RUTCOR
RESEARCH
REPORT

Generalization Error Bounds for Logical Analysis of Data

Martin Anthony[a]

RRR 1-2011, January 2011

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/∼rrr

[a]Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. m.anthony@lse.ac.uk

Rutcor Research Report
RRR 1-2011, January 2011

Generalization Error Bounds for Logical Analysis of Data

Martin Anthony

Abstract. This report analyses the predictive performance of standard techniques for the 'logical analysis of data' (LAD), within a probabilistic framework. Improving and extending earlier results, we bound the generalization error of classifiers produced by standard LAD methods in terms of their complexity and how well they fit the training data. We also obtain bounds on the predictive accuracy which depend on the extent to which the underlying LAD discriminant function achieves a large separation (a 'large margin') between (most of) the positive and negative observations.

Acknowledgements: This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Thanks to Endre Boros and John Shawe-Taylor for useful discussions. Part of this work was carried out while the author was visiting RUTCOR, Rutgers University.

1 Logical analysis of data

1.1 Boolean functions

A Boolean function (of n variables) is usually taken to be a function from $\{0,1\}^n$ to $\{0,1\}$. Sometimes it is useful to regard a Boolean function as a mapping from $\{-1,1\}^n$ to $\{0,1\}$. When taking the first approach, we say that we are using the *standard* convention, and we shall refer to the latter as the *nonstandard* convention. Transforming from standard to nonstandard conventions is simple. Recall that 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 known as *negated literals*. A disjunctive normal formula is one of the form

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

where each T_l is a *term* of the form

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

for some disjoint subsets P, N of $\{1, 2, \ldots, n\}$. A Boolean function is said to be an l-DNF if it has a disjunctive normal formula in which, for each term, the number of literals, $|P \cup N|$, 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 T_i is at most k.

1.2 Polynomial threshold functions

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 shall denote the product of the x_i for $i \in S$. For example, $x_{\{1,2,3\}} = x_1 x_2 x_3$. When $S = \emptyset$, the empty set, we interpret x_S as the constant 1. With this notation, a Boolean function f defined on $\{0,1\}^n$ is a *polynomial threshold function* of *degree* (at most) d if there are real numbers w_S, one for each $S \in [n]^{(d)}$, such that

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

This may be written

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

where the *sign function* sgn is such that $\text{sgn}(x) = 1$ if $x > 0$ and $\text{sgn}(x) = 0$ if $x \leq 0$. The set of polynomial threshold functions on $\{0,1\}^n$ of degree d will be denoted by $\mathcal{P}(n,d)$. The class $\mathcal{P}(n,1)$ is usually known simply as the set of *threshold functions* on $\{0,1\}^n$. It is easy to see that any l-DNF f on $\{0,1\}^n$ is in $\mathcal{P}(n,l)$, as follows. 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 form the expression

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

We do this for each term $T_1, T_2, \ldots, T_k$ and expand the algebraic expression $A_1+A_2+\cdots+A_k$ according to the normal rules of algebra, until we obtain a linear combination of the form $\sum_{S\in[n]^{(l)}} w_S x_S$. Then, since $f(x) = 1$ if and only if $A_1 + A_2 + \cdots + A_k > 0$, it follows that

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

so $f \in \mathcal{P}(n,l)$.

The class $\mathcal{B}(n,d)$ of *binary-weight* polynomial threshold functions to be the functions in $\mathcal{P}(n,d)$ for which the weights w_S all belong to $\{-1,0,1\}$ for $S \neq \emptyset$, and for which $w_\emptyset \in \mathbb{N}$ (where $\mathbb{N}$ is the set of natural numbers). Next, 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 $S \neq \emptyset$; thus a function is in $\mathcal{P}_j(n,d)$ if and only if 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 $w_0, w_1, w_2, \ldots, w_j$ such that

$$f(x) = 1 \Longleftrightarrow w_0 + \sum_{i=1}^{j} w_i x_{S_i} > 0.$$

We shall say that the functions in $\mathcal{P}_j(n,d)$ *involve at most j product terms*. In an analogous way we can define $\mathcal{B}_j(n,d)$, the class of binary-weight polynomial threshold functions involving at most j terms w_S where $S \neq \emptyset$, and which have $w_\emptyset \in \{0,1,\ldots,j-1\}$. We have remarked that any l-DNF function lies in $\mathcal{P}(n,l)$; in fact, it lies in the subclass $\mathcal{B}(n,l)$. When using the standard convention for Boolean functions, it is not generally true that a k-term-l-DNF lies in $\mathcal{B}_k(n,l)$; all that can be said is that it lies in $\mathcal{B}(n,l)$; however, if we use the nonstandard convention, it *is* the case that $f \in \mathcal{P}_k(n,l)$. For, instead of replacing a negated literal $\bar{u}_i$ in a term by the algebraic expression $1 - x_i$, we replace it simply by $-x_i$; it is clear that the product terms of the resulting polynomial threshold function are in one-to-one correspondence with the terms of the DNF formula and that they have precisely the same degree. (We take $w_\emptyset = j - 1$ where $j \leq k$ is the number of terms in the DNF.)

1.3 Standard LAD methods

In the simplest LAD framework, one is given some elements of $\{0,1\}^n$, 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, simple, 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 generalises well to other data points, so far unseen. That is, we should like it to be the case that for most $y \in \{0,1\}^n$ which are not in D, h classifies y correctly, by which we mean $h(y) = t(y)$.

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

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

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

In the basic 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 the coordinates $i_1, i_2, \ldots, i_s$. 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. Once a support set has been found, one then looks for *positive patterns*. 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. Suppose that the chosen support set has cardinality s, that each positive pattern is a conjunction of at most $d \leq s$ literals, and that the number of patterns is P; then the resulting function is a P-term-d-DNF formula.

There are some variants on this method. In particular, we can also make use of *negative patterns*. A (pure) negative pattern is a conjunction of literals which is satisfied by at least one negative example and by no positive example. 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 the function

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

is easily seen to be an extension of D. If each pattern and negative pattern is a conjunction of at most d literals, then the resulting 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 extension 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 the subclass of $\mathcal{P}(n,d)$ consisting of homogenous polynomial threshold functions, those in which the constant term $w_\emptyset$ is 0. Without any loss, we may suppose that the weights are normalised so that $\sum_{i=1}^{q} |w_i| + \sum_{j=1}^{r} |w'_j| = P$. (The non-weighted discriminant can be thought of also as a homogeneous polynomial threshold function having ± 1 weights and will also be normalised according to this definition.) If we use weights in this manner, it may be easier to 'update' the extension should we subsequently be presented with more classified data points. Note that the total number of patterns used by the LAD method described above is certainly no more than m, the number of data points.

2 Generalisation from random data

Given an extension of a fairly large data set determined by LAD techniques, it is important to know how well it would classify further data points. We can apply some probabilistic techniques to analyse the performance of LAD algorithms on random data. Following the PAC model of computational learning theory, 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 hidden function t. Thus, 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 extension h of a data set D (which it will be presumed belongs to some hypothesis space), we measure how well h performs on further examples by means of its *error*

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

which is the probability that h incorrectly classifies an $x \in \{0,1\}^n$ drawn randomly according to μ. (Note that such a random x may be one of the data points of D.)

The following results are improvements of ones from [1].

Theorem 2.1 *Suppose that D is a data set of m points, each generated at random according to a fixed probability distribution on $\{0,1\}^n$. Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$:* for any $d, P \geq 1$, *if h is any extension of D which is either a P-term-d-DNF or a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$, then the error of h is less than*

$$\frac{1}{m}\left(dP\ln\left(\frac{en}{d}\right)+P\ln\left(\frac{2e}{P}\right)+\ln\left(\frac{4}{\delta}\right)+2\ln d+3\ln P\right),$$

for $n \geq 2$.

Note that if $P \geq 5$, then the second term in the bound is negative.

Theorem 2.2 *Suppose that D is a data set of m points, each generated at random according to a fixed probability distribution on $\{0,1\}^n$. Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$:* for any $d, P \geq 1$ *with $P \leq 2m$, if h is an extension of D which is a polynomial threshold function in $\mathcal{P}_P(n,d)$, then 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 2.1: We use a standard bound (which can be found in [6], for example): given a class of hypotheses H, for a random data set D 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 $|H|\exp(-\epsilon m)$. We observe that, in the non-standard convention, the class of P-term-d-DNF functions is a subclass of the class of binary-weight polynomial threshold functions $\mathcal{B}_P(n,d)$. We now bound the cardinality of this latter class. 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 $w_0 \in \{0,1,\ldots,P-1\}$ such that

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

The number of possible such x_S is

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

which is at most $(en/d)^d$. 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 P choices for w_0. Therefore

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

It follows that

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

So, fixing P, d and taking H equal either to the class of P-term-d-DNF or to $\mathcal{B}_P(n,d)$, with probability at least $1-\delta$, if $h \in H$ is an extension of a random data set D of size m, then

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

It follows that with probability at most $1-\delta/(4d^2P^2)$, there will 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

$$\epsilon(d,P,n,m) = \frac{1}{m}\left(dP\ln(en/d) + P\ln(2e/P) + \ln\left(\frac{4d^2P^3}{\delta}\right)\right).$$

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 2.2: We use a bound from [6], which follows [11]. 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 cardinality of H restricted to any 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

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

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 the restriction of H to domain C. Then, for any subset C of $\{0,1\}^n$, of cardinality 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 $r \leq P$. We can use the theory of the *Vapnik-Chervonenkis dimension.* This was introduced in [12] and has been used extensively in computational learning theory. Given a set G of functions from a (not necessarily finite) set X to $\{0,1\}$, the *VC-dimension* of G, $\operatorname{VCdim}(G)$, is defined to be the largest integer D such that for some set C of cardinality k, $|G|_C| = 2^k$. From Sauer's inequality [10], if $k \geq D \geq 1$, $\Pi_G(k) \leq (ek/D)^D$. 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,

$$\Pi_{H^{\mathcal{R}}}(k) \leq \left(\frac{ek}{P}\right)^P.$$

Hence,

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

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

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

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 $\delta/(4d^2P^2)$. As above, the result then follows. □

3 Bounds involving observed error

We now develop some more general results. In particular, we bound the error in terms of the observed error for non-extensions. 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 3.1 *Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$:* for any $d, P \geq 1$, *if h is any P-term-d-DNF or a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$, then*

$$\mathrm{er}(h) < \mathrm{er}_D(h) + \sqrt{\frac{1}{2m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + 2\ln(dP) + \ln\left(\frac{8P}{\delta}\right)\right)}.$$

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 $H = \mathcal{B}_P(n,d)$,

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

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

$$\epsilon = \sqrt{\frac{1}{2m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + 2\ln(dP) + \ln\left(\frac{8P}{\delta}\right)\right)}.$$

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

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

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

Proof: We use the following result of Vapnik and Chervonenkis [12, 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 3.3 *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. Let δ be any positive number less than one. Then the following holds with probability at least $1 - \delta$: for all $h \in H$,*

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

where Π_H is the growth function of H.

Proof: A theorem of Vapnik [11] shows that, for any η, with probability at least $1 - 4\,\Pi_H(2m)\,e^{-m\eta^2/4}$, for all $h \in H$,

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

It follows, therefore, that with probability at least $1 - \delta$, for all $h \in H$,

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

where

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

This means

$$\mathrm{er}(h) - \alpha\sqrt{\mathrm{er}(h)} - \mathrm{er}_D(h) < 0.$$

Thinking of this as a quadratic inequality in the nonnegative quantity $\sqrt{\mathrm{er}(h)}$, we therefore must have

$$\sqrt{\mathrm{er}(h)} < \frac{\alpha}{2} + \frac{\sqrt{\alpha^2 + 4\,\mathrm{er}_D(h)}}{2},$$

and so

$$\begin{aligned}
\mathrm{er}(h) &< \left(\frac{\alpha}{2} + \frac{\sqrt{\alpha^2 + 4\,\mathrm{er}_D(h)}}{2}\right)^2 \\
&= \frac{\alpha^2}{4} + \frac{1}{4}\left(\alpha^2 + 4\,\mathrm{er}_D(h)\right) + \frac{\alpha}{2}\sqrt{\alpha^2 + 4\,\mathrm{er}_D(h)} \\
&\leq \frac{\alpha^2}{2} + \mathrm{er}_D(h) + \frac{1}{2}\left(\alpha^2 + 4\,\mathrm{er}_D(h)\right) \\
&= \alpha^2 + 3\,\mathrm{er}_D(h),
\end{aligned}$$

as required. □

We then have the following bounds.

Theorem 3.4 *Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Let δ be a positive number less than one. Then the following holds with probability at least $1 - \delta$:* for any $d, P \geq 1$, *if h is any P-term-d-DNF or a binary-weight polynomial threshold function in $\mathcal{B}_P(n,d)$, then*

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

Proof: We first note that $\Pi_H(2m) \leq |H|$ and then observe that, by Theorem 3.3, and using our earlier bound for the cardinality of H, 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

$$\epsilon = \frac{4}{m}\left(dP\ln\left(\frac{en}{d}\right) + P\ln\left(\frac{2e}{P}\right) + \ln\left(\frac{16P^3d^2}{\delta}\right)\right).$$

□

Theorem 3.5 *Suppose that D is a data set of m labeled points, each generated at random according to a fixed probability distribution on $Z = \{0,1\}^n \times \{0,1\}$. Let δ be a positive number less than one. Then the following holds with probability at least $1-\delta$:* for any $d, P \geq 1$ *with $P \leq 2m$, if h is a polynomial threshold function in $\mathcal{P}_P(n,d)$, then*

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

Proof: We observe that, by Theorem 3.3, 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

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

□

4 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 discriminant

$$f = \sum_{i=1}^{q} T_i - \sum_{j=1}^{r} T'_j$$

or, more generally, the discriminant obtained when weighting the patterns:

$$f = \sum_{i=1}^{q} w_i T_i - \sum_{j=1}^{r} w'_j T'_j.$$

Suppose, then, that $h = \text{sgn}(f)$ where $f = \sum_{i=1}^{q} w_i T_i - \sum_{j=1}^{r} w'_j T'_j$. For $\gamma > 0$, we define the error of h on D at margin γ to be

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

So, this is the proportion of data points in D for which either $h(x) = \text{sgn}(f(x)) \neq y$, 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, $\text{er}_D^{\gamma}(h) \geq \text{er}_D(h)$.

We can bound the generalization error of homogeneous polynomial threshold classifiers in terms of their margin error. However, it is possibly more useful to obtain a different type of error bound which doesn't 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 $h = \text{sgn}(f)$, 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. For this reason, $\hat{\phi}_D^{\gamma}(h) \leq \text{er}_D^{\gamma}(h)$. For, in the case where $0 < y_i f(x_i) < \gamma$, we obtain a contribution of $1/m$ to $\text{er}^{\gamma}(h)$ 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

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

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

We have the following result, obtained using results from [8, 5, 9]. This bound is better than the comparable bound, that of Theorem 3.2, if we can take γ to be larger than of order $\sqrt{P}$, while having $\hat{\phi}_D^{\gamma}(h)$ close to $\text{er}_D(h)$, as will be the case, for instance, if we are using an unweighted discriminant and most observations are covered by many of the patterns.

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

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

where

$$\epsilon'(m,d,P,n,\gamma) = \frac{4P}{\gamma}\sqrt{\frac{2d}{m}\ln\left(\frac{2en}{d}\right)} + \sqrt{\frac{1}{2m}\left(\ln\left(\frac{8}{\delta}\right) + 2\ln\log_2\left(\frac{4P}{\gamma}\right) + 2\ln(dP)\right)}.$$

Proof: Let H be the set of all homogeneous polynomial threshold functions. Let $F_{d,P}$ denote the set of normalised discriminants involving at most P patterns, of degree at most d. Thus, it is the set of all functions of the form $f = \sum_{i=1}^{q} w_i T_i - \sum_{j=1}^{r} w'_j T'_j$ where $q+r = P$, each T_i and T'_j is of degree at most d, and $\sum_{i=1}^{q} |w_i| + \sum_{j=1}^{r} |w'_j| = P$. 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, d, P, and for any $\delta \in (0,1)$, the following holds with probability at least $1-\delta$: for all $h \in H$,

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

where $R_m(F_{d,P})$ is the *Rademacher complexity* of $F_{d,P}$. Consider, for $x \in \{0,1\}^n$, the vector $x^{(d)}$ whose entries are (in some prescribed order) x_S for all non-empty S of cardinality at most d. The set of all such $x^{(d)}$ forms a subset of $\{0,1\}^N$ where $N = \sum_{i=1}^{d}\binom{n}{i}$. We may consider the function

$$f(x) = \sum_{i=1}^{q} w_i T_i - \sum_{j=1}^{r} w'_j T'_j$$

as being of the form

$$f(x) = \sum_{1 \leq |S| \leq d} \alpha_S x_S,$$

where the α_S are $\pm w_i$ or $\pm w'_j$. Thus the set $F_{d,P}$ can be thought of as a (domain-restriction of) a subset of the set $\mathcal{G}$ of all linear functions defined on $\{0,1\}^N$ defined by weight vectors α with $\|\alpha\|_1 = P$ (this because of normalisation). 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 $R_m(F_{d,P}) \leq R_m(\mathcal{G})$. To bound $R_m(\mathcal{G})$ we use a result from [8]. This shows that

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

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

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

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 $\gamma \leq P$ since if $\gamma > P$, then $\hat{\phi}_D^\gamma(h) = 1$ (by the normalisation assumption) and the error bound is then trivially true. We obtain the following, for fixed d, P: with probability at least $1 - \delta$, for *all* $\gamma \in (0, P]$, if $h = \mathrm{sgn}(f)$ where $f \in F_{d,P}$ then

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

The theorem now follows by using the same sort of methods as before to move to a bound in which d, P are not prescribed in advance: we simply replace δ by $\delta/(4d^2P^2)$. □

Acknowledgements

Thanks to Endre Boros and John Shawe-Taylor for helpful discussions.

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

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

[12] 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.