler, Littlestone and Warmuth (1988) show that, for every $d \geq 1$ there is a hypothesis space H_d and a consistent learning algorithm L for H_d with sample complexity meeting the upper bound. On the other hand, it is an open problem to decide whether for every d and for every concept space C of VC dimension d , there is *some* hypothesis space H and *some* (C, H) learning algorithm L for which the sample complexity meets the lower bound.