Chapter 8: Learning and the VC Dimension

8.1 INTRODUCTION

In the previous chapter we discussed the theory of VC dimension, with the promis that this theory would prove useful in the study of learning. The results to b proved in this chapter fulfil that promise. We show that, for any hypothesis spac H, the condition that H has finite VC dimension is both necessary and sufficien for potential learnability. Thus we have a complete characterisation of potentiall learnable hypothesis spaces: they are precisely those of finite VC dimension.

The details of this characterisation provide a general upper bound for the sampl complexity of a consistent learning algorithm, when the hypothesis space is potentiall learnable. We shall also give two general lower bounds for the sample complexity o pac learning algorithms, one in terms of the VC dimension and accuracy, the othe in terms of confidence and accuracy.

8.2 VC DIMENSION AND POTENTIAL LEARNABILITY

We shall find it useful to introduce some slight elaborations of our standard notation We use the notation $\mathbf{s} = (\mathbf{x}, \mathbf{b})$ for the training sample

$$\mathbf{s} = ((x_1, b_1), (x_2, b_2), \ldots, (x_m, b_m))$$

in $(X \times \{0,1\})^m$. If t is a target concept and $\mathbf{s}$ is a training sample for t (that is $b_i = t(x_i)$ for each i), then we denote $\mathbf{s}$ by $(\mathbf{x}, t(\mathbf{x}))$. This notation emphasises th fact that, when $\mathbf{s}$ belongs to the set $S(m,t)$ of training samples of length m for t, onl the values of t on the elements of the sample $\mathbf{x}$ are given. However, for the sake o compactness, we shall denote the subset of H which agrees with $\mathbf{s}$ by $H[\mathbf{x}, t]$, rathe than $H[(\mathbf{x}, t(\mathbf{x}))]$.

Given $\mathbf{s} = (\mathbf{x}, \mathbf{b})$, the *observed error* of a hypothesis $h \in H$ on $\mathbf{s}$ is defined to be

$$\mathrm{er}_{\mathbf{s}}(h) = \frac{1}{m} |\{i : h(x_i) \neq b_i\}| .$$

Note that $H[\mathbf{s}]$ is the set of hypotheses having observed error zero on $\mathbf{s}$. If $\mathbf{s} = (\mathbf{x}, t(\mathbf{x})$

to r from the fact that it is 'small enough'.

This discussion suggests that, for any hypothesis space $H = \bigcup H_r$ and for any learning algorithm L for H, we should define the *effective hypothesis space* $L(m, H_r)$ to be the set of all hypotheses $L(\mathbf{s})$ obtained as $\mathbf{s}$ ranges through all training samples of length m for hypotheses in H_r,

$$L(m, H_r) = \bigcup_{t \in H_r} \{L(\mathbf{s}) \mid \mathbf{s} \in S(m, t)\}.$$

Thus the Occam algorithms for boolean spaces are consistent learning algorithms which have effective hypothesis spaces with 'small enough' cardinalities. The appropriate generalisation to general (and, in particular, real) hypothesis spaces, is to define an Occam algorithm to be a consistent learning algorithm for which the effective hypothesis spaces have 'small enough' VC dimension. Following Blumer *et al.* (1989), we make the following definition.

We say that a learning algorithm L for H is *Occam* with respect to the representation $\Omega \to H$ if

- L is consistent;
- $\text{VCdim}(L(m, H_r)) \leq m^\alpha r^\beta$, where $0 < \alpha < 1$ and $\beta \geq 1$ are constants.

As for boolean spaces, we have the following result.

Theorem 9.6.1 Let H be a space of real or boolean hypotheses having representation $\Omega \to H$. If L is an Occam learning algorithm (with respect to the representation) then, for each r, L is a pac learning algorithm for (H_r, H), with sample complexity $m_L(H_r, \delta, \epsilon)$ polynomial in r, δ^* and ϵ^{-1}.

Proof Let $t \in H_r$ be a given target concept, μ any distribution on X, and δ and ϵ given confidence and accuracy parameters. Consider these quantities as fixed but arbitrary in what follows. For convenience, denote $L(m, H_r)$ by H^*. By definition of the effective hypothesis space, L is a learning algorithm for (H_r, H^*). It is easy to see that Proposition 8.2.3 can be modified to yield

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

We are given that H^* has VC dimension at most $D = m^\alpha r^\beta$. If $m > D$ then, by Sauer's Lemma, the quantity on the right-hand side of the inequality is less than

$$2\left(\frac{2em}{D}\right)^D 2^{-\epsilon m/2}.$$

Let U denote the collection of all subsets of $\mathbf{R}$ which can be expressed as a finite union of closed intervals, and let $J = \{\chi_A \mid A \in U\}$, the *interval union space.*

In order to show that VCdim(J) is infinite, let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ be any sample of distinct points in $\mathbf{R}$, and let $E_{\mathbf{x}}$ denote the corresponding set of examples. Given any $S \subseteq E_{\mathbf{x}}$ we can construct a set $A \in U$ such that $S \subseteq A$ and $(E_{\mathbf{x}} \setminus S) \cap A = \emptyset$, as follows. For each $x_i \in S$ let A_i be a closed interval which contains x_i but no other element of $E_{\mathbf{x}}$, and let A be the union of all such A_i. The set A is a finite union of closed intervals, and χ_A is 1 on S and 0 on $E_{\mathbf{x}} \setminus S$. In other words, J shatters $\mathbf{x}$. Since this argument works for any finite sample, of whatever length, we conclude that VCdim(J) is infinite. □

Note that the space H constructed in this example is contained in the space of (characteristic functions of) closed sets in $\mathbf{R}$. Thus the latter space also has infinite VC dimension. It follows from Theorem 8.2.1 that neither space is 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) (see also Vapnik (1982)). The work of Blumer *et al.* (1986, 1989), showed that it is one of the key results in Computational Learning Theory. The proof is rather involved, and the details will be given in the next section. For the moment, we shall describe only the underlying ideas.

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 (Section 3.2), $\mu^m(Q_m^\epsilon)$. 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 $f(m, \epsilon)$. In the next section we shall prove the following result, which, in this specific form, is due to Blumer *et al.* (1986, 1989), and generalises a result of Haussler and Welzl (1987).

Proposition 8.2.3 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 $m \to \infty$. 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.

8.3 PROOF OF THE FUNDAMENTAL THEOREM

In this section, we present a proof of the key result that finite VC dimension implies potential learnability. The proof is rather involved, and it is worth giving first a very informal explanation of the method.

We aim to bound the probability that a given sample of length m is 'bad', in the sense that there is some hypothesis which is consistent with the target concept on the sample but which has actual error greater than ϵ. We transform this problem into a slightly more manageable one involving samples of length $2m$. For such a sample, the sub-sample $\mathbf{x}$ comprising the first half of the sample may be thought of as a randomly drawn sample of length m, while the second half may be thought of as a 'testing' sample on which to evaluate the performance of a hypothesis consistent with the target concept on $\mathbf{x}$. We obtain a bound on the probability that some hypothesis consistent with the target on the first half of the sample is 'bad', in the sense that it has observed error greater than $\epsilon/2$ on the second half of the sample. A given example is just as likely to occur in the first half as in the second half. A group action based on this idea enables us to find the required bound by solving a simple counting problem.

We shall assume some measure-theoretic properties of the hypothesis spaces without explicit comment. These were mentioned in the Further Remarks of Chapter 3, and the details are discussed fully by Pollard (1984) and Blumer *et al.* (1989).

Theorem 8.3.1 If a hypothesis space has finite VC dimension, then it is potentially learnable.

Proof We use the notation introduced at the end of the previous section. There are four stages.

- Bound $\mu^m(Q_m^\epsilon)$ by the probability (with respect to μ^{2m}) of a certain subset R_m^ϵ of X^{2m}.
- Using a group action, bound the probability of R_m^ϵ in finite terms.
- Express this bound in terms of Π_H by combinatorial arguments.
- Apply the argument given in the last paragraph of Section 8.2 to conclude that $\mu^m(Q_m^\epsilon)$ tends to zero as m tends to infinity.

Stage 1 Given samples $\mathbf{x}, \mathbf{y} \in X^m$, let $\mathbf{xy} \in X^{2m}$ denote the sample of length $2m$ obtained by concatenating $\mathbf{x}$ and $\mathbf{y}$. With this notation, define

$$R_m^\epsilon = \left\{ \mathbf{xy} \in X^{2m} \;\middle|\; \exists h \in B_\epsilon \text{ for which } \mathrm{er}_{\mathbf{x}}(h) = 0 \text{ and } \mathrm{er}_{\mathbf{y}}(h) > \frac{\epsilon}{2} \right\}.$$

Lemma 8.3.2 For all $m \geq 8/\epsilon$,

$$\mu^m(Q_m^\epsilon) \leq 2\mu^{2m}(R_m^\epsilon).$$

Proof Let χ_Q be the characteristic function of Q_m^ϵ; that is, $\chi_Q(\mathbf{x}) = 1$ if $\mathbf{x} \in Q_m^\epsilon$ and $\chi_Q(\mathbf{x}) = 0$ otherwise. If we define the characteristic function χ_R similarly, then

$$\chi_R(\mathbf{xy}) = \chi_Q(\mathbf{x})\psi_{\mathbf{x}}(\mathbf{y}),$$

where

$$\psi_{\mathbf{x}}(\mathbf{y}) = \begin{cases} 1, & \text{if } \exists h \in H[\mathbf{x}] \cap B_\epsilon \text{ with } \mathrm{er}_{\mathbf{y}}(h) > \epsilon/2; \\ 0, & \text{otherwise.} \end{cases}$$

Now we have

$$\mu^{2m}(R_m^\epsilon) = \int \chi_R(\mathbf{xy}) = \int \left(\chi_Q(\mathbf{x}) \int \psi_{\mathbf{x}}(\mathbf{y}) \right),$$

where the integrals are taken over the whole of the relevant spaces, with respect to the product measures. The inner integral is the probability that, given $\mathbf{x}$, there is some $h \in B_\epsilon$ which is consistent with $\mathbf{x}$ and satisfies $\mathrm{er}_{\mathbf{y}}(h) > \epsilon/2$. This is certainly not less than the probability that a particular $h \in B_\epsilon$ which is consistent with $\mathbf{x}$ satisfies $\mathrm{er}_{\mathbf{y}}(h) > \epsilon/2$.

Thus it suffices to show that the above-mentioned quantity is at least $\frac{1}{2}$, for then we have

$$\mu^{2m}(R_m^\epsilon) \geq \int \frac{1}{2}\chi_Q(\mathbf{x}) = \frac{1}{2}\mu^m(Q_m^\epsilon).$$

In order to prove this, we use the following bound on the 'tail' of the binomial distribution. Let $0 \leq p \leq 1$ and let $LE(p, m, s)$ denote the probability of at most s successes in m independent trials each of which has a probability p of success. Then

$$LE\,(p, m, (1-\beta)mp) \leq e^{-\beta^2 mp/2},$$

for any $0 \leq \beta \leq 1$. This is often known as a Chernoff bound, since it follows from a special case of a result of Chernoff (1952). (See also Angluin and Valiant (1979) and, for a generalisation of this result, McDiarmid (1989).)

Let $h \in B_\epsilon$, so that $\mathrm{er}_\mu(h) = \epsilon_h > \epsilon$. For $\mathbf{y} \in X^m$, $m\,\mathrm{er}_{\mathbf{y}}(h)$ is the number of components of $\mathbf{y}$ on which h and t disagree, and so it is a binomially distributed random variable. Now, applying the above Chernoff bound, we have

$$\begin{aligned}
\mu^m\left\{\mathbf{y} \;\middle|\; \mathrm{er}_{\mathbf{y}}(h) \leq \frac{\epsilon}{2}\right\} &= \mu^m\left\{\mathbf{y} \;\middle|\; m\,\mathrm{er}_{\mathbf{y}}(h) \leq \frac{\epsilon}{2}m\right\} \\
&\leq \mu^m\left\{\mathbf{y} \;\middle|\; m\,\mathrm{er}_{\mathbf{y}}(h) \leq \frac{\epsilon_h}{2}m\right\} \\
&= LE\left(\epsilon_h, m, \left(1-\frac{1}{2}\right)m\epsilon_h\right) \\
&\leq \exp\left(-\frac{\epsilon_h m}{8}\right) \\
&< \exp\left(-\frac{\epsilon m}{8}\right).
\end{aligned}$$

For $m \geq 8/\epsilon$, this is at most $1/e$. It follows that for any $h \in B_\epsilon$,

$$\mu^m\left\{\mathbf{y} \;\middle|\; \mathrm{er}_{\mathbf{y}}(h) > \frac{\epsilon}{2}\right\} > 1 - \frac{1}{e} > \frac{1}{2}.$$

This completes the proof of Stage 1. □

Stage 2 The next stage is to bound the probability of R_m^ϵ by using a group action on X^{2m}. Following Pollard (1984), we use the 'swapping group' to convert the problem into an easy counting problem.

For $i \in \{1, \ldots, m\}$ let τ_i be the permutation of $\{1, \ldots, 2m\}$ which switches i and $m+i$. There is an induced transformation of X^{2m} defined by letting τ_i act on the coordinates, and we use τ_i to denote this transformation also. Thus, for example, if $m = 4$,

$$\tau_2(z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8) = (z_1, z_6, z_3, z_4, z_5, z_2, z_7, z_8).$$

Let G_m be the group generated by the permutations τ_i $(1 \leq i \leq m)$. As an abstract group G_m is just the direct product of m copies of the group of order 2, so $|G_m| = 2^m$.

Lemma 8.3.3 Given $\mathbf{z} \in X^{2m}$, let $\Gamma(\mathbf{z})$ denote the number of $\sigma \in G_m$ for which $\sigma\mathbf{z}$ is in R_m^ϵ. Then

$$|G_m|\, \mu^{2m}(R_m^\epsilon) \leq \max \Gamma(\mathbf{z}),$$

where this maximum is taken over all $\mathbf{z} \in X^{2m}$.

Proof The proof is quite general, applying to any finite group G of transformations of a space X^n induced by coordinate-permutations, and any subset S of X^n. Let χ_S be the characteristic function of S. Since G is finite we can interchange summation and integration as follows (where the integral sign represents integration over the entire space with respect to the product measure derived from μ):

$$\sum_{\sigma \in G} \int \chi_S(\sigma\mathbf{z}) = \int \sum_{\sigma \in G} \chi_S(\sigma\mathbf{z}).$$

The left-hand side is the sum over σ of the measure of $\sigma^{-1}(S)$, which is the same as the measure of S, since coordinate-permutations preserve the product measure. Hence the left-hand side is just $|G|\, \mu^n(S)$. The integrand on the right-hand side is just the number of σ in G for which $\sigma\mathbf{z} \in S$. Since the total weight of a probability measure is 1, the integral is bounded by the maximum of this quantity, taken over $\mathbf{z}$. Putting $n = 2m$, $G = G_m$, and $S = R_m^\epsilon$, the result follows. □

Stage 3 Given any $h \in B_\epsilon$, let

$$R_m^\epsilon(h) = \left\{ \mathbf{xy} \in X^{2m} \;\middle|\; \mathrm{er}_{\mathbf{x}}(h) = 0 \text{ and } \mathrm{er}_{\mathbf{y}}(h) > \frac{\epsilon}{2} \right\}.$$

Also, for $\mathbf{z} \in X^{2m}$, let $\Gamma(h, \mathbf{z})$ denote the number of $\sigma \in G_m$ which transform $\mathbf{z}$ to a vector in $R_m^\epsilon(h)$.

Lemma 8.3.4 Suppose that m is any positive integer and that $h \in B_\epsilon$. Then

$$\Gamma(h, \mathbf{z}) < 2^{m(1-\epsilon/2)},$$

for all $\mathbf{z} \in X^{2m}$.

Proof Suppose that $\Gamma(h, \mathbf{z}) \neq 0$. If $\mathbf{z} \notin R_m^\epsilon(h)$, then for some $\tau \in G_m$, $\tau\mathbf{z} \in R_m^\epsilon(h)$. But the number of σ such that $\sigma\mathbf{z} \in R_m^\epsilon(h)$ is precisely the number of σ for which $\sigma\tau\mathbf{z} \in R_m^\epsilon(h)$ (since G_m is a group). Hence, we may, without loss of generality, suppose that $\mathbf{z} \in R_m^\epsilon(h)$.

Now, $\mathbf{z} = \mathbf{xy}$ where $\mathrm{er}_{\mathbf{x}}(h) = 0$ and $\mathrm{er}_{\mathbf{y}}(h) > \epsilon/2$. To simplify notation, let us suppose that the $r > m\epsilon/2$ entries of $\mathbf{z}$ on which h and the target concept t disagree are

$$z_{m+1}, z_{m+2}, \ldots, z_{m+r}.$$

Recall that a transformation $\sigma \in G_m$ interchanges some pairs (z_j, z_{m+j}). If $\sigma\mathbf{z}$ is in $R_m^\epsilon(h)$ then σ does not interchange (z_j, z_{m+j}) for $1 \leq j \leq r$. Conversely, any σ which satisfies this condition is in $R_m^\epsilon(h)$. Since σ is uniquely determined by the set of j for which $\sigma(z_j) = z_{m+j}$, the number of such σ is just the number of subsets of $\{r+1, r+2, \ldots, m\}$; that is, $\Gamma(h, \mathbf{z}) = 2^{m-r}$. Since $r > \epsilon m/2$, we have

$$\Gamma(h, \mathbf{z}) < 2^{m-\epsilon m/2},$$

as required. □

Lemma 8.3.5 For any positive integer m,

$$\mu^{2m}(R_m^\epsilon) < \Pi_H(2m)\, 2^{-\epsilon m/2}.$$

Proof Let $\mathbf{z} \in X^{2m}$ be fixed but arbitrary, and let $s = \Pi_{B_\epsilon}(\mathbf{z})$. Then there are hypotheses $h_1, \ldots, h_s$ in B_ϵ which give s different classifications of $\mathbf{z}$ and, further, any classification of $\mathbf{z}$ by a hypothesis in B_ϵ is one of these s classifications. We have

$$s = \Pi_{B_\epsilon}(\mathbf{z}) \leq \Pi_H(\mathbf{z}) \leq \Pi_H(2m).$$

Suppose $\sigma\mathbf{z} = \mathbf{ab}$ is in R_m^ϵ. This means that there is some $h \in H$ such that $\mathrm{er}_{\mathbf{a}}(h) = 0$ and $\mathrm{er}_{\mathbf{b}}(h) > \epsilon/2$. Since all classifications of $\mathbf{z}$, and hence of its rearrangement $\sigma\mathbf{z} = \mathbf{ab}$, are realised by some h_i $(1 \leq i \leq s)$, it follows that $\sigma\mathbf{z}$ is in one of the sets $R_m^\epsilon(h_i)$. Thus the set of σ for which $\sigma\mathbf{z}$ is in R_m^ϵ is the union of the sets of those σ for which $\sigma\mathbf{z}$ is in $R_m^\epsilon(h_i)$. In terms of the notation previously introduced, we therefore have

$$\Gamma(\mathbf{z}) \leq \sum_{i=1}^{s} \Gamma(h_i, \mathbf{z}).$$

The last expression is the sum of $s \leq \Pi_H(2m)$ terms and, by Lemma 8.3.4, each of them is bounded above by $2^{m(1-\epsilon/2)}$. Thus, from Lemma 8.3.3, we have

$$\mu^{2m}(R_m^\epsilon) \leq |G_m|^{-1} \max_{\mathbf{z}} \Gamma(\mathbf{z}) \leq 2^{-m} \Pi_H(2m)\, 2^{m(1-\epsilon/2)} = \Pi_H(2m)\, 2^{-m\epsilon/2},$$

as claimed. □

Stage 4 The bound

$$\mu^m(Q_m^\epsilon) < 2\, \Pi_H(2m)\, 2^{-m\epsilon/2}$$

follows by combining Lemmas 8.3.2 and 8.3.5. If H has finite VC dimension then, by Sauer's Lemma, $\Pi_H(2m)$ is bounded by a polynomial function of m. The right-hand side is eventually dominated by the negative exponential term, and tends to 0 as m tends to infinity; so it can be made less than any given $\delta > 0$ by choosing $m \geq m_0(\delta, \epsilon)$, a quantity depending only on δ and ϵ. Thus H is potentially learnable. □