Chapter 7: The VC Dimension

7.1 MOTIVATION

Suppose that, as in the framework of previous chapters, we have a hypothesis space H defined on an example space X. In Chapter 4 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. A pertinent comment is that most hypothesis spaces which occur 'naturally' have a high degree of structure, and even if the space is infinite it may contain functions only of a special type. This is true, almost by definition, for any hypothesis space H which is constructed by means of a representation $\Omega \to H$.

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'. In this chapter 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). The development of this notion is probably the most significant contribution that mathematics has made to Computational Learning Theory.

In order to illustrate some of the ideas, we consider the *real perceptron*. This is a machine which operates in the same manner as the linear threshold machine of Section 2.5, but with real-valued inputs. Thus, as shown in Figure 7.1, there are n inputs and a single active node. The arcs carrying the inputs have real-valued weights $\alpha_1, \alpha_2, \ldots, \alpha_n$ and there is a real threshold value θ at the active node. As with the linear threshold machine, 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.

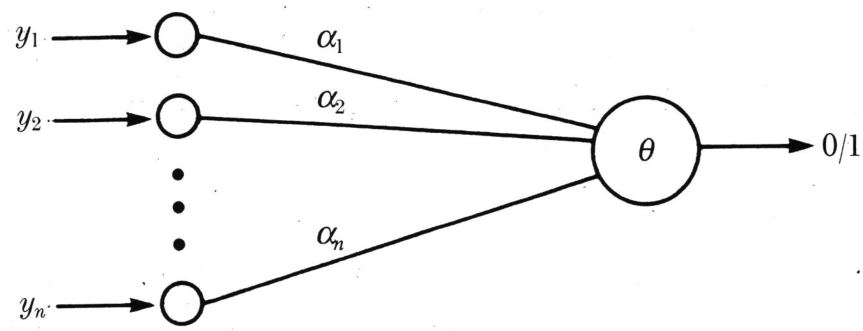

Figure 7.1: The real perceptron P_n

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

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

It should be noted that $\omega \mapsto h_\omega$ is not an injection: for any $\lambda > 0$ the state $\lambda\omega$ defines the same function as ω.

Example 7.1.1 As an example, consider P_2, the real perceptron with two inputs. In state $\omega = (\alpha_1, \alpha_2, \theta)$, P_2 computes the boolean-valued function h_ω for which

$$h_\omega(y_1, y_2) = 1 \iff \alpha_1 y_1 + \alpha_2 y_2 \geq \theta.$$

It is useful to describe this geometrically (Figure 7.2). The example $y = (y_1, y_2)$, considered as a point in the plane $\mathbf{R}^2$, is a positive example of h_ω if and only if y lies on the straight line l_ω with equation $\alpha_1 y_1 + \alpha_2 y_2 = \theta$ or on the side of l_ω consisting of points with $\alpha_1 y_1 + \alpha_2 y_2 > \theta$.

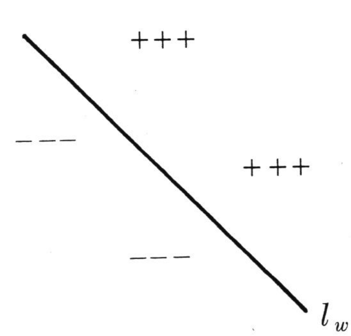

Figure 7.2: Geometrical interpretation of a hypothesis in P_2

Given a sample of m points in $\mathbf{R}^2$, the machine P_2 can only achieve certain classifications of the sample into positive and negative examples: precisely those for which, as above, the positive examples are separated from the negative examples by a line in the plane. When a classification of the sample can be realised in this way, we shall say that it is *linearly separable.* The fact that relatively few classifications are linearly separable is an indication of the restricted 'expressive power' of P_2. □

7.2 THE GROWTH FUNCTION

Suppose that H is a hypothesis space defined on the example space X, and let $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ 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

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

as h runs through all hypotheses of H. Although H may be infinite, we observe that $H|E_{\mathbf{x}}$, the hypothesis space obtained by restricting the hypotheses of H to domain $E_{\mathbf{x}} = \{x_1, x_2, \ldots, x_m\}$, is finite and is of cardinality $\Pi_H(\mathbf{x})$. Note that for any sample $\mathbf{x}$ of length m, $\Pi_H(\mathbf{x}) \leq 2^m$. An important quantity, and one which shall turn out to be crucial in applications to potential learnability, is the maximum possible number of classifications by H of a sample of a given length. We define the *growth function* Π_H by

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

We have used the notation Π_H for both the number of classifications and the growth function, but this should cause no confusion.

Example 7.2.1 Let $X = \mathbf{R}$ be the real line and let H be the set of rays, as defined in Chapter 2. Suppose that m is a positive integer and that $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is a sample of length m, in which the examples are arranged in strictly increasing order:

$$x_1 < x_2 < \ldots < x_m.$$

Given $\theta \in \mathbf{R}$, $r_\theta(x_i) = 1$ if and only if $x_i \geq \theta$. Therefore, for any $h = r_\theta$ and any k between 1 and $m-1$, $h(x_k) = 1$ implies $h(x_{k+1}) = 1$. Thus the set of 'classification vectors' (vectors of the form $(h(x_1), h(x_2), \ldots, h(x_m))$ for some $h \in H$) consists only of the $m+1$ vectors

$$(111\ldots11),\ (011\ldots11),\ (001\ldots11),\ \ldots,\ (000\ldots00).$$

Now any sample in which the examples are distinct can be obtained from one in which the examples are in strictly increasing order by a permutation, and this permutation of the sample will simply give another set of $m+1$ classification vectors. If not all the examples are distinct, there will clearly be fewer possible classifications. Thus $\Pi_H(m)$, the maximum number of classifications, is $m+1$. □

In general, it is difficult to find an exact formula for the growth function of a hypothesis space. In the next section 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.

7.3 THE VC DIMENSION

We noted above that the number of possible classifications by H of a sample of length m is at most 2^m, this being the number of binary vectors of length m. We say that a sample $\mathbf{x}$ of length m is *shattered* by H, or that H *shatters* $\mathbf{x}$, if this maximum possible value is attained; that is, if H gives all possible classifications of $\mathbf{x}$. Note that if the examples in $\mathbf{x}$ are not distinct then $\mathbf{x}$ cannot be shattered by any H. When the examples are distinct, $\mathbf{x}$ is shattered by H if and only if for any subset S of $E_{\mathbf{x}}$, there is some hypothesis h in H such that for $1 \le i \le m$,

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

S is then the subset of $E_{\mathbf{x}}$ comprising the positive examples of h.

Based on the intuitive notion that a hypothesis space H has high expressive power if it can achieve all possible classifications of a large set of examples, we use as a measure of this power the *Vapnik-Chervonenkis dimension*, or *VC dimension*, of H, defined as follows. The 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. Using the notation introduced in the previous section, we can say that the VC dimension of H, denoted VCdim(H), is given by

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

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

Example 7.3.1 Consider again the case in which X is the real line and H is the space of rays. Given a sample (y, y') of length 2, we may suppose without loss that $y < y'$. Then there is no ray $h = r_\theta$ such that $h(y) = 1$ and $h(y') = 0$, because if such a ray were to exist, we should have $y' < \theta \le y$. Therefore H shatters no sample of length 2. Clearly H shatters any sample consisting of just one example, and therefore VCdim(H) = 1. □

Example 7.3.2 Let X be the plane $\mathbf{R}^2$, and H the hypothesis space of P_2. Suppose that $\mathbf{x} = (x_1, x_2, x_3)$ is any sample consisting of three distinct non-collinear points.

We observed earlier that H can achieve precisely those classifications of a sample into positive and negative examples which are linearly separable. Thus, $\mathbf{x}$ is shattered by H if and only if for any subset S of $E_{\mathbf{x}} = \{x_1, x_2, x_3\}$, S and $E_{\mathbf{x}} \setminus S$ are linearly separable. This is easily seen to be true in this case (Figure 7.3), and hence $\text{VCdim}(H) \geq 3$.

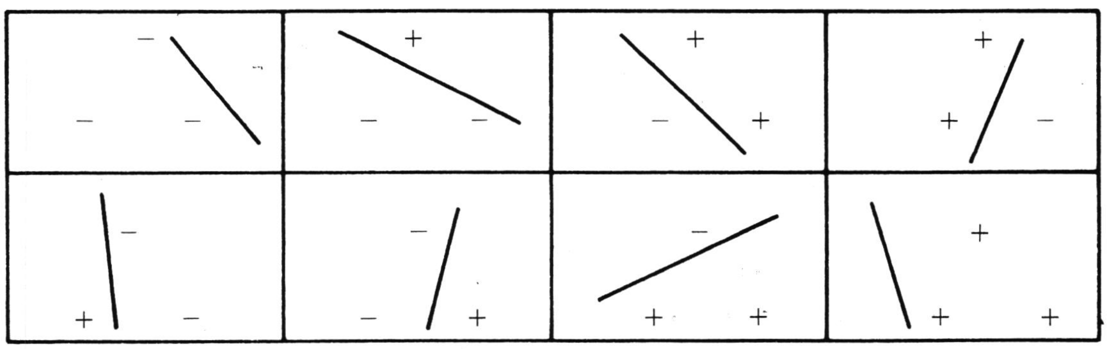

Figure 7.3: P_2 shatters three non-collinear points

In order to prove that $\text{VCdim}(H) = 3$, we have to show that *no* sample of length 4 is shattered by H. Suppose, by way of contradiction, that the sample $\mathbf{x} = (x_1, x_2, x_3, x_4)$ of length 4 is shattered by H. Then for every $S \subseteq E_{\mathbf{x}}$, S and $E_{\mathbf{x}} \setminus S$ are linearly separable and so, in particular, no three of x_1, x_2, x_3, x_4 can be collinear. There are two cases to consider: either all four points are boundary points of the smallest closed polygonal region containing $E_{\mathbf{x}}$, or one of the points (without loss, x_4) lies in the interior of this region. Typical examples of these cases are illustrated in Figure 7.4.

In the first case, $\{x_1, x_3\}$ and $\{x_2, x_4\}$ (for example) are not linearly separable, while in the second case $\{x_4\}$ and $\{x_1, x_2, x_3\}$ are not linearly separable. Therefore H shatters no sample of length 4 and, consequently, as claimed, $\text{VCdim}(H) = 3$. □

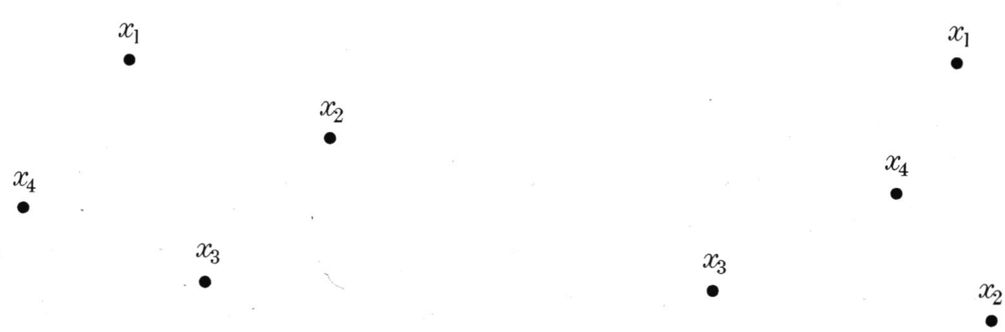

Figure 7.4: The two cases for a sample of four points

When the hypothesis space H is the set of functions defined by some representatior $\Omega \to H$, we shall take the VC dimension of the representation to be the VC dimensior of H. Thus, we have shown that the VC dimension of P_2 is 3.

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

Proposition 7.3.3 If H is a finite hypothesis space, then

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

Proof The VC dimension of H is the greatest integer d for which $\Pi_H(d) = 2^d$. Bu the number of classifications by a finite hypothesis space H of a sample of any lengtl is certainly at most the number of distinct hypotheses in H. Hence, for any positiv integer m, $\Pi_H(m) \leq |H|$. In particular,

$$2^d = \Pi_H(d) \leq |H| \, .$$

Taking logarithms gives the result. [

Example 7.3.4 Using the foregoing Proposition, we can obtain an upper boun on the VC dimension of M_n, the hypothesis space of monomial concepts defined o $\{0,1\}^n$. Recall that $|M_n| = 3^n$ and therefore, by the Proposition, the VC dimensio of M_n is at most $\lg 3^n$. That is

$$\mathrm{VCdim}(M_n) \leq (\lg 3)\, n.$$

In order to get a lower bound, we claim that M_n shatters the sample $(e_1, e_2, \ldots, e_n$ where, for i between 1 and n, e_i is the point in $\{0,1\}^n$ with 1 as entry in position i an with all other entries 0. It will follow immediately from this that the VC dimensio of M_n is at least n. To prove our claim, suppose that

$$q = (q_1, q_2, \ldots, q_n) \in \{0,1\}^n \, .$$

We have to show that there is h in M_n such that

$$h(e_1) = q_1, \; h(e_2) = q_2, \ldots, h(e_n) = q_n.$$

If q is the all-1 vector, we take h to be the empty monomial in which no liter appears; otherwise we take h to be the conjunction of those literals $\overline{u}_j$ for whic $q_j = 0$. Summarising, we have

$$n \leq \mathrm{VCdim}(M_n) \leq (\lg 3)\, n$$

for any n.

7.4 THE VC DIMENSION OF THE REAL PERCEPTRON

We have seen that the VC dimension of P_2 is 3. Furthermore, if one interprets P_1 in the obvious way (Exercise 2), then it is easy to verify that P_1 has VC dimension 2. We shall prove in this section that, more generally, for any positive integer n, the VC dimension of P_n is precisely $n+1$. In order to do so, we need some geometrical ideas.

Consider the perceptron P_n with n inputs. In state

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

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

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

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

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

bounded by the *hyperplane*

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

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

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

Roughly speaking, l_ω divides $\mathbf{R}^n$ into the set of positive examples of h_ω and the set of negative examples of h_ω

A subset C of $\mathbf{R}^n$ 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 notation here is the standard one for the real vector space $\mathbf{R}^n$.) It is clear that 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 conv(S), is called the *convex hull* of S; 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 conv(S) is the smallest closed region which is bounded by a polygon and which contains S.

We shall find the following result, known as *Radon's Theorem*, extremely useful. Le n be any positive integer, and let E be any set of $n+2$ points in $\mathbf{R}^n$. Then there i a non-empty subset S of E such that

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

A proof is given by Grunbaum (1967).

Theorem 7.4.1 For any positive integer n, let P_n be the real perceptron with inputs. Then

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

Proof Let $\mathbf{x} = (x_1, x_2, \ldots, x_{n+2})$ be any sample of length $n+2$. As we have notec if two of the examples are equal then $\mathbf{x}$ cannot be shattered. Suppose then that th 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 a hypothesis h_ω in P_n such that S is the set of positive example 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

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

We deduce that no such h_ω exists and therefore that $\mathbf{x}$ is not shattered by P_n. Thu *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, fc $1 \leq i \leq n$, let e_i be the point with a 1 in the ith coordinate and all other coordinate 0. We shall show that P_n shatters the sample

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

of length $n+1$.

Suppose that S is a subset of $E_\mathbf{x} = \{o, e_1, \ldots, e_n\}$. For $i = 1, 2, \ldots, n$, 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

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

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, $\text{VCdim}(P_n) \geq n + 1$. Combining these two results, we have the stated equality. □

7.5 SAUER'S LEMMA

In this section we assume that H has finite VC dimension. The growth function $\Pi_H(m)$ is a measure of how many different classifications of an m-sample into positive and negative examples can be achieved by the hypotheses of H, while the VC dimension of H is the maximum value of m for which $\Pi_H(m) = 2^m$. Clearly these two quantities are related, because the VC dimension is defined in terms of the growth function. But there is another, less obvious, relationship: the growth function $\Pi_H(m)$ can be bounded by a polynomial function of m, and the degree of the polynomial is the VC dimension d of H. Explicitly, we have the following theorem, due to Sauer (1972) and Shelah (1972) independently (see Assouad (1983)). In combinatorial circles it is usually known as *Sauer's Lemma*.

Theorem 7.5.1 (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

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

where the binomial numbers are defined by

$$\binom{m}{i} = \frac{m(m-1)\ldots(m-i+1)}{1.2\ldots m}.$$

□

Before we give the proof, it may be helpful to interpret the result. First, it should be noted that the explicit definition of the binomial numbers means that $\binom{a}{b}$ is zero whenever $b > a \geq 1$. Thus for values of m not exceeding d the result asserts only that

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

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

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

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.

For convenience, we let $\Phi(d,m)$ denote this sum of binomial numbers for *any* $d \geq 0$ and $m \geq 1$. We have:

$$\Phi(0,m) = 1 \ \ (m \geq 1); \quad \Phi(d,1) = 2 \ \ (d \geq 1).$$

The binomial numbers satisfy the identity

$$\binom{a}{b} = \binom{a-1}{b} + \binom{a-1}{b-1},$$

which can be verified explicitly using the formula. From this we can immediately derive the identity

$$\Phi(d,m) = \Phi(d,m-1) + \Phi(d-1,m-1),$$

which is valid for all $d \geq 1$ and $m \geq 2$ (Exercise 5).

Proof of Sauer's Lemma If H is a hypothesis space with $d = \text{VCdim}(H) = 0$ then for any example x, $h(x)$ is the same (either 0 or 1) for all hypotheses $h \in H$. It follows that $\Pi_H(\mathbf{x}) = 1$ for any sample $\mathbf{x}$ of any length m. Thus $\Pi_H(m) = 1 = \Phi(0,m)$, and the theorem is true in the case $d = 0$.

If $m = 1$ and $d \geq 1$, then for any H we have $\Pi_H(1) \leq 2 = \Phi(d,1)$, so the theorem is true in this case also.

Using these 'boundary conditions' we can prove the theorem by induction on $d+m$. The case $d+m=2$ is covered explicitly by the boundary conditions. Suppose the result holds for all cases with $d+m \leq k$, where $k \geq 2$, and let H be a hypothesis space of VC dimension d and $\mathbf{x}$ a sample of length m, where $d+m = k+1$. The cases $(d,m) = (0,k+1)$ and $(d,m) = (k,1)$ are covered by the boundary conditions so we may assume that $d \geq 1, m \geq 2$.

If the given sample $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ contains repeated examples, then we can remove the repetitions and obtain a shorter sample. The result then follows by the induction hypothesis. So we may suppose that $\mathbf{x}$ contains m distinct examples. Let E be the set of examples in $\mathbf{x}$ and let $H_E = H|E$ be the hypothesis space on E obtained by restricting the hypotheses of H to the domain E. Then, as remarked earlier, H_E is finite and $\Pi_H(\mathbf{x}) = |H_E|$. We shall show that $|H_E| \leq \Phi(d, m)$.

Let $F = E \setminus \{x_m\}$ and consider the hypothesis space $H_F = H|F$. Two distinct hypotheses h, g of H_E give, on restriction to F, the same hypothesis of H_F precisely when h and g agree on F and disagree on x_m. Denote by H_* the set of hypotheses of H_F which arise in this manner from two distinct hypotheses of H_E. Thus, if $h_* \in H_*$ then both possible extensions of h_* to a $\{0,1\}$-function on E are hypotheses of H_E. It follows that

$$|H_E| = |H_F| + |H_*|.$$

We now bound $|H_F|$ and $|H_*|$.

Let $\mathbf{x}' = (x_1, x_2, \ldots, x_{m-1})$ be the sample consisting of the first $m-1$ examples of $\mathbf{x}$. Then H_F is a hypothesis space on F and therefore

$$|H_F| = \Pi_H(\mathbf{x}') \leq \Pi_H(m-1).$$

Using the induction hypothesis we can conclude that

$$|H_F| \leq \Pi_H(m-1) \leq \Phi(d, m-1),$$

since $d + (m-1) \leq k$.

We claim that VCdim(H_*) is at most $d-1$. Indeed, suppose that H_* shatters some sample $\mathbf{z} = (z_1, z_2, \ldots, z_d)$ of length d of examples from F. For each $h_* \in H_*$, there are $h_1, h_2 \in H_E$ such that h_1 and h_2 agree with h_* on F, and $h_1(x_m) = 0, h_2(x_m) = 1$. It follows that H_E, and hence H, shatters the sample $(z_1, \ldots, z_d, x_m)$ of length $d+1$, an impossibility since VCdim$(H) \leq d$. Hence VCdim$(H_*) \leq d-1$. Using the induction hypothesis again we have

$$|H_*| = \Pi_{H_*}(\mathbf{x}') \leq \Pi_{H_*}(m-1) \leq \Phi(d-1, m-1),$$

since $(d-1) + (m-1) \leq k$.

Combining the results obtained, we have

$$\Pi_H(\mathbf{x}) = |H_E| = |H_F| + |H_*| \leq \Phi(d, m-1) + \Phi(d-1, m-1) = \Phi(d, m),$$

as required. □

Example 7.5.2 Let H be the hypothesis space of the real perceptron P_n. Then H has VC dimension $n+1$ and therefore, for any positive integer m, $\Pi_H(m) \leq \Phi(n+1, m)$. For example, when $n = 2$

$$\Pi_H(4) \leq \Phi(3,4) = 1 + 4 + 6 + 4 = 15.$$

This corresponds to the fact, illustrated in Figure 7.4, that not all the 2^4 classifications of a 4-sample can be realised by P_2. In fact, careful analysis of the cases shows that $\Pi_H(4) = 14$ (Exercise 3). □

We shall now elaborate on the fact that $\Phi(d, m)$ is bounded by a polynomial function of m, of degree d. A simple form of this result, $\Phi(d, m) \leq m^d$ for $m \geq d > 1$, is fairly easy to prove (Exercise 6). But there is some advantage in having a better bound, as given by the following result of Blumer *et al.* (1989).

Proposition 7.5.3 For all $m \geq d \geq 1$,

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

where e is the base of natural logarithms.

Proof The proof is in two stages. First, we claim that for all positive integers d,

$$\Phi(d, m) \leq \frac{2m^d}{d!}$$

for all $m \geq d$. This can be proved by an inductive argument, as follows. If $d = 1$ then $\Phi(d, m) = m + 1 \leq 2m$. If $m = d > 1$ then $\Phi(d, m) = \Phi(d, d) = 2^d$. Now, for $d \geq 1$, we have

$$\left(1 + \frac{1}{d}\right)^d \geq 1 + d\frac{1}{d} = 2.$$

This justifies the induction step in the following argument:

$$2^{d+1} \leq \left(\frac{d+1}{d}\right)^d 2^d \leq 2\left(\frac{d+1}{d}\right)^d \frac{d^d}{d!} = 2\frac{(d+1)^{d+1}}{(d+1)!},$$

and verifies the claim for $m = d > 1$.

Suppose that $m > d \geq 1$. Since

$$\Phi(d+1, m+1) = \Phi(d+1, m) + \Phi(d, m),$$

it suffices to prove that

$$2\frac{m^d}{d!} + 2\frac{m^{d+1}}{(d+1)!} \leq 2\frac{(m+1)^{d+1}}{(d+1)!}.$$

It is straightforward to verify that this is true if and only if

$$1+\left(\frac{d+1}{m}\right) \leq \left(1+\frac{1}{m}\right)^{d+1},$$

which follows from the binomial theorem. Thus, for all $m \geq d$, $\Phi(d,m) \leq 2m^d/d!$.

It remains to show that, for all $m \geq d \geq 1$,

$$2\left(\frac{d}{e}\right)^d < d!.$$

The result clearly holds when $d = 1$. Suppose it holds for a given value of $d \geq 1$: then

$$(d+1)! = (d+1)\,d! > (d+1)\,2\left(\frac{d}{e}\right)^d.$$

Thus it suffices to prove that

$$(d+1)\,2\left(\frac{d}{e}\right)^d > 2\left(\frac{d+1}{e}\right)^{d+1};$$

that is,

$$\left(1+\frac{1}{d}\right)^d \leq e,$$

which is indeed true for any $d \geq 1$. The result follows. □

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

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

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

The following consequence of the results in this section will be of use to us later.

Proposition 7.5.4 Let H be any hypothesis space consisting of at least two hypotheses and defined on a finite example space X. Then

$$\mathrm{VCdim}(H) > \frac{\ln|H|}{1+\ln|X|}.$$

Proof Observe that two hypotheses of H are distinct precisely when they give different classifications of the whole example space X into positive and negative examples. Since there are $\Pi_H(|X|)$ such classifications, we have $|H| = \Pi_H(|X|)$. It follows from Sauer's Lemma and Proposition 7.5.3 that

$$|H| = \Pi_H(|X|) < \left(\frac{e|X|}{d}\right)^d,$$

where $d \geq 1$ is the VC dimension of H. Now,

$$|H| < \left(\frac{e|X|}{d}\right)^d \Longrightarrow d(1 + \ln|X|) - d\ln d > \ln|H|$$

$$\Longrightarrow d > \frac{\ln|H|}{1 + \ln|X|},$$

as required. □

We remark that if $\text{VCdim}(H) \geq 2$, then this result can be improved to

$$\text{VCdim}(H) \geq \frac{\ln|H|}{\ln|X|},$$

using the result $\Phi(d, m) \leq m^d$ for $m \geq d > 1$.

FURTHER REMARKS

For any positive integer n, let G_n be the subset of the hypothesis space of P_n consisting of the hypotheses for which the zero vector (the origin) is a negative example. Thus, G_n is the set of characteristic functions of all those closed half-spaces of $\mathbf{R}^n$ which do not contain the origin. Then one can show that $G = G_n$ has VC dimension n (Exercise 10) and that for any m, $\Pi_G(m) = \Phi(n, m)$ (see Vapnik and Chervonenkis (1971)). Thus the major result of this chapter, $\Pi_H(m) \leq \Phi(d, m)$ is the best possible result of its kind.

EXERCISES

1. Show that if $X = \mathbf{R}$ and H is the set of all closed intervals, then

$$\Pi_H(m) = 1 + m + \frac{1}{2}m(m-1).$$

2. Describe explicitly the hypothesis space of P_1 and show that the VC dimension of P_1 is 2.

3. Show that when H is the hypothesis space of the real perceptron P_2, $\Pi_H(4) = 14$

4. Let H be a hypothesis space of finite VC dimension. For $h \in H$, define the $\{0,1\}$-valued function $\bar{h}$ by

$$\bar{h}(x) = 1 \iff h(x) = 0,$$

and let the *complement* of H be the space $\left\{\bar{h} \mid h \in H\right\}$. Prove that this space has the same VC dimension as H.

5. Prove that $\Phi(d,m) = \Phi(d,m-1) + \Phi(d-1,m-1)$ for $d \geq 1$ and $m \geq 2$.

6. Prove that $\Phi(d,m) \leq m^d$, for all $m \geq d > 1$.

7. A monomial is *monotone* if it contains no negated literals. Prove that the space of monotone monomials defined on $\{0,1\}^n$ has VC dimension precisely n.

8. A hypothesis space H is *linearly ordered* if it has at least two hypotheses and if for any $h, g \in H$, either

$$h(x) = 1 \Longrightarrow g(x) = 1$$

or

$$g(x) = 1 \Longrightarrow h(x) = 1.$$

Prove that if H is linearly ordered then $\text{VCdim}(H) = 1$. (This is a result of Wenocur and Dudley (1981).) Deduce that the space of rays has VC dimension 1.

9. Suppose that H contains the identically-0 function and the identically-1 function, and that $\text{VCdim}(H) = 1$. Prove that H is linearly ordered. (This is a result of Wenocur and Dudley (1981).)

10. Let G_n be the set of hypotheses of P_n for which the zero vector o is a negative example. Suppose that the sample $\mathbf{x} = (x_1, x_2, \ldots, x_m)$ is shattered by G_n. Why can none of the x_i be o? Prove that the sample $(x_1, \ldots, x_m, o)$ is shattered by P_n. Using this, prove that $\text{VCdim}(G_n) = n$.

11. Use the result on G_n stated in the Further Remarks to prove that for $m \geq 2$, $\Pi_{P_n}(m) = 2\Phi(n, m-1)$.
[Hint: Let $\mathbf{x}$ be a sample of length m for which $\Pi_{P_n}(\mathbf{x}) = \Pi_{P_n}(m)$. Without loss of generality, we may assume that the origin o is one of the examples in $\mathbf{x}$, since clearly the number of classifications by P_n of a vector is unchanged if the vector is translated. Thus, $\mathbf{x} = (x_1, \ldots, x_{m-1}, o)$. How are $\Pi_{P_n}(\mathbf{x})$ and $\Pi_{G_n}((x_1, \ldots, x_{m-1}))$ related?]