Classification by Polynomial Surfaces

Martin Anthony
Department of Statistical and Mathematical Sciences
London School of Economics and Political Science
Houghton Street, London WC2A 2AE
anthony@vax.lse.ac.uk

Abstract

Linear threshold functions (for real and Boolean inputs) have received much attention, for they are the component parts of many artificial neural networks. Linear threshold functions are exactly those functions such that the positive and negative examples are separated by a hyperplane. One extension of this notion is to allow separators to be surfaces whose equations are polynomials of at most a given degree (linear separation being the degree-1 case). We investigate the representational and expressive power of polynomial separators. Restricting to the Boolean domain, by using an upper bound on the number of functions defined on $\{0,1\}^n$ by polynomial separators having at most a given degree, we show, as conjectured by Wang and Williams [26], that for almost every Boolean function, one needs a polynomial surface of degree at least $\lfloor n/2 \rfloor$ in order to separate the negative examples from the positive examples. Further, we show that, for odd n, at most half of all Boolean functions are realisable by a separating surface of degree $\lfloor n/2 \rfloor$. We then compute the Vapnik-Chervonenkis dimension of the class of functions realised by polynomial separating surfaces of at most a given degree, both for the case of Boolean inputs and real inputs. In the case of linear separators, the VC dimensions coincide for these two cases, but for surfaces of higher degree, there is a strict divergence. We then use these results on the VC dimension to quantify the sample size required for valid generalisation in Valiant's probably approximately correct framework [24, 6].

1 Introduction

A $\{0,1\}$-valued function f on $\mathbf{R}^n$ is said to be a *linear threshold function* if there are real numbers $w_1, w_2, \ldots, w_n$ and θ such that $f(x) = 1$ if and only if the inner product $\langle w, x \rangle = \sum_{i=0}^{n} w_i x_i$ is greater than θ. The coefficients w_i $(1 \leq i \leq n)$ are known as the *weights*, the vector w is known as the *weight-vector*, and θ is called the *threshold*. When $\theta = 0$, we say that t is a *homogeneous* linear threshold function. Linear threshold functions have been studied extensively: see, for example, [19, 13, 17, 15, 3]. Observe that a function is a linear threshold function if and only if the positive examples can be separated from the negative examples by a separating hyperplane. (In the case of a homogeneous linear threshold function, this hyperplane contains the origin.) One may consider different kinds of *separators*, generalising the notion of a linear threshold function. As in [7, 20, 26] (for example), one may consider separators which are quadratic, cubic, or in general multilinear. A surface in $\mathbf{R}^n$ is said to be a *polynomial discriminator* of order m if it can be described by a multinomial equation of degree m in the variables $x_1, x_2, \ldots, x_n$. For example, the set of discriminators of order 2 in $\mathbf{R}^3$ consists of all surfaces whose equations are of the form

$$w_1x_1 + w_2x_2 + w_3x_3 + w_4x_1^2 + w_5x_2^2 + w_6x_3^2 + w_7x_1x_2 + w_8x_1x_3 + w_9x_2x_3 = w_{10},$$

for some contants $w_i, (1 \leq i \leq 10)$, where at least one of the terms of order 2 has a non-zero coefficient. (This last condition ensures that the order of a discriminator is well-defined.)

Linear and polynomial discriminators have been studied in the the context of pattern classification (see, for example, [9, 8, 20]), where the aim is to classify a given set of test data points into two categories correctly, this classification being used as a (hopefully valid) means of classifying further points. Although we shall discuss only the method of separation by polynomial surfaces in this paper, it should be mentioned that other methods have usefully been applied when the desired classification of the data points cannot be achieved by linear separators; see, for example [16, 27]. The representation and approximation of boolean functions by polynomials has useful applications in a number of areas of computer science; a discussion of some of the problems studied and the techniques used may be found in the survey of Saks [23]. In addition, polynomial discriminators have recently been employed in signal processing [22]. It is therefore an important problem to determine the 'power' of classification achievable by such discriminators.

Here, we investigate the representational and expressive power of polynomial separators. Restricting to the Boolean domain, and bounding the number of functions defined on $\{0,1\}^n$ by polynomial separators having at most a given degree, we show that for almost every Boolean function, one needs a polynomial surface of degree at least $\lfloor n/2 \rfloor$ in order to separate the negative examples from the positive examples. We then compute the Vapnik-Chervonenkis dimension of the class of functions realised by polynomial separating surfaces of at most a given degree; we do this both for the case of Boolean inputs and real inputs. In the case of linear separators, the VC dimensions coincide for these two cases, but for surfaces of higher degree, there is a strict divergence. Having computed the VC dimensions, we obtain an indication of how large a set of test data should be used for valid further classification of previously unseen points (within the criteria of the 'probably approximately correct' model of machine learning [24, 6, 1]).

2 Definitions and Notation

We now introduce some notation which will prove useful. Let us denote the set $\{1, 2, \ldots, n\}$ by $[n]$. We shall denote the set of all subsets of at most m objects from $[n]$ by $[n]^{(m)}$ and we shall denote by $[n]^m$ the set of all selections, in which repetition is allowed, of at most m objects from $[n]$. Thus, $[n]^m$ is a collection of 'multi-sets'. For example, $[3]^{(2)}$ consists of the sets

$$\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\},$$

while $[3]^2$ consists of the multisets

$$\emptyset, \{1\}, \{1,1\}, \{2\}, \{2,2\}, \{3\}, \{3,3\}, \{1,2\}, \{1,3\}, \{2,3\}.$$

In general, $[n]^{(m)}$ consists of $\sum_{i=0}^{m} \binom{n}{i}$ sets, $[n]^m$ consists of $\binom{n+m}{m}$ multisets, and, allowing a slight abuse of notation, $[n]^{(m)} \subseteq [n]^m$. For each $S \in [n]^m$, and for any $x = (x_1, x_2, \ldots, x_n) \in \mathbf{R}^n$, x_S denotes the product of the x_i for $i \in S$ (with repetitions as required). For example, $x_{\{1,2,3\}} = x_1 x_2 x_3$ and $x_{\{1,1,2\}} = x_1^2 x_2$. When $S = \emptyset$, the empty set, we interpret x_S as the constant 1.

With this notation, a $\{0,1\}$-valued function f defined on $\mathbf{R}^n$ is a *polynomial discriminant function* of degree at most m if there are constants w_S, one for each $S \in [n]^m$, such that

$$f(x) = 1 \Longleftrightarrow \sum_{S \in [n]^m} w_S x_S > 0.$$

The set of polynomial discriminant functions on $\mathbf{R}^n$ of degree at most m will be denoted by $P(n,m)$.

If we restrict attention to binary inputs (which we take to be 0 or 1) then any terms x_S in which S contains a repetition are redundant, simply because for $x = 0$ or 1, $x^k = x$ for all k; thus, for example, for binary inputs, $x_1 x_2^2 x_3^3 = x_1 x_2 x_3$. Therefore, we arrive at the following definition from [26], which is a special case of that of a polynomial discriminant function. A Boolean function f of n variables (that is, a $\{0,1\}$-valued function on $\{0,1\}^n$) is an *order-at-most-m threshold function* if there are constants w_S, one for each $S \in [n]^{(m)}$, such that

$$f(x) = 1 \Longleftrightarrow \sum_{S \in [n]^{(m)}} w_S x_S > 0.$$

If, in addition, f is not an order-at-most-$(m-1)$ threshold function, then we say that f has *threshold order m*. Of course, each such function is the restriction to $\{0,1\}^n$ of a polynomial discriminant function; what we have emphasised here is that in considering binary inputs only, some redundancy can be eliminated immediately. The set of all Boolean functions on $\{0,1\}^n$ of threshold order at most m will be denoted by $T(n,m)$. We remark that $T(n,m)$ is, in the terminology of [17], the set of functions computable by a *mask perceptron* of order at most m, on *retina* $\{0,1\}^n$.

For $x \in \mathbf{R}^n$, define the *extended vector* $\phi_m(x)$ to be the vector of length $\binom{n+m}{m}$ whose entries are x_S for $S \in [n]^m$ in some prescribed order. To be precise, we shall suppose the entries are in *lexicographic (dictionary) order*. For example, $x_{\{1,1,3\}}$ is before $x_{\{1,2,3\}}$ in the lexicographic ordering. (By default, the constant 1 preceeds all other x_S.) By way of illustration, consider the case $n = 3$ and $m = 2$. In this case,

$$\phi_2(x_1 x_2 x_3) = \left(1, x_1, x_1^2, x_1 x_2, x_1 x_3, x_2, x_2^2, x_2 x_3, x_3, x_3^2\right).$$

We observe (see [7], for example) that a function f is a polynomial discriminant function of degree at most m if and only if there is some homogeneous linear threshold function h_f, defined on real vectors of length $\binom{n+m}{m}$, such that

$$f(x) = 1 \Longleftrightarrow h_f(\phi_m(x)) = 1;$$

that is, if and only if the extended vectors corresponding to the positive examples of f and those corresponding to the negative examples can be separated by a hyperplane. Analogously, for the case of Boolean functions, for $x \in \{0,1\}^n$, let $\psi_m(x)$ be the $\{0,1\}$-vector of length $\sum_{i=0}^{m} \binom{n}{i}$ whose entries are x_S for $S \in [n]^{(m)}$ arranged in lexicographic order. Wang and Williams [26] call $\psi_m(x)$ the *m-augment* of x. Then a Boolean function f has threshold order at most m if and only if the ψ_m-vectors corresponding to the positive examples of f and those corresponding to the negative examples can be separated by a hyperplane.

3 Representational Power

In this section we quantify the representational power of the class $T(n,m)$ of boolean functions having threshold order at most m.

For a Boolean function f, define f^{m+} to be the set $\{\psi_m(x) : f(x) = 1\}$ and define f^{m-} in the analogous way, $f^{m-} = \{\psi_m(x) : f(x) = 0\}$. The following result is clear.

Lemma 1 *The Boolean function f has threshold order at most m if and only if the subsets f^{m-} and f^{m+} of the $\sum_{i=0}^{m} \binom{n}{i}$-dimensional Boolean hypercube are linearly separable.* □

Denote by $C(N,d)$ the quantity

$$C(N,d) = 2\sum_{k=0}^{d-1} \binom{N-1}{k}.$$

The following well-known result can be found in [7, 20], for example, and is known as the 'function-counting theorem'.

Lemma 2 *The number of ways in which N points in d-dimensional real space can be partitioned into two sets in such a way that the blocks of this partition are linearly separable by a hyperplane passing through the origin is at most $C(N,d)$.* □

In fact, this bound is tight, equality being obtained if the N points in question have the property that no $d+1$ of them lie in a $(d-1)$-dimensional flat; that is, if the points are in *general position.* We immediately have the following upper bound, which is implicit in [5].

Theorem 3 *The number of Boolean functions of threshold order at most m satisfies*

$$|T(n,m)| \le C\left(2^n, \sum_{i=0}^{m} \binom{n}{i}\right)$$

for all m, n with $1 \le m \le n$. □

It is fairly easy to deduce from this that $\log |T(n,m)|$ is at most $n\binom{n}{m} + O(n^m)$ as $n \to \infty$, with $m = o(n)$.

In his work, Cover [7] gives a similar upper bound on the number of ways a finite set of points in real n-dimensional space can be classified by separating surfaces described by equations of degree at most m. This upper bound is obtained in much the same way as that above, using the vectors $\phi_m(x)$ instead of $\psi_m(x)$. Cover proves that if N points $\{x_1, x_2, \ldots, x_N\}$ are such that the points $\phi_m(x_1), \phi_m(x_2), \ldots, \phi_m(x_N)$ are in general position in $\binom{n+m}{m}$-dimensional real space then the number of ways in which the x_i can be classified by means of separating surfaces of degree at most m is exactly $C(N, \binom{n+m}{m})$. This quantify is therefore an upper bound on the number of classifications of any N points (not necessarily in general position). Specialising to the case that $N = 2^n$ and the x_i are all points of $\{0,1\}^n$, the bound $|T(n,m)| \le C\left(2^n, \binom{n+m}{m}\right)$, weaker than that of Theorem 3, is obtained.

Let $\sigma(n,m)$ denote the proportion of Boolean functions of n variables with threshold order (exactly) m; thus,

$$\sigma(n,m) = \frac{|T(n,m)| - |T(n,m-1)|}{2^{2^n}}.$$

Wang and Williams [26] conjectured that for any fixed m, $\sigma(n,m) \to 0$ as $n \to \infty$. They also conjectured (as computational results seemed to suggest) that for even values of n, $\sigma(n, n/2) \to 1$ as $n \to \infty$ and that for odd values of n, $\sigma(n, (n \pm 1)/2) \to 1/2$ as $n \to \infty$. The following result proves the first conjecture and partially proves the second.

Theorem 4 *We have*

$$\lim_{n\to\infty} \frac{\left|T\left(n, \lfloor \frac{n}{2} \rfloor - 1\right)\right|}{2^{2^n}} = 0,$$

and so, for $m = m(n) \le \lfloor n/2 \rfloor - 1$, $\sigma(n, m(n)) \to 0$ as $n \to \infty$. Further, for odd n,

$$\frac{\left|T\left(n, \lfloor \frac{n}{2} \rfloor\right)\right|}{2^{2^n}} \le \frac{1}{2}.$$

Proof: We know that $|T(n,m)| \le 2\sum_{i=0}^{d(m)} \binom{2^n - 1}{i}$, where $d(m) = \sum_{i=1}^{m} \binom{n}{i}$. Now,

$$d(\lfloor n/2 \rfloor - 1) = \sum_{i=1}^{\lfloor n/2 \rfloor - 1} \binom{n}{i} \le \frac{2^n}{2} - \binom{n}{\lfloor \frac{n}{2} \rfloor} \le \frac{2^n}{2} - c\,\frac{2^n}{\sqrt{n}},$$

where c is a constant independent of n. Now, for any N and any $\lambda \le N/2$,

$$\sum_{i < \frac{N}{2} - \lambda} \binom{N}{i} < 2^N \exp(-2\lambda^2/N).$$

This result may be found in [18] and is a direct consequence of a bound of Chernoff (and also of Hoeffding's inequality). It follows that

$$|T(n, \lfloor n/2 \rfloor - 1)| < 2 \sum_{i=0}^{d\left(\lfloor \frac{n}{2} \rfloor - 1\right)} \binom{2^n}{i} < 2\, 2^{2^n} \exp\left(-2(c2^n/\sqrt{n})^2/2^n\right),$$

from which

$$\frac{\left|T\left(n, \lfloor \frac{n}{2} \rfloor - 1\right)\right|}{2^{2^n}} < 2 \exp\left(-2c^2 2^n/n\right) \to 0,$$

as $n \to \infty$. Furthermore, for n odd, we have $d\left(\lfloor n/2 \rfloor\right) = 2^n/2 - 1$ and so

$$\left| T\left(n, \left\lfloor \frac{n}{2} \right\rfloor\right) \right| < 2 \sum_{i=0}^{2^n/2-1} \binom{2^n - 1}{i} = 2\,\frac{2^{2^n - 1}}{2} = \frac{1}{2} 2^{2^n},$$

and the result follows. $\square$

It has come to my attention that Noga Alon (unpublished) has independently obtained the main part of the above result; namely, that almost all boolean functions have threshold order $\lfloor n/2 \rfloor$.

This result shows, among other things, that that the representational power of $T(n, m)$ is limited unless m is of the same order as n. The analysis shows that if a Boolean function is chosen uniformly at random from the set of all Boolean functions on n variables, then

$$\text{Prob}\left(\text{threshold order of } f < \lfloor n/2 \rfloor\right) < \exp\left(-K(2^n/n)\right),$$

for a positive constant K.

We believe that

$$\text{Prob}\left(\text{threshold order of } f > \lceil n/2 \rceil\right) \to 0,$$

as $n \to \infty$, as conjectured by Wang and Williams [26]. This could be shown to be true by proving a lower bound on $|T(n, m)|$ of $(2 - o(1)) \sum_{k=0}^{d(m)} \binom{2^n - 1}{k}$ as $n \to \infty$.

Saks [23] has reported an interesting result of Noga Alon, obtained using a result of Gotsman [12]: for some $\epsilon > 0$, almost all boolean functions of n variables have threshold order at most $(1 - \epsilon)n$.

4 The Vapnik-Chervonenkis Dimension

Suppose that H is a set of $\{0, 1\}$-valued functions defined on a set X. A finite set $T \subseteq X$ is said to be *shattered* by H if for any subset T^+ of T there is $h \in H$ such that $h(x) = 1$ for all $x \in T^+$ and $h(y) = 0$ for all $y \in T \setminus T^+$. Thus T is shattered if the restriction of H to T consists of all the $2^{|T|}$ possible $\{0, 1\}$-valued functions on T. The *Vapnik-Chervonenkis dimension* (or VC dimension) of H [25, 6], denoted $\text{VCdim}(H)$, is defined to be the largest integer k such that there is some subset of X of cardinality k shattered by H. (If no such largest k exists, we say that H has infinite Vapnik-Chervonenkis dimension.

It is well-known (see [6, 1], for example) that the Vapnik-Chervonenkis dimension of the set of homogeneous linear threshold functions is n. Indeed, we have the following useful characterisation of the shattered sets, a proof of which we include for completeness.

Lemma 5 *A subset $T = \{x_1, x_2, \ldots, x_k\}$ of $\mathbf{R}^n$ can be shattered by the set of homogeneous linear threshold functions if and only if it is a linearly independent set of vectors.*

Proof: Suppose that the vectors are linearly dependent. Then at least one of the vectors is a linear combination of the others. Without loss, suppose $x_1 = \sum_{i=2}^{k} \lambda_i x_i$ for some constants λ_i, $(1 \le i \le k)$. Suppose the vector w is such that for $2 \le j \le k$, $\langle w, x_j \rangle$ is positive if $\lambda_j > 0$ and non-positive if $\lambda_j \le 0$. Then $\langle w, x_1 \rangle = \sum_{i=2}^{k} \lambda_i \langle w, x_j \rangle \ge 0$. It follows that there is no homogeneous linear threshold function for which x_1 is a negative example and, for $2 \leq j \leq k$, x_j is a positive example if and only if $\lambda_j > 0$. That is, the set of vectors is not shattered.

For the converse, it suffices to prove the result when $k = n$. Let A be the matrix whose rows are the vectors $x_1, x_2, \ldots, x_n$ and let v be any of the 2^n vectors with entries $1, -1$. Then A is nonsingular and so the matrix equation $Aw = v$ has a solution. The homogeneous linear threshold function t defined by this solution weight-vector w satisfies $t(x_j) = 1$ if and only if entry j of v is 1. Thus all possible classifications of the set of vectors can be realised, and the set is shattered. $\square$

Recall that a set $\{f_1, f_2, \ldots, f_k\}$ of functions defined on a set X is *linearly dependent* if there are constants λ_i $(1 \leq i \leq k)$, not all zero, such that, *for all* $x \in X$,

$$\lambda_1 f_1(x) + \lambda_2 f_2(x) + \ldots + \lambda_k f_k(x) = 0;$$

that is, some non-trivial linear combination of the functions is the zero function on X. The following result is due to Dudley [10]; we present here a proof based on the idea of 'extended vectors'.

Theorem 6 *Let $\mathcal{F}$ be a real vector space of real-valued functions defined on a set X. Suppose that $\mathcal{F}$ has (vector space) dimension d. For any $f \in \mathcal{F}$, define the $\{0,1\}$-valued function f_+ on X by*

$$f_+(x) = \begin{cases} 1 & \text{if } f(x) > 0 \\ 0 & \text{if } f(x) \leq 0, \end{cases}$$

and let $\mathrm{pos}(\mathcal{F}) = \{f_+ : f \in \mathcal{F}\}$. *Then the VC dimension of* $\mathrm{pos}(\mathcal{F})$ *is d.*

Proof: Suppose that $\{f_1, f_2, \ldots, f_d\}$ is a basis for $\mathcal{F}$ and, for $x \in X$, let $x^{\mathcal{F}} = (f_1(x), f_2(x), \ldots, f_k(x))$. The subset T of X is shattered by $\mathrm{pos}(\mathcal{F})$ if and only if for each $T^+ \subseteq T$ there is $f \in \mathcal{F}$ such that $f(x) > 0$ if $x \in T^+$ and $f(x) \leq 0$ if $x \in T \setminus T^+$. But, since $\{f_1, \ldots, f_d\}$ is a basis of $\mathcal{F}$, there are constants $\alpha_i (1 \leq i \leq d)$ such that $f = \sum_{i=1}^d \alpha_i f_i$, whence this last condition is equivalent to the existence of constants α_i such that $\sum_{i=0}^d \alpha_i f_i(x) > 0$ if $x \in T^+$ and $\sum_{i=0}^d \alpha_i f_i(x) \leq 0$ if $x \in T \setminus T^+$. But this is true if and only if there is a homogeneous linear threshold function (given by the vector whose entries are the α_i) such that $t(x^{\mathcal{F}}) = 1$ if $x \in T^+$ and $t(x^{\mathcal{F}}) = 0$ if $x \in T \setminus T^+$. It follows, first, that the VC dimension of $\mathrm{pos}(\mathcal{F})$ is at most d. Secondly, it is equal to d if and only if there is a set $\{x_1^{\mathcal{F}}, \ldots, x_d^{\mathcal{F}}\}$ of linearly independent extended vectors. Suppose that this is not so. Then the vector subspace spanned by the set $\{x^{\mathcal{F}} : x \in X\}$ is of dimension at most $d-1$ and therefore is contained in some hyperplane. Hence there are constants $\lambda_1, \lambda_2, \ldots, \lambda_d$, not all zero, such that for every $x \in X$, $\sum_{i=1}^d \lambda_i (x^{\mathcal{F}})_i = 0$; that is, $\sum_{i=1}^d \lambda_i f_i(x) = 0$ for all x. But this contradicts the fact that the functions $f_1, \ldots, f_d$ are linearly independent. It follows that the VC dimension of $\mathrm{pos}(\mathcal{F})$ is d. $\square$

Let $B(n,m) = \{x_S : S \in [n]^m\}$, regarded as a set of real functions on $\mathbf{R}^n$. (Thus, we identify $x_{\{1,2\}}$ with the function $x \mapsto x_1 x_2$, for example.)

Proposition 7 *For all n and m, the set $B(n,m)$ is a linearly independent set of real functions on $\mathbf{R}^n$.*

Proof: We prove by induction on n that for all m, $B(n,m)$ is a linearly independent set of real functions on $\mathbf{R}^n$. The base case $n = 1$ is straightforward; it is well-known (see [21], for example) that the functions $1, x, x^2, x^3, \ldots, x^m$ are linearly independent.

Suppose now that the assertion is true for a value of $n \geq 1$ and let m be any positive integer. By the inductive assumption, the set $\{x_S : S \in [n]^m\}$ is a linearly independent set. For $0 \leq k \leq m$, let $S_k \subseteq [n+1]^m$ be the set of selections containing $n+1$ exactly k times. Suppose that for some constants α_S, for all $x \in \mathbf{R}^{n+1}$,

$$\sum_{S \in [n+1]^m} \alpha_S x_S = 0.$$

Then,

$$\sum_{k=0}^{m} x_{n+1}^k \sum_{S \in S_k} \alpha_S x_S^* = 0$$

for all x, where, for $S \in S_k$, x_S^* is x_S with the k factors equal to x_{n+1} deleted. (So, x_S^* is of the form x_T for some $T \in [n]^m$; that is, $x_S^* \in B(n,m)$.) It follows, from the linear independence of $1, x_{n+1}, x_{n+1}^2, \ldots, x_{n+1}^m$, that for all $x_1, x_2, \ldots, x_n$, we have

$$\sum_{S \in S_k} \alpha_S x_S^* = 0$$

for each k. But the inductive assumption then implies that for all k and for all $S \in S_k$, $\alpha_S = 0$; that is, all the coefficients α_S are zero. Hence the functions are linearly independent. $\Box$

Now, let $m \leq n$ and let $C(n,m) = \{x_S : S \in [n]^{(m)}\}$, regarded as a set of real functions on domain $\{0,1\}^n$.

Proposition 8 *For all n, m with $m \leq n$, $C(n,m)$ is a linearly independent set of real functions defined on $\{0,1\}^n$.*

Proof: Let $n \geq 1$ and suppose that for some constants α_S and for all $x \in \{0,1\}^n$,

$$A(x) = \sum_{S \in [n]^{(m)}} \alpha_S x_S = 0.$$

Set x to be the all-0 vector to deduce that $\alpha_\emptyset = 0$. Let $1 \leq k \leq m$ and assume, inductively, that $\alpha_S = 0$ for all $S \subseteq [n]$ with $|S| < k$. Let $S \subseteq [n]$ with $|S| = k$. Setting $x_i = 1$ if $i \in S$ and $x_j = 0$ if $j \notin S$, we deduce that $A(x) = \alpha_S = 0$. Thus for all S of cardinality k, $\alpha_S = 0$. Hence $\alpha_S = 0$ for all S, and the functions are linearly independent. $\Box$

Corollary 9 *For all n, m,*

$$\mathrm{VCdim}(P(n,m)) = \binom{n+m}{m},$$

and for all n, m with $m \leq n$,

$$\mathrm{VCdim}(T(n,m)) = \sum_{i=0}^{m} \binom{n}{i}.$$

Proof: Let $\mathcal{F}$ be the vector subspace of the space of all real functions on $\mathbf{R}^n$ spanned by $B(n,m)$. Since $B(n,m)$ is a linearly independent set of $\binom{n+m}{m}$ functions, it follows that $\mathcal{F}$ has vector-space dimension $|B(n,m)| = \binom{n+m}{m}$. But, clearly, $P(n,m) = \mathrm{pos}(\mathcal{F})$, in the notation of Theorem 6. By that result, $P(n,m)$ has the given VC dimension.

In a similar way, let $\mathcal{H}$ be the vector subspace of the space of all real functions defined on $\{0,1\}^n$ spanned by $C(n,m)$. Then $\mathcal{H}$ has vector-space dimension $|C(n,m)| = \sum_{i=0}^{m}\binom{n}{i}$. But $T(n,m) = \text{pos}(\mathcal{H})$ and therefore $T(n,m)$ has VC dimension $\sum_{i=0}^{m}\binom{n}{i}$. □

Note that if all inputs are binary, the VC dimension is lower than if the inputs are allowed to be arbitrary real numbers. We remark that the VC dimensions coincide for $m = 1$, the case of linear threshold functions.

Note that, since $B(n,m) = \{x_S : S \in [n]^m\}$ is a linearly independent set, it follows that for any $B' \subseteq B(n,m)$, the vector space $\langle B' \rangle$ of functions spanned by B' has dimension $|B'|$ and $\text{pos}(\langle B' \rangle)$ has VC dimension $|B'|$. For example, taking $n = 2$, the set $\text{pos}(\langle 1, x_1, x_2, x_1^2, x_2^2 \rangle)$ of 'conic' discriminators on the plane has VC dimension 5.

The determination of the VC dimension of $T(n,m)$ allows us to give a lower bound on $|T(n,m)|$. Immediate from the observation that if a set of functions H has VC dimension d then $|H| \geq 2^d$, we obtain

$$|T(n,m)| \geq 2^{\sum_{i=0}^{m}\binom{n}{i}}$$

for all m, n with $1 \leq m \leq n$. For fixed m, this is $2^{\Omega(n^m)}$. This lower bound has been obtained by Baldi [5] in a different manner. Recently, the lower bound has been improved by Saks [23], who shows that $|T(n,m)| \geq |T(n-1,m)||T(n-1,m-1)|$ for $2 \leq m \leq n-2$, from which it follows that, for fixed m, $|T(n,m)| = 2^{\Omega(n^{m+1})}$.

5 Generalisation Ability of Polynomial Surfaces

We have examined in two ways the classifying power of polynomial discriminators. First, we obtained upper bounds on the number of ways a set of points can be classified by polynomial surfaces of a given order and, secondly, we computed the VC dimension of the set of polynomial discriminants on $\mathbf{R}^n$ having at most a given degree and the VC dimension of this set restricted to $\{0,1\}^n$. But this latter measure — the VC dimension — is of great importance for another reason. It quantifies, in a precise manner described below, the *generalisation* performance of the set of classifiers.

We define 'generalisation performance' as in the *probably approximately correct* model of learning introduced by Valiant [24] and subsequently further developed by many researchers (see [6, 1]). Here, it is assumed that there is a probability measure P on (some appropriate σ-algebra of subsets of) $\mathbf{R}^n$, which is fixed but unknown to us. Suppose that we are given a training sample of s correctly classified points in $\mathbf{R}^n$, each chosen independently and at random from $\mathbf{R}^n$ according to P, so that the sample is chosen according to the product probability distribution P^s. Based on this sample and the classification of the points it contains, we wish to predict with high accuracy the classification of other randomly drawn points. More precisely, suppose that there is an underlying 'target' classification $t \in P(n,m)$ of all the data points; that is, t is represented by a polynomial surface of degree at most m which separates all positively classified points from all negatively classified points. Let us denote the training sample by $\mathbf{x} \in (\mathbf{R}^n)^s$. Suppose that we have a deterministic means — a *learning algorithm* — for hypothesising a classifier $L(\mathbf{x}) \in P(n,m)$ based on the sample $\mathbf{x}$. For simplicity, we shall denote $L(\mathbf{x})$ by $L_{\mathbf{x}}$. For $0 < \epsilon < 1$, $L_{\mathbf{x}}$ is said to be *ϵ-good* for t if the probability $P(h(x) \neq t(x))$ that h misclassifies a P-random example is less than ϵ. For L to be a valid means of generalisation, we want it to be the case that there is a high P^s-probability that $\mathbf{x}$ leads to an ϵ-good hypothesis $L_{\mathbf{x}}$.

We refer to [24, 6, 1] for further discussion of the theory of pac learning; here we simply state the following result, in which the upper bound follows from [6] in conjunction with our results on the VC dimensions, and the lower bound follows from [11] in conjunction with the VC dimensions.

Theorem 10 *Let L be any algorithm which takes as input a sample $\mathbf{x}$ of s points from $\mathbf{R}^n$, together with their classifications determined by some target function $t \in P(n,m)$, and which returns a function $L_{\mathbf{x}}$ in $P(n,m)$ which correctly classifies the points in the sample. There is a constant $K > 0$ such that for any $0 < \delta, \epsilon < 1$ and*

$$s \geq s(\delta, \epsilon) = \frac{K}{\epsilon}\left(\binom{n+m}{m}\log\left(\frac{1}{\epsilon}\right) + \log\left(\frac{1}{\delta}\right)\right),$$

the following holds: for any $t \in P(n,m)$ and any probability measure P on $\mathbf{R}^n$,

$$P^s\left(\{\mathbf{x} : P\left(L_{\mathbf{x}}(x) \neq t(x)\right) < \epsilon\}\right) > 1 - \delta.$$

Further, there is a constant $c > 0$ such that for any $0 < \delta < 1/100$ and $0 < \epsilon < 1/8$, for

$$s < \frac{1}{\epsilon}\left(c\binom{n+m}{m} + \log\left(\frac{1}{\delta}\right)\right)$$

there is a probability measure P on $\mathbf{R}^n$ and a target classification $t \in P(n,m)$ such that this P^m probability is less than $1 - \delta$. □

This indicates that for fixed desired confidence and accuracy measures δ and ϵ, one should take a random sample of size proportional to $\binom{n+m}{m}$ in order to guarantee, for any target $t \in P(n,m)$ and any probability distribution P, valid generalisation within this probabilistic framework.

An analogous result holds for $T(n,m)$, with $\binom{m+n}{m}$ replaced by $\sum_{i=0}^{m}\binom{n}{i}$. □

We remark that values of c and K are known; see the results of Blumer *et al.* [6], subsequently improved in [2].

Now that we have determined what size of sample to take, it is natural to ask how we design a suitable learning algorithm L. It is clear from the above result that any L which produces a hypothesis agreeing with the target on the sample — that is, a *consistent-hypothesis-finder* [6] — will suffice. We have, throughout this paper and in common with earlier work on separators [7, 20], made much use of the fact that polynomial separation is possible if and only if linear separation of a transformed set of 'extended vectors' is possible in a higher-dimensional space. Consequently, any consistent-hypothesis-finder for linear threshold functions can be used as a consistent-hypothesis-finder [6] for the spaces $P(n,m)$ and $T(n,m)$. Many such algorithms are known. For example, any linear programming algorithm could be used to find a function in $P(n,m)$ consistent with a sample of such a function: take the sample, form the vectors $\phi_m(x)$ for x in the sample, obtaining a training sample for a linear threshold function in $\binom{n+m}{m}$ dimensions; use the linear programming algorithm to find a suitable hyperplane, and transform back to obtain the equation of a suitable polynomial surface of degree m. (Proceed analogously for the case of Boolean inputs, using the vectors $\psi_m(x)$.) Using Karmarkar's algorithm [14], this procedure can be implemented in time polynomial in n^m. This approach will work for any consistent-hypothesis-finder for linear threshold functions, such as the perceptron learning algorithm [17]. (However, this particular algorithm is not in general efficient as a consistent-hypothesis-finder [4].)

6 Conclusions and Further Work

We have quantified in a number of ways the expressive and representational power of classification methods based on polynomial separating surfaces of given degree. We have shown that almost all Boolean functions of n variables have threshold order at least $\lfloor n/2 \rfloor$ and that in the case of odd n, at most half of the functions have threshold order $\lfloor n/2 \rfloor$. The Vapnik-Chervonenkis dimension of the set of functions realisable by polynomial separators of a given degree has been determined, both in the real case and in the Boolean case. In addition, we have quantified the size of training set required for valid generalisation within the framework of the probably approximately correct learning model and discussed how such generalisation may be achieved.

Part of the conjecture of Wang and Williams remains open. In particular, while we have shown that almost all Boolean functions on n variables have threshold order at least $\lfloor n/2 \rfloor$, we have not shown that almost all Boolean function of n variables have threshold order at most $\lceil n/2 \rceil$.

Acknowledgements I thank Graham Brightwell for helpful discussions and comments. I thank Michael Saks for helpful communications regarding the results of Alon.

References

[1] M. Anthony and N. Biggs, *Computational Learning Theory: An Introduction*, Cambridge University Press, Cambridge, UK, 1992.

[2] M. Anthony, N. Biggs and J. Shawe-Taylor, The learnability of formal concepts, in *COLT'90, Proceedings of the Third Workshop on Computational Learning Theory*, Morgan Kaufman, San Mateo, CA, 1990.

[3] M. Anthony, G. Brightwell, D. Cohen and J. Shawe-Taylor, On exact specification by examples, in *COLT'92, Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, ACM Press, New York, NY, 1992.

[4] M. Anthony and J. Shawe-Taylor, On the running time of the perceptron algorithm as a consistent-hypothesis-finder, LSE Mathematics Preprint Series LSE-MPS-29, London School of Economics, 1992.

[5] P. Baldi, Neural networks, orientations of the hypercube, and algebraic threshold functions, *IEEE Transactions on Information Theory*, 34 (3), 1988: 523–530.

[6] A. Blumer, A. Ehrenfeucht, D. Haussler and M. Warmuth, Learnability and the Vapnik-Chervonenkis Dimension. *Journal of the ACM*, 36(4), 1989: 929–965.

[7] T.M. Cover, Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition, *IEEE Trans. Electronic Computers* 14, 1965: 326–334.

[8] L. Devroye, Automatic pattern recognition: a study of the probability of error, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10(4), 1988: 530–543.

[9] R. Duda and P. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[10] R.M. Dudley, Central limit theorems for empirical measures, *Ann. Probability* 6, 1978: 899–929.

[11] A. Ehrenfeucht, D. Haussler, M. Kearns and L.G. Valiant, A general lower bound on the number of examples needed for learning, *Information and Computation* 82, 1989: 247–261.

[12] C. Gotsman, On boolean functions, polynomials and algebraic threshold functions. Technical report TR-89-18, Department of Computer Science, Hebrew University, 1989.

[13] S-T. Hu, *Threshold Logic*, University of California Press, Berkeley, 1965.

[14] N. Karmarkar, A new polynomial time algorithm for linear programming, *Combinatorica*, 4, 1984: 373–395.

[15] N. Littlestone, Learning quickly when irrelevant attributes abound: a new linear threshold learning algorithm. *Machine Learning*, 2(4), 1988: 285–318.

[16] O.L. Mangasarian, R. Setiono and W.H. Wolberg, Pattern recognition via linear programming: theory and application to medical diagnosis, in *Large-Scale Numerical Optimization*, Thomas F. Coleman and Yuying Li (Eds.), SIAM, Philadelphia 1990: 20–30.

[17] M. Minsky and S. Papert, *Perceptrons.* MIT Press, Cambridge, MA., 1969. (Expanded edition 1988.)

[18] J.W. Moon, Four combinatorial problems, in *Combinatorial Mathematics and its Applications* (ed. D.J.A. Welsh), 185-190, Academic Press, London, 1971.

[19] S. Muroga, *Threshold Logic and its Applications*, Wiley, New York, 1971.

[20] N.J. Nilsson, *Learning Machines*, McGraw-Hill, New York, 1965.

[21] A. Ostaszewski, *Advanced Mathematical Methods*, Oxford University Press, Oxford, 1991.

[22] P.J.W. Rayner and M.R. Lynch, A new connectionist model based on a non-linear adaptive filter. In *Proceedings, IEEE International Conference on Acoustics, Speech and Signal Processing, 1989.*

[23] M. Saks, Slicing the hypercube, to appear in *Surveys in Combinatorics, 1993*, a volume of invited talks at the 1993 British Combinatorial Conference, to be published by Cambridge University Press, July 1993.

[24] L.G. Valiant, A theory of the learnable. *Communications of the ACM*, 27(11), 1984: 1134–1142.

[25] V.N. Vapnik and A. Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2), 1971: 264–280.

[26] C. Wang and A.C. Williams, The threshold order of a boolean function, *Discrete Applied Mathematics*, 31, 1991: 51–69.

[27] W.H. Wolberg and O.L. Mangasarian, Multisurface method of pattern separation for medical diagnosis applied to breast cytology, *Proc. Nat. Acad. Sci. USA*, 87, 1990: 9193–9196.