Generalization Error Bounds for the Logical Analysis of Data

Martin Anthony

Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A2AE, U.K.

Abstract

This paper analyzes the predictive performance of standard techniques for the 'logical analysis of data' (LAD), within a probabilistic framework. It does so by bounding the generalization error of related polynomial threshold functions in terms of their complexity and how well they fit the training data. We also quantify the predictive accuracy in terms of the extent to which there is a large separation (a 'large margin') between (most of) the positive and negative observations.

Keywords: Logical analysis of data, LAD methods, Generalization error, Machine Learning, Learning algorithms, polynomial threshold functions

Email address: m.anthony@lse.ac.uk (Martin Anthony)

Appears in Discrete Applied Mathematics, 2012

1. Introduction

In this paper, we analyze the predictive performance of standard techniques for the 'logical analysis of data' (LAD), within a probabilistic model of learning theory. The key aim in LAD is, on the basis of some observed data points, each labeled 0 or 1, to find a way of classifying all possible data points that, it is hoped, will be largely correct. The types of classifiers produced by LAD are, at their simplest, Boolean DNF functions and, more generally, they are polynomial threshold functions (which we often refer to as LAD-type classifiers). We describe LAD and these types of classifiers in Section 2. In order to be able to make precise statements about how well these classifiers perform, we work in a standard probabilistic model of learning, which is described in Section 3. Section 3 presents results (improving on earlier results from [1]) on the predictive performance of LAD-type classifiers that agree with all the observed data points. Section 4 contains more general results that apply when some observed data might be incorrectly classified. In Section 5, we provide generalization error bounds that involve the extent to which the polynomial underlying the LAD-type classifier achieves a large separation (a 'large margin') between (most of) the positive and negative observations.

2. Logical analysis of data

We start by describing the key ingredients in the classifiers produced by LAD methods. These are Boolean functions and polynomial threshold functions.

2.1. Boolean functions

A Boolean function is simply a function from $\{0,1\}^n$ to $\{0,1\}$, for some $n \in \mathbb{N}$. Any Boolean function can be expressed by a *disjunctive normal formula* (or DNF), using *literals* $u_1, u_2, \ldots, u_n, \bar{u}_1, \ldots, \bar{u}_n$, where the $\bar{u}_i$ are *negated literals*. A disjunctive normal formula is one of the form

$$T_1 \vee T_2 \vee \cdots \vee T_k,$$

where each T_r is a *term* of the form

$$T_r = \left(\bigwedge_{i \in P} u_i \right) \bigwedge \left(\bigwedge_{j \in N} \bar{u}_j \right),$$

for disjoint subsets P, N of $\{1, 2, \ldots, n\}$. The Boolean function is said to be an l-DNF if it has a disjunctive normal formula in which the number of literals, $|P \cup N|$, in each term is at most l; it is said to be a k-term-l-DNF if there is such a formula in which, furthermore, the number of terms is at most k. We say that the DNF is *monotone* if each term contains no negated literals.

2.2. Polynomial threshold functions

It will be useful in our analysis to think of Boolean functions as being represented by the signs of polynomial expressions in the underlying variables. Let $[n]^{(d)}$ denote the set of all subsets of at most d objects from $[n] = \{1, 2, \ldots, n\}$. For any $x = (x_1, x_2, \ldots, x_n) \in \{0,1\}^n$, x_S denotes the product of the x_i for $i \in S$. For example, $x_{\{1,2\}} = x_1 x_2$. We interpret $x_\emptyset$ as the constant 1. Then, a Boolean function f of n variables is a *polynomial threshold function* of *degree* at most d (see [10, 2] for instance) if there are real numbers w_S, one for each $S \in [n]^{(d)}$, such that

$$f(x) = \operatorname{sgn} \left(\sum_{S \in [n]^{(d)}} w_S x_S \right),$$

where $\operatorname{sgn}(z) = 1$ if $z > 0$ and $\operatorname{sgn}(z) = 0$ if $z \leq 0$. We will denote the set of polynomial threshold functions on $\{0,1\}^n$ of degree d by $\mathcal{P}(n, d)$. The set $\mathcal{P}(n, 1)$ is simply known as the set of *threshold functions* on $\{0,1\}^n$. Any l-DNF f on $\{0,1\}^n$ is in $\mathcal{P}(n, l)$. For, given a term $T_j = u_{i_1} u_{i_2} \ldots u_{i_r} \bar{u}_{j_1} \bar{u}_{j_2} \ldots \bar{u}_{j_s}$ of the DNF, we can set

$$A_j = x_{i_1} x_{i_2} \ldots x_{i_r} (1 - x_{j_1})(1 - x_{j_2}) \ldots (1 - x_{j_s}).$$

Then, expanding the expression $A_1 + A_2 + \cdots + A_k$ according to the normal rules of algebra, we obtain a linear combination of the form $\sum_{S \in [n]^{(l)}} w_S x_S$.

Since $f(x) = 1$ if and only if $A_1 + A_2 + \cdots + A_k > 0$, it follows that $f(x) = \operatorname{sgn}\left(\sum_{S \in [n]^{(l)}} w_S x_S\right)$, showing that $f \in \mathcal{P}(n, l)$.

The subclass $\mathcal{B}(n, d)$ of $\mathcal{P}(n, d)$ of *binary-weight* polynomial threshold functions consists of those for which the weights w_S all belong to $\{-1, 0, 1\}$ for $S \neq \emptyset$, and for which $w_\emptyset \in \mathbb{N}$. For $1 \leq j \leq \sum_{i=0}^{d} \binom{d}{i}$, define $\mathcal{P}_j(n, d)$ to be the set of all functions in $\mathcal{P}(n, d)$ with at most j of the weights w_S non-zero for $S \neq \emptyset$. We say that the functions in $\mathcal{P}_j(n, d)$ *involve at most* j *product terms*. In an analogous way we define $\mathcal{B}_j(n, d)$, the class of binary-weight polynomial threshold functions involving at most j terms and having $w_\emptyset \in \{-j, -j+1, \ldots, -1, 0, 1, \ldots, j-1, j\}$. We have remarked that any l-DNF function lies in $\mathcal{P}(n, l)$. It is not generally true that a k-term-l-DNF lies in $\mathcal{P}_k(n, l)$, though. However. as can be seen from the translation described above between a DNF and a polynomial threshold function, if we have a monotone k-term-l-DNF, then the resulting polynomial threshold function will be in $\mathcal{B}_k(n, l)$. (For, it is simply the sum of monomials, one corresponding to each of the monotone terms of the DNF.)

2.3. Standard LAD methods

In the basic LAD framework, we are given elements of $\{0, 1\}^n$, some *observations*, classified according to some *hidden function* t: a given $x \in \{0, 1\}^n$ in the data set is classified as *positive* if $t(x) = 1$ and *negative* if $t(x) = 0$. The observations, together with the positive/negative classifications will be denoted D. The aim is to find a function h of a particular type (called a hypothesis) which fits the observations well. In a sense, such a hypotheses 'explains' the given data well and it is to be hoped that it generalizes well to other data points, so far unseen. That is, we want it to be the case that for most $y \in \{0, 1\}^n$, h classifies y correctly, meaning $h(y) = t(y)$.

The *observed error* of a hypothesis on a data set D is the proportion of observations in D incorrectly classified by the hypothesis:

$$\operatorname{er}_D(h) = \frac{1}{|D|} \left|\{x \in D : h(x) \neq t(x)\}\right|.$$

An *extension* of D (or a hypothesis *consistent* with D) is a hypothesis with zero observed error.

In the standard LAD method described in [7], a DNF is produced. First, a *support set* of variables is found. This is a set $S = \{i_1, i_2, \ldots, i_s\}$ such that no positive data point agrees with a negative data point in all the coordinates in S. If S is a support set then there is some extension of D which depends only on the literals $u_i, \bar{u}_i$ for $i \in S$ (and conversely). In the technique described in [7], a small support set is found by solving a set-covering problem derived from the data set D. This is framed in [7] as an integer linear programming problem. (This can be solved exactly or, given the well-known greedy heuristic for set-covering, a relatively small support set can be found efficiently.) Once a support set has been found, *positive patterns* are then found. A (pure) positive pattern is a conjunction of literals which is satisfied by at least one positive example in D but by no negative example. We then take as hypothesis h the disjunction of a set of positive patterns. If these patterns together cover all positive examples, then h is an extension of D. If the chosen support set has cardinality s, and each positive pattern is a conjunction of at most $d \leq s$ literals, and the number of patterns is P, then the resulting function is a P-term-d-DNF formula. There are a number of different pattern-generation algorithms, and one could look for patterns satisfying particular additional properties; see [12, 13], for instance.

There are some variants on this method. In particular, we can also make use of *negative patterns*, to make use of any commonalities among the observations that have been classified with label 0. A (pure) negative pattern is a conjunction of literals which is satisfied by at least one negative example and by no positive example. Negative patterns can be detected or generated in analogous ways to positive patterns. Suppose that $T_1, T_2, \ldots, T_q$ are patterns covering all positive examples in D and that $T'_1, T'_2, \ldots, T'_r$ are negative patterns covering all negative examples in D. Then we can form the hypothesis

$$h = \operatorname{sgn}\left(\sum_{i=1}^{q} T_i - \sum_{j=1}^{r} T'_j\right),$$

which will be an extension of D if each T_i is a pure positive pattern and each T'_j is a pure negative pattern. If each pattern is a conjunction of at most d literals, then the resulting extension lies in $\mathcal{P}(n, d)$. If, furthermore, all the positive and negative patterns involved are monotone (meaning they contain no negated literals), then the extension lies in $\mathcal{B}_P(n, d)$, where $P = q + r$ is the number of patterns. More generally, we might consider 'impure' patterns.

For instance, a particular conjunction of literals may cover many positive observations (that is, they satisfy the conjunction) but may also cover a small number of negative observations. We might well want to make use of such a pattern.

There might be some advantage in 'weighting' the patterns, assigning positive weights to the patterns and negative weights to the negative patterns; that is, we take as hypothesis a function of the form

$$h = \operatorname{sgn}\left(\sum_{i=1}^{q} w_i T_i - \sum_{j=1}^{r} w'_j T'_j\right),$$

where the w_i, w'_i are positive. For instance, we might take the weight associated to a pattern to be proportional to the number of observations it covers. Such classifiers will lie in $\mathcal{P}(n,d)$ if all patterns are of degree at most d. Without any loss, we may suppose that the representation of such a classifier as a polynomial threshold function is such that in the underlying polynomial, $f(x) = \operatorname{sgn}\left(\sum_{S\in[n]^{(d)}} w_S x_S\right)$, the vector $w = (w_S)$ is normalized, so that $\|w\|_1 = \sum_{S\in[n]^{(d)}} |w_S| = 1$.

3. Generalization from random data

It is important to know how well a hypothesis will classify further data. A standard framework for addressing this is the 'PAC' model of learning. In this framework, we assume that the data points are generated randomly according to a fixed probability distribution μ on $\{0,1\}^n$ and that they are classified by some 'target' function t. If there are m data points in D, then we may regard the data points as a vector in $(\{0,1\}^n)^m$, drawn randomly according to the product probability distribution μ^m. Given any hypothesis h, a measure of how well h performs in classification is its *error*

$$\operatorname{er}(h) = \mu\left(\{x \in \{0,1\}^n : h(x) \neq t(x)\}\right).$$

This is simply the probability that h incorrectly classifies $x \in \{0,1\}^n$ drawn randomly according to μ. (Note that such a random x may be one of the data points of D.) An important aspect of the PAC model of learning is to

formalize the intuitive idea that if a simple hypothesis is an extension of (or, at least, a good fit to) a large set of training data, then it is likely to have small error. We have the following results for LAD-type classifiers.

Theorem 3.1. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m points, each generated at random according to a fixed probability distribution on $\{0,1\}^n$. Then,* for any *$d, P \geq 1$, if h is any extension of D which is a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$ (and, in particular, if h is a monotone P-term-d-DNF), then the error of h is less than*

$$\frac{1}{m}\left(dP\ln\left(\frac{en}{d}\right)+P\ln\left(\frac{2e}{P}\right)+\ln\left(\frac{12}{\delta}\right)+2\ln d+3\ln P\right),$$

for $n \geq 2$.

Note that if $P \geq 6$, then the second term in the bound is negative.

Theorem 3.2. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m points, each generated at random according to a fixed probability distribution on $\{0,1\}^n$. Then,* for any *$d, P \geq 1$ with $P \leq 2m$, if h is an extension of D which is a polynomial threshold function in $\mathcal{P}_P(n,d)$, the error of h is less than*

$$\frac{1}{m}\left(2dP\log_2\left(\frac{en}{d}\right)+2P\log_2(2m)+4P\log_2\left(\frac{e}{P}\right)+2\log_2\left(\frac{8}{\delta}\right)+2\log_2(dP)\right).$$

Note that P and d are *not* specified in advance in these results, and may be observed after learning. (Note also that since we certainly have $P \leq m$ for the standard LAD methods, the restriction $P \leq 2m$ is benign.)

Proof of Theorem 3.1: We use a standard bound (which can be found in [6], for example): given a class of hypotheses H, for a random sample of m points, each generated according to μ, the probability that there is

some extension $h \in H$ which has error at least ϵ is less than $|H| \exp(-\epsilon m)$. Recall that $h \in \mathcal{B}_P(n,d)$ if for some $j \leq P$ there are non-empty subsets $S_1, S_2, \ldots, S_j$ of $\{1,2,\ldots,n\}$, each of cardinality at most d, and constants $w_1, w_2, \ldots, w_j \in \{-1,1\}$ and $w_0 \in \{-P, \ldots, P\}$ such that

$$h(x) = \mathrm{sgn}\left(w_0 + \sum_{i=1}^{j} w_i x_{S_i}\right).$$

The number of possible such x_S is

$$N = \binom{n}{\leq d} = \sum_{i=0}^{d} \binom{n}{i}.$$

We will make use of the following inequality (see [6], for instance):

$$\sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d.$$

To count the number of functions in $\mathcal{B}_P(n,d)$, we observe that, given the (non-empty) product terms which such an h involves, there are two choices for the weight assigned to each (either -1 or 1). Furthermore, there are $2P+1 \leq 3P$ choices for w_0. Therefore

$$\begin{aligned} |\mathcal{B}_P(n,d)| &\leq 3P \sum_{j=0}^{P} \binom{N}{j} 2^j \\ &< 3P\, 2^P \sum_{j=0}^{P} \binom{N}{j} \\ &\leq 3P\, 2^P \left(\frac{eN}{P}\right)^P. \end{aligned}$$

It follows that

$$\ln |\mathcal{B}_P(n,d)| \leq \ln(3P) + P \ln\left(\frac{2e}{P}\right) + P \ln N \leq \ln(3P) + P \ln\left(\frac{2e}{P}\right) + Pd \ln\left(\frac{en}{d}\right).$$

So, fixing P, d, taking H to be $\mathcal{B}_P(n,d)$, and using the bound mentioned at the start of the proof, we have that, with probability at least $1-\delta$, if $h \in H$

is an extension of a random data set D of size m, then

$$\mathrm{er}(h) < \frac{dP\ln(en/d) + P\ln(2e/P) + \ln(3P/\delta)}{m}.$$

It follows that only with probability at most $1 - \delta/(4d^2P^2)$, will there be some $h \in \mathcal{B}_P(n,d)$ which is an extension of D and which satisfies $\mathrm{er}(h) > \epsilon(d,P,n,m)$ where

$$\epsilon(d,P,n,m) = \frac{1}{m}\left(dP\ln(en/d) + P\ln(2e/P) + \ln\left(\frac{12d^2P^3}{\delta}\right)\right).$$

So, the probability that for *some* $d, P \geq 1$, there will be some such h is no more than

$$\sum_{d=1}^{\infty}\sum_{P=1}^{\infty}\frac{\delta}{4d^2P^2} = \frac{\delta}{4}\sum_{d=1}^{\infty}\frac{1}{d^2}\sum_{P=1}^{\infty}\frac{1}{P^2} = \frac{\delta}{4}\left(\frac{\pi^2}{6}\right)^2 < \delta.$$

The result follows. □

Proof of Theorem 3.2: We use a bound from [6], which follows [14]. With the notation as above, the bound states that for any positive integer $m \geq 8/\epsilon$ and any $\epsilon \in (0,1)$, the probability that there exists $h \in H$ with $\mathrm{er}(h) \geq \epsilon$ and such that h is consistent with a randomly generated data set of size m is less than $2\Pi_H(2m)2^{-\epsilon m/2}$, where for a positive integer k, $\Pi_H(k)$ is the maximum possible cardinality of H domain-restricted to a k-subset of $\{0,1\}^n$. (The function Π_H is known as the growth function.) We now bound the growth function of $H = \mathcal{P}_P(n,d)$.

As usual, let $[n]^{(d)}$ be the set of all subsets of $\{1,2,\ldots,n\}$ of cardinality at most d and, for $\mathcal{R} \subseteq [n]^{(d)}$, let $H^{\mathcal{R}}$ be the set of polynomial threshold functions of the form

$$\mathrm{sgn}\left(\sum_{S\in\mathcal{R}} w_S x_S\right).$$

Then

$$H = \bigcup_{\mathcal{R}\subseteq[n]^{(d)},|\mathcal{R}|\leq P} H^{\mathcal{R}}.$$

For a subset C of $\{0,1\}^n$, let $H|_C$ denote H restricted to domain C. Then, for C such that $|C| = k$,

$$|H|_C| = \left| \bigcup_{\mathcal{R} \subseteq [n]^{(d)}, |\mathcal{R}| \leq P} H^{\mathcal{R}}|_C \right| \leq \sum_{\mathcal{R} \subseteq [n]^{(d)}, |\mathcal{R}| \leq P} \left| H^{\mathcal{R}}|_C \right| \leq \sum_{\mathcal{R} \subseteq [n]^{(d)}, |\mathcal{R}| \leq P} \Pi_{H^{\mathcal{R}}}(k),$$

from which it follows that

$$\Pi_H(k) \leq \sum_{\mathcal{R} \subseteq [n]^{(d)}, |\mathcal{R}| \leq P} \Pi_{H^{\mathcal{R}}}(k).$$

The number of such $\mathcal{R}$ is $\sum_{r=0}^{P} \binom{N}{r}$ where $N = \sum_{i=1}^{d} \binom{n}{i}$. Fix $\mathcal{R} \subseteq [n]^{(d)}$, of cardinality $r \leq P$. Given a set G of functions from a (not necessarily finite) set X to $\{0,1\}$, the *VC-dimension*, $\mathrm{VCdim}(G)$, of G (introduced in [15]) is the largest integer Δ such that for some set C of cardinality Δ, $|G|_C| = 2^{\Delta}$. From Sauer's inequality [11], if $m \geq \Delta \geq 1$,

$$\Pi_G(m) \leq \sum_{i=0}^{\Delta} \binom{m}{i} \leq \left(\frac{em}{\Delta}\right)^{\Delta}.$$

It can be shown (see [2], for example) that the VC-dimension of $H^{\mathcal{R}}$ is $|\mathcal{R}| = r \leq P$, so, for each $\mathcal{R}$ under consideration,

$$\Pi_{H^{\mathcal{R}}}(k) \leq \sum_{i=0}^{r} \binom{k}{i} \leq \sum_{i=0}^{P} \binom{k}{i} \leq \left(\frac{ek}{P}\right)^{P}.$$

Hence,

$$\Pi_H(k) \leq \sum_{\mathcal{R} \subseteq [n]^{(d)}, |\mathcal{R}| \leq P} \left(\frac{ek}{P}\right)^{P} \leq \sum_{i=0}^{P} \binom{N}{i} \left(\frac{ek}{P}\right)^{P} \leq \left(\frac{eN}{P}\right)^{P} \left(\frac{ek}{P}\right)^{P},$$

so

$$\ln \Pi_H(k) \leq P \ln k + Pd \ln\left(\frac{en}{d}\right) + 2P \ln\left(\frac{e}{P}\right),$$

where we have used the fact that $N \leq (en/d)^d$.

So, with probability at least $1 - \delta$, if $h \in H$ is an extension of a random data set D of size m, then

$$\mathrm{er}(h) < \frac{2Pd \log_2(en/d) + 2P \log_2(2m) + 4P \log_2(e/P) + 2 \log_2(2/\delta)}{m}.$$

So, the probability that for *some* $d, P \geq 1$, there will be some $h \in \mathcal{P}_P(n,d)$ consistent with D and with error at least

$$\frac{1}{m}\left(2Pd\log_2(en/d) + 2P\log_2(2m) + 4P\log_2(e/P) + 2\log_2\left(8d^2P^2/\delta\right)\right)$$

is less than $\delta/(4d^2P^2)$. The proof may now be completed as in the previous proof. □

4. Bounds involving observed error

We now develop some more general results. In particular, we bound the error for non-extensions in terms of the observed error. We also jettison the assumption that there is a deterministic target concept giving correct classifications: we do this by assuming that D is now a set of labeled data points and that the labeled data are generated by a fixed probability distribution μ on the set $Z = X \times \{0,1\}$ (rather than just on X), where $X = \{0,1\}^n$. Then, the error of a hypothesis h is simply $\mathrm{er}(h) = \mu\{(x,y) : h(x) \neq y\}$ and the observed error is

$$\mathrm{er}_D(h) = \frac{1}{|D|}\left|\{(x,y) \in D : h(x) \neq y\}\right|.$$

We present two types of results. The first type of (high-probability) bound takes the form $\mathrm{er}(h) < \mathrm{er}_D(h) + \epsilon_1$ and the second $\mathrm{er}(h) < 3\,\mathrm{er}_D(h) + \epsilon_2$ where, generally, $\epsilon_2 < \epsilon_1$.

Theorem 4.1. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Then,* for any *$d, P \geq 1$, if h is a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$ (and, in particular, if h is a monotone P-term-d-DNF),*

$$\mathrm{er}(h) < \mathrm{er}_D(h) + \sqrt{\frac{1}{2m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + 2\ln(dP) + \ln\left(\frac{24P}{\delta}\right)\right)}.$$

Proof: We use the fact (which follows from a Hoeffding bound: see [4] for instance) that, for a finite hypothesis class H, with probability at least $1 - 2|H|e^{-2m\epsilon^2}$, for all $h \in H$, we have $|\mathrm{er}(h) - \mathrm{er}_D(h)| < \epsilon$. Using the fact that when $H = \mathcal{B}_P(n, d)$,

$$\ln |H| \leq \ln(3P) + P \ln \left(\frac{2e}{P} \right) + Pd \ln \left(\frac{en}{d} \right),$$

we see that, for any d, P, with probability only at most $1 - \delta/(4d^2P^2)$ will there be some $h \in \mathcal{B}_P(n, d)$ with $\mathrm{er}(h) \geq \mathrm{er}_D(h) + \epsilon$, where

$$\epsilon = \sqrt{\frac{1}{2m} \left(dP \ln \left(\frac{en}{d} \right) + P \ln \left(\frac{2e}{P} \right) + 2 \ln(dP) + \ln \left(\frac{24P}{\delta} \right) \right)}.$$

The result follows since $\sum_{d,P=1}^{\infty} \delta/(4d^2P^2) < \delta$. □

Theorem 4.2. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Then,* for any $d, P \geq 1$ *with $P \leq 2m$, if h is a polynomial threshold function in $\mathcal{P}_P(n, d)$,*

$$\mathrm{er}(h) < \mathrm{er}_D(h) + \sqrt{\frac{8}{m} \left(dP \ln \left(\frac{en}{d} \right) + P \ln(2m) + 2P \ln \left(\frac{e}{P} \right) + 2 \ln(dP) + \ln \left(\frac{16}{\delta} \right) \right)}.$$

Proof: We use the following result of Vapnik and Chervonenkis [15, 4]: with probability at least $1 - 4\Pi_H(2m)e^{-\epsilon^2 m/8}$, for all $h \in H$, $|\mathrm{er}(h) - \mathrm{er}_D(h)| < \epsilon$. Using the fact that when $H = \mathcal{P}_P(n, d)$,

$$\ln \Pi_H(k) \leq P \ln k + Pd \ln \left(\frac{en}{d} \right) + 2P \ln \left(\frac{e}{P} \right),$$

we see that, for any d, P, with probability only at most $1 - \delta/(4d^2P^2)$ will there be some $h \in \mathcal{P}_P(n, d)$ with $\mathrm{er}(h) \geq \mathrm{er}_D(h) + \epsilon'$, where

$$\epsilon' = \sqrt{\frac{8}{m} \left(dP \ln \left(\frac{en}{d} \right) + P \ln(2m) + 2P \ln \left(\frac{e}{P} \right) + 2 \ln(dP) + \ln \left(\frac{16}{\delta} \right) \right)}.$$

The result follows.

We now remove the square roots in the second (more general) bound, at the expense of replacing $\mathrm{er}_D(h)$ by $3\,\mathrm{er}_D(h)$. If the observed error is small, the resulting bound will be better. We use the following result.

Theorem 4.3. *Let δ be any positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose H is some set of functions from a domain X into $\{0,1\}$. Suppose D is a data set of m labeled points (x,b) of $Z = X \times \{0,1\}$, each generated at random according to a fixed probability distribution on Z. Then, for all $h \in H$,*

$$\mathrm{er}(h) < 3\,\mathrm{er}_D(h) + \frac{4}{m}\left(\ln(\Pi_H(2m)) + \ln\left(\frac{4}{\delta}\right)\right)$$

where Π_H is the growth function of H.

Proof: A theorem of Vapnik [14] shows that, for any ξ, with probability at least $1 - 4\,\Pi_H(2m)\,e^{-m\xi^2/4}$, for all $h \in H$,

$$\frac{\mathrm{er}(h) - \mathrm{er}_D(h)}{\sqrt{\mathrm{er}(h)}} < \xi.$$

So, with probability at least $1-\delta$, for all $h \in H$,

$$\mathrm{er}(h) < \mathrm{er}_D(h) + \alpha\sqrt{\mathrm{er}(h)},$$

where

$$\alpha = \sqrt{\frac{4}{m}\left(\ln(\Pi_H(2m)) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

Fix h and let $\beta = \mathrm{er}_D(h)$ and $z = \sqrt{\mathrm{er}(h)}$. Then, if $\mathrm{er}(h) < \mathrm{er}_D(h) + \alpha\sqrt{\mathrm{er}(h)}$, we have $z^2 - \alpha z - \beta < 0$, and

$$\left(z - \frac{\alpha}{2}\right)^2 = z^2 - \alpha z + \frac{\alpha^2}{4} = (z^2 - \alpha z - \beta) + \left(\frac{\alpha^2}{4} + \beta\right) < \frac{\alpha^2}{4} + \beta.$$

It follows that

$$\begin{aligned}
\mathrm{er}(h) &= z^2 = \left(\left(z - \frac{\alpha}{2}\right) + \frac{\alpha}{2}\right)^2 \\
&\leq \left(z - \frac{\alpha}{2}\right)^2 + \frac{\alpha^2}{4} + \alpha\left(z - \frac{\alpha}{2}\right) \\
&< \frac{\alpha^2}{4} + \beta + \frac{\alpha^2}{4} + \alpha\sqrt{\frac{\alpha^2}{4} + \beta} \\
&\leq \frac{\alpha^2}{2} + \beta + 2\sqrt{\frac{\alpha^2}{4} + \beta}\sqrt{\frac{\alpha^2}{4} + \beta} \\
&= \alpha^2 + 3\beta \\
&= \frac{4}{m}\left(\ln(\Pi_H(2m)) + \ln\left(\frac{4}{\delta}\right)\right) + 3\,\mathrm{er}_D(h).
\end{aligned}$$

So, with probability at least $1 - \delta$, for all $h \in H$,

$$\mathrm{er}(h) < 3\,\mathrm{er}_D(h) + \frac{4}{m}\left(\ln(\Pi_H(2m)) + \ln\left(\frac{4}{\delta}\right)\right),$$

as required. □

We then have the following bounds.

Theorem 4.4. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Then,* for any *$d, P \geq 1$, if h is a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$ (and, in particular, if h is a monotone P-term-d-DNF),*

$$\mathrm{er}(h) < 3\,\mathrm{er}_D(h) + \frac{4}{m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + 2\ln(dP) + \ln\left(\frac{48P}{\delta}\right)\right).$$

Proof: We first note that $\Pi_H(2m) \leq |H|$ and then observe that, by Theorem 4.3, and using our earlier bound for the cardinality of $H = \mathcal{B}_P(n,d)$, the following holds: for each possible choice of d, P, with probability only at most

$\delta/(4d^2P^2)$ will there be some $h \in H = \mathcal{B}_P(n,d)$ such that $\mathrm{er}(h) \geq 3\,\mathrm{er}_D(h)+\epsilon$ where

$$\epsilon = \frac{4}{m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + \ln\left(\frac{48P^3d^2}{\delta}\right)\right).$$

□

Theorem 4.5. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Then,* for any *$d, P \geq 1$ with $P \leq 2m$, if h is a polynomial threshold function in $\mathcal{P}_P(n,d)$,*

$$\mathrm{er}(h) < 3\,\mathrm{er}_D(h)+\frac{4}{m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln(2m) + 2P\ln\left(\frac{e}{P}\right) + 2\ln(dP) + \ln\left(\frac{16}{\delta}\right)\right).$$

Proof: We observe that, by Theorem 4.3, and using our earlier bound on growth function, for each possible choice of d, P, with probability only at most $\delta/(4d^2P^2)$ will there be some $h \in \mathcal{P}_P(n,d)$ such that $\mathrm{er}(h) \geq 3\,\mathrm{er}_D(h)+\epsilon$ where

$$\epsilon = \frac{4}{m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln(2m) + 2P\ln\left(\frac{e}{P}\right) + \ln\left(\frac{16d^2P^2}{\delta}\right)\right).$$

□

5. Margin-based results

We now turn attention to bounding the error when we take into account the margin, which involves the value (and not just the sign) of the polynomial $\sum_{S\in[n]^{(d)}} w_S x_S$ that underlies the LAD-type classifier in $\mathcal{P}(n,d)$. (This, recall, is the polynomial arising from the discriminant $\sum_{i=1}^q T_i - \sum_{j=1}^r T'_j$ or, more generally, the weighted discriminant $\sum_{i=1}^q w_i T_i - \sum_{j=1}^r w'_j T'_j$.)

Suppose, then, that $h = \mathrm{sgn}(f)$ where $f = \sum_{S\in[n]^{(d)}} w_S x_S$. Without any loss of generality, we may assume that the coefficients in f have been normalized,

so that the 1-norm of the weight vector satisfies $\|w\|_1 = \sum_{S \in [n]^{(d)}} |w_S| = 1$. For $\gamma > 0$, we define the error of h on D at margin γ to be

$$\mathrm{er}_D^\gamma(h) = \frac{1}{|D|} \left| \{(x,y) \in D : yf(x) < \gamma\} \right|.$$

So, this is the proportion of data points in D for which either $h(x) = \mathrm{sgn}(f(x)) \neq y$, or for which $h(x) = y$ *but* $|f(x)| < \gamma$. (So, for (x,y) to contribute nothing to the margin error we need not only that the sign of $f(x)$ be correct, but that its value $|f(x)|$ be at least γ.) Clearly, $\mathrm{er}_D^\gamma(h) \geq \mathrm{er}_D(h)$.

We can bound the generalization error of polynomial threshold classifiers in terms of their margin error. However, it is possibly more useful to obtain a different type of error bound which does not involve the 'hard' margin error just described, but which instead takes more account of the distribution of the margins among the sample points. (A bound involving standard margin error then directly follows.)

For a fixed $\gamma > 0$, let $\phi^\gamma : \mathbb{R} \to [0,1]$ be given by

$$\phi^\gamma(z) = \begin{cases} 1 & \text{if } z \leq 0 \\ 1 - z/\gamma & \text{if } 0 < z < \gamma \\ 0 & \text{if } z \geq \gamma, \end{cases}$$

For a data-set D of size m, consisting of labeled points (x_i, y_i) and for a hypothesis $h = \mathrm{sgn}(f)$, let

$$\hat{\phi}_D^\gamma(h) = \frac{1}{m} \sum_{i=1}^{m} \phi^\gamma(y_i f(x_i)).$$

If h misclassifies (x_i, y_i) (that is, $h(x_i) \neq y_i$), then $\phi^\gamma(y_i f(x_i)) = 1$. If h classifies (x_i, y_i) correctly and with margin at least γ, so that $y_i f(x_i) \geq \gamma$, then $\phi^\gamma(y_i f(x_i)) = 0$. If, however, h classifies (x_i, y_i) correctly but *not* with margin at least γ, so that $0 < y_i f(x_i) < \gamma$, then $\phi^\gamma(y_i f(x_i)) = 1 - (y_i f(x_i))/\gamma$, which is strictly between 0 and 1. We have $\hat{\phi}_D^\gamma(h) \leq \mathrm{er}_D^\gamma(h)$. For, in the case where $0 < y_i f(x_i) < \gamma$, we obtain a contribution of $1/m$ to $\mathrm{er}_D^\gamma(h)$ but only a contribution of $(1/m)(1 - y_i f(x_i)/\gamma)$ to $\hat{\phi}_D^\gamma(h)$. We now obtain (high-probability) generalization error bounds of the form

$$\mathrm{er}(h) < \hat{\phi}_D^\gamma(h) + \epsilon.$$

Such bounds are potentially more useful when h achieves a large margin on many (though not necessarily all) of the data points.

We have the following result, obtained using results from [8, 5, 9].

Theorem 5.1. *Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$. Suppose that D is a data set of m points, each generated at random according to a fixed probability distribution on $\{0,1\}^n$. Then,* for any $d \geq 1$ *and* for any $\gamma > 0$*, if h is a polynomial threshold function in $\mathcal{P}(n,d)$ (normalized as indicated above, so that the weight vector has 1-norm equal to 1), then*

$$\mathrm{er}(h) < \hat{\phi}_D^\gamma(h) + \epsilon'(m,d,P,n,\gamma),$$

where

$$\epsilon'(m,d,P,n,\gamma) = \frac{4}{\gamma}\sqrt{\frac{2d}{m}\ln\left(\frac{2en}{d}\right)} + \sqrt{\frac{1}{2m}\left(\ln\left(\frac{4}{\delta}\right) + 2\ln\log_2\left(\frac{4}{\gamma}\right) + 2\ln d\right)}.$$

Proof: Let $H = \mathcal{P}(n,d)$ be the set of polynomial threshold functions of degree at most d on $\{0,1\}^n$. Let F_d denote the set of polynomials of the form $f = \sum_{S\in[n]^{(d)}} w_S x_S$, where $\|w\|_1 = 1$. As noted in [8], a result from [5] implies (on noting that ϕ^γ has a Lipschitz constant of $1/\gamma$) that, for fixed γ and d, and for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$: for all $h \in H$,

$$\mathrm{er}(h) < \hat{\phi}_D^\gamma(h) + \frac{2}{\gamma}R_m(F_d) + \sqrt{\frac{\ln(2/\delta)}{2m}},$$

where $R_m(F_d)$ is the *Rademacher complexity* of F_d. Consider, for $x \in \{0,1\}^n$, the vector $x^{(d)}$ whose entries are (in some prescribed order) x_S for all S of cardinality at most d. The set of all such $x^{(d)}$ forms a subset of $\{0,1\}^N$ where $N = \sum_{i=0}^d \binom{n}{i}$. We may consider the set F_d as a (domain-restriction of) a subset of the set $\mathcal{G}$ of all linear functions $y \mapsto \langle \alpha, y\rangle$ on $\{0,1\}^N$ defined by weight vectors α with $\|\alpha\|_1 = 1$ (this because of normalization). It will then follow by the definition of Rademacher complexity and the fact that it is non-decreasing with respect to containment of the function class [5] that

$R_m(F_d) \leq R_m(\mathcal{G})$. To bound $R_m(\mathcal{G})$ we use a result from [8]. This shows that

$$R_m(\mathcal{G}) \leq \sqrt{\frac{2\ln(2N)}{m}},$$

which, since $N \leq (en/d)^d$, gives

$$R_m(F_d) \leq \sqrt{\frac{2d}{m}\ln\left(\frac{2en}{d}\right)}.$$

To obtain a result that holds simultaneously for all γ, one can use the technique deployed in the proof of Theorem 2 in [8], or use Theorem 9 of [3]. Note that we may assume $\gamma \leq 1$ since if $\gamma > 1$, then $\hat{\phi}_D^\gamma(h) = 1$ (by the normalization assumption) and the error bound is then trivially true. We obtain the following, for fixed d: with probability at least $1-\delta$, for *all* $\gamma \in (0,1]$, if $h = \mathrm{sgn}(f)$ where $f \in F_d$ then

$$\mathrm{er}(h) < \hat{\phi}_D^\gamma(h) + \frac{4}{\gamma}\sqrt{\frac{2d}{m}\ln\left(\frac{2en}{d}\right)} + \sqrt{\frac{1}{2m}\left(\ln\left(\frac{2}{\delta}\right) + 2\ln\log_2\left(\frac{4}{\gamma}\right)\right)}.$$

The theorem now follows by using the same sort of methods as before to move to a bound in which d is not prescribed in advance: we simply replace δ by $\delta/(2d^2)$. □

Acknowledgements

This work was carried out while I was visiting RUTCOR, Rutgers University and I thank Endre Boros for helpful discussions. My work is supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. An earlier version of this paper contained some errors, and I thank Hans Simon for drawing this to my attention. (There was a corresponding error also in [1]. Theorems 3.1 and 3.2 correct and extend the results of that paper.)

References

[1] M. Anthony. Accuracy of techniques for the logical analysis of data. *Discrete Applied Mathematics* 96, 247–257, 1999.

[2] M. Anthony. Classification by polynomial surfaces. *Discrete Applied Mathematics*, 61 (1995): 91–103.

[3] M. Anthony. Generalization error bounds for threshold decision lists. *Journal of Machine Learning Research* 5, 2004, 189–217.

[4] M. Anthony and P. L. Bartlett (1999). *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge UK.

[5] P.L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research* 3, 463–482, 2002.

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

[7] Y. Crama, P.L. Hammer and T. Ibaraki. Cause-effect relationships and partially defined Boolean functions. *Annals of Operations Research*, 16: 299–325, 1988.

[8] Sham Kakade, Karthik Sridharan, Ambuj Tewari. On the Complexity of Linear Prediction: Risk Bounds, Margin Bounds and Regularization. In Daphne Koller, Dale Schuurmans, Yoshua Bengio, Léon Bottou (Eds.), *Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 8-11, 2008.* MIT Press 2009.

[9] V. Koltchninskii and D. Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics* 30(1), 1–50, 2002.

[10] M. Saks. Slicing the hypercube. In *Surveys in Combinatorics, 1993* (ed. Keith Walker) (1993), pp 211-255. Cambridge University Press, Cambridge, UK.

[11] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.

[12] Gabriela Alexe and Peter L. Hammer. Spanned patterns for the logical analysis of data. *Discrete Applied Mathematics*, 154 (7): 1039-1049, 2006.

[13] Sorin Alexe and Peter L. Hammer. Accelerated algorithm for pattern detection in logical analysis of data. *Discrete Applied Mathematics*, 154 (7): 1050-1063, 2006.

[14] V. N. Vapnik, *Estimation of Dependences Based on Empirical Data*, New York: Springer-Verlag, 1982.

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