Margin-based Generalization Error Bounds for Threshold Decision Lists

Martin Anthony
Department of Mathematics
and Centre for Discrete and Applicable Mathematics
London School of Economics
London WC2A 2AE, UK.
m.anthony@lse.ac.uk

May 2003
CDAM Research Report LSE-CDAM-2003-09

Abstract

This paper concerns the use of *threshold decision lists* for classifying data into two classes. The use of such methods has a natural geometrical interpretation and can be appropriate for an iterative approach to data classification, in which some points of the data set are given a particular classification, according to a linear threshold function (or hyperplane), are then removed from consideration, and the procedure iterated until all points are classified. We analyse theoretically the generalization properties of data classification techniques that are based on the use of threshold decision lists and the subclass of *multilevel threshold functions*. We obtain bounds on the generalization error that depend on the levels of separation — or *margins* — achieved by the successive linear classifiers.

1 Introduction

This paper concerns the use of *threshold decision lists* for classifying data into two classes. The use of such methods has a natural geometrical interpretation and can be appropriate for an iterative approach to data classification, in which some points of the data set are given a particular classification, according to a linear threshold function (or hyperplane), are then removed from consideration, and the procedure iterated until all points are classified. We analyse theoretically the generalization properties of data classification techniques that are based on the use of threshold decision lists and the subclass of *multilevel threshold functions*. We obtain bounds on the generalization error that depend on the levels of separation — or *margins* — achieved by the successive linear classifiers. The results we obtain are easily modified to give generalization bounds for *perceptron decision trees* [17, 6], improving upon previous such results.

1.1 Threshold decision lists

Suppose that F is any set of functions from $\mathbb{R}^n$ to $\{0,1\}$, for some fixed $n \in \mathbb{N}$. A function $f : \mathbb{R}^n \to \{0,1\}$ is a *decision list* based on F if it can be evaluated as follows, for some $k \in \mathbb{N}$, some functions $f_1, f_2, \ldots, f_k \in F$, some $c_1, c_2, \ldots, c_k \in \{0,1\}$, and all $y \in \mathbb{R}^n$: if $f_1(y) = 1$, then $f(y) = c_1$; if not, we evaluate $f_2(y)$, and if $f_2(y) = 1$, then $f(y) = c_2$; otherwise we evaluate $f_3(y)$, and so on. If y fails to satisfy any f_i then $f(y)$ is given the default value 0. We can regard a decision list based on F as a finite sequence

$$f = (f_1, c_1),\ (f_2, c_2), \ldots,\ (f_r, c_r),$$

such that $f_i \in F$ and $c_i \in \{0,1\}$ for $1 \leq i \leq r$. The values of f are defined by $f(y) = c_j$ where $j = \min\{i \mid f_i(y) = 1\}$, or 0 if there are no j such that $f_j(y) = 1$. We call each f_j a *test*, and the pair (f_j, c_j) a *term* of the decision list. Decision lists were introduced by Rivest [15], in the context of learning Boolean functions (and where the tests were conjunctions of literals).

A function $t : \mathbb{R}^n \to \{0,1\}$ is a *threshold function* if there are $w \in \mathbb{R}^n$ and $\theta \in \mathbb{R}$ such that

$$t(x) = \begin{cases} 1 & \text{if } \langle w, x\rangle \geq \theta \\ 0 & \text{if } \langle w, x\rangle < \theta, \end{cases}$$

where $\langle w, x\rangle$ is the standard inner product of w and x. Thus, $t(x) = \mathrm{sgn}(\langle w, x\rangle - \theta)$, where $\mathrm{sgn}(z) = 1$ if $z \geq 0$ and $\mathrm{sgn}(z) = 0$ if $z < 0$. Given such w and θ, we say that t is represented by $[w, \theta]$ and we write $t \leftarrow [\alpha, \theta]$. The vector w is known as the *weight vector*, and θ is known

as the *threshold*. Geometrically, a threshold function is defined by a hyperplane: all points lying to one side of the plane and on the plane are given the value 1, and all points on the other side are given the value 0.

Threshold decision lists are decision lists in which the tests are threshold functions. (These have also been called *neural* decision lists [12] and *linear* decision lists [20].) Formally, a threshold decision list

$$f = (f_1, c_1), (f_2, c_2), \ldots, (f_r, c_r)$$

has each $f_i : \mathbb{R}^n \to \{0, 1\}$ of the form $f_i(x) = \mathrm{sgn}(\langle w_i, x \rangle - \theta_i)$ for some $w_i \in \mathbb{R}^n$ and $\theta_i \in \mathbb{R}$. The value of f on $y \in \mathbb{R}^n$ is $f(y) = c_j$ if $j = \min\{i \mid f_i(y) = 1\}$ exists, or 0 otherwise (that is, if there are no j such that $f_j(y) = 1$).

There is a natural geometrical interpretation of the use of threshold decision lists. Suppose we are given some data points in $\mathbb{R}^n$, each one of which is labeled 0 or 1. Of course, since there are very few threshold functions, it is unlikely that the positive and negative points can be separated by a hyperplane. But we can use a hyperplane to separate off a set of points all having the same classification (either all are positive points or all are negative points). These points can then be removed from consideration and the procedure iterated until no points remain. This procedure is similar in nature to one of Jeroslow [11], but at each stage in his procedure, only positive examples may be 'chopped off' (not positive *or* negative). The classifier constructed by this iterative procedure is a threshold decision list.

If we consider threshold decision lists in which the hyperplanes are parallel, we obtain a special subclass, known as the *multilevel threshold functions*. These have been considered in a number of papers, such as [8, 13, 19], for instance. A *k-level threshold function* f is one that is representable by a threshold decision list of length k in which the test hyperplanes are parallel to each other. Any such function is defined by k parallel hyperplanes, which divide $\mathbb{R}^n$ into $k + 1$ regions. The function assigns points in the same region the same value, either 0 or 1. Without any loss, we may suppose that the classifications assigned to points in neighbouring regions are different (for, otherwise, at least one of the planes is redundant); thus, the classifications alternate as we traverse the regions in the direction of the normal vector common to the hyperplanes.

2 Generalization error and covering numbers

Following a form of the PAC model of computational learning theory (see [4, 21, 7]), we assume that labelled data points (x, b) (where $x \in \mathbb{R}^n$ and $b \in \{0, 1\}$) have been generated randomly (perhaps from some larger corpus of data) according to a fixed probability distribution P on $Z = \mathbb{R}^n \times \{0, 1\}$. (Note that this includes as a special case the situation in which x is drawn according to a fixed distribution μ on $\mathbb{R}^n$ and the label b is then given by $b = t(x)$ where t is some fixed function.) Thus, if there are m data points, we may regard the data set as a *sample* $s = ((x_1, b_1), \ldots, (x_m, b_m)) \in Z^m$, drawn randomly according to the product probability distribution P^m. Suppose that H is the set of threshold decision lists of some fixed length, k. Given any function $f \in H$, we can measure how well f matches the sample s through its *sample error*

$$\mathrm{er}_s(f) = \frac{1}{m}|\{i : f(x_i) \neq b_i\}|$$

(the proportion of points in the sample incorrectly classified by f). An appropriate measure of how well f would perform on further examples is its *error*,

$$\mathrm{er}_P(f) = P\left(\{(x, b) \in Z : f(x) \neq b\}\right),$$

the probability that a further randomly drawn labelled data point would be incorrectly classified by f.

Much effort has gone into obtaining high-probability bounds on $\mathrm{er}_P(f)$ in terms of the sample error. A typical result would state that, for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, for all $h \in H$, $\mathrm{er}_P(h) < \mathrm{er}_s(h) + \epsilon(m, \delta)$, where $\epsilon(m, \delta)$ (known as a *generalization error bound*) is decreasing in m and δ. Such results can be derived using uniform convergence theorems from probability theory [22, 14, 10], in which case $\epsilon(m, \delta)$ would typically involve the VC-dimension; see [22, 7, 21, 2].

Recently, some emphasis has been placed in practical machine learning techniques (such as Support Vector Machines; see [9], for instance) on 'learning with a large margin'; see [18, 2, 3, 16], for example. This paper obtains results of this type for the class of threshold decision lists. Precise formulations will be given shortly, but, broadly speaking, the rationale behind margin-based generalization error bounds is that if a classifier has managed to achieve a 'wide' separation between (most of) the points of different classification, then this indicates that it is a good classifier, and it is possible that a better (that is, smaller) generalization error bound can be obtained. The classical example of this is linear separation, where the classifier is a linear threshold function. If we have found a linear threshold function that classifies the points of a sample correctly *and*, moreover, the points of opposite classifications are separated by

a wide margin (so that the hyperplane achieves not just a correct, but a 'definitely' correct classification), then this function might be a better classifier of future, unseen, points than one which 'merely' separates the points correctly, but with a small margin.

A key tool in the derivation of margin-based generalization error bounds is the *covering number* of a class of real functions. Suppose that $F : X \to \mathbb{R}$ is a set of real-valued functions with domain X, and that $x = (x_1, x_2, \ldots, x_m)$ is an unlabelled sample of m points of X. Then, for $\epsilon > 0$, $C \subseteq F$ is an *ϵ-cover of F with respect to the d_∞^x-metric* if for all $f \in F$ there is $\hat{f} \in C$ such that $d_\infty^x(f, \hat{f}) < \epsilon$, where $d_\infty^x(f, g) = \max_{1 \le i \le m} |f(x_i) - g(x_i)|$. (Coverings with respect to other metrics derived from x can also be defined, but this paper needs only the present definition.) The class F is said to be totally bounded if it has a finite ϵ-cover with respect to the d_∞^x metric, for all $\epsilon > 0$ and all $x \in X^m$ (for all m). In this case, given $x \in X^m$, we define the *d_∞^x-covering numbers* $\mathcal{N}_\infty(F, \epsilon, x)$ to be the minimum cardinality of an ϵ-cover of F with respect to the d_∞^x-metric. We then define the *d_∞-covering numbers* $\mathcal{N}_\infty(F, \epsilon, m)$ by

$$\mathcal{N}_\infty(F, \epsilon, m) = \sup\{\mathcal{N}_\infty(F, \epsilon, x) : x \in X^m\}.$$

Many bounds on covering numbers for specific classes have been obtained (see [2] for an overview), and general bounds on covering numbers in terms of a generalization of the VC-dimension, known as the *fat-shattering dimension*, have been given [1].

In this paper, we use a recent bound of Zhang [23] for the d_∞-covering numbers of bounded linear mappings. For $R > 0$, let $B_R = \{x \in \mathbb{R}^n : \|x\| \le R\}$ be the closed ball in $\mathbb{R}^n$ of radius R, centred on the origin. For $w \in \mathbb{R}^n$, let $f_w : B_R \to \mathbb{R}$ be given by $f_w(x) = \langle w, x \rangle$, and let

$$L_R = \{f_w : w \in \mathbb{R}^n, \|w\| = 1\}.$$

Zhang [23] has shown that

$$\log_2 \mathcal{N}_\infty(L_R, \epsilon, m) \le 36 \frac{R^2}{\epsilon^2} \log_2 \left(2 \lceil 4R/\epsilon + 2 \rceil m + 1\right). \tag{1}$$

One thing of note is that this bound is dimension-independent: it does not depend on n. This bound differs from previous bounds [5, 2, 16] for the logarithm of the d_∞-covering numbers in that it involves a factor of order $\ln m$ rather than $(\ln m)^2$. (Previous approaches to bounding the d_∞-covering numbers first bounded the *fat-shattering dimension* and then used a result of Alon *et al.* [1] relating the covering numbers to the fat-shattering dimension. An additional $\ln m$ factor appears when this route is taken.)

3 Error bounds for threshold decision lists

Suppose that h is a threshold decision list, with k terms, and suppose that the tests in h are the threshold functions $t_1, t_2, \ldots, t_k$, and that t_i is represented by weight vector w_i and threshold θ_i. We say that h classifies the labelled example (x, b) (correctly, and) with margin $\gamma > 0$ if $h(x) = b$ and, for all $1 \leq i \leq k$, $|\langle w_i, x\rangle - \theta_i| \geq \gamma$. In other words, h classifies x with margin γ if, overall, the classification of x given by the threshold decision list h is correct and, additionally, x is distance at least γ from *all* of the k hyperplanes defining h. Note that we do not simply stipulate that x is distance at least γ from the single hyperplane involved in the first test that x passes: rather, we require x to be distance at least γ from all of the hyperplanes. (In this sense, the classification given to x by h is not just correct, but 'definitely' correct.) Given a labelled sample $s = ((x_1, b_1), \ldots, (x_m, b_m))$, the error of h on s at margin γ, denoted $\mathrm{er}_s^\gamma(h)$, is the proportion of labelled examples in s that are *not* classified by h with margin γ. Thus, $\mathrm{er}_s^\gamma(h)$ is the fraction of the sample points that are either misclassified by h, or are classified correctly but are distance less than γ from one of the planes.

Shawe-Taylor and Cristianini [17] and Bennett *et al.* [6] considered *perceptron decision trees*, decision trees in which the nodes compute threshold functions. Following their approach, rather than considering one margin parameter γ, we could have margin parameters $\gamma_1, \gamma_2, \ldots, \gamma_k$ for each of the k terms of the decision list. Given $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, we can modify the definition just given by saying that h classifies the labelled example (x, b) correctly with margin Γ if $h(x) = b$ and, for all $1 \leq i \leq k$, $|\langle w_i, x\rangle - \theta_i| \geq \gamma_i$. We define $\mathrm{er}_s^\Gamma(h)$ to be the proportion of labelled examples in the sample s that are not classified with margin Γ. Following a method used in [17, 6], together with the covering number bound from [23], we can obtain the following two results. (In these results, it simplifies matters to assume that $R \geq 1$ and $\gamma_i \leq 1$. But it will be clear how to modify them otherwise.)

Theorem 3.1 *Suppose $R \geq 1$ and $Z = B_R \times \{0, 1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Fix $k \in \mathbb{N}$ and let H be the set of all threshold decision lists with k terms, defined on domain B_R. Let $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0, 1]$ be given. Then, with probability at least $1 - \delta$, the following holds for $s \in Z^m$: if $h \in H$ and $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(576\, R^2 D(\Gamma) \log_2(8m) + \ln\left(\frac{2}{\delta}\right)\right)},$$

where $D(\Gamma) = \sum_{i=1}^k (1/\gamma_i^2)$.

Proof: The proof follows a technique from [17, 6] (where the case of zero margin error was the

focus), and is a modification of proofs in [2, 3, 5, 16], which in turn are based on [22]. For those not familiar with these proofs, and for the sake of completeness, some details are included that the cognoscenti will recognise as standard.

Given $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$, it can fairly easily be shown that if

$$Q = \{s \in Z^m : \exists h \in H \text{ with } \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon\}$$

and

$$R = \{(s, s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon/2\},$$

then $P^m(Q) \leq 2\, P^{2m}(R)$. Let G be the permutation group (the 'swapping group') on the set $\{1, 2, \ldots, 2m\}$ generated by the transpositions $(i, m+i)$ for $i = 1, 2, \ldots, m$. Then G acts on Z^{2m} by permuting the coordinates: for $\sigma \in G$, $\sigma(z_1, z_2, \ldots, z_{2m}) = (z_{\sigma(1)}, \ldots, z_{\sigma(m)})$. Now, by invariance of P^{2m} under the action of G, $P^{2m}(R) \leq \max\{\Pr(\sigma z \in R) : z \in Z^{2m}\}$, where $\Pr$ denotes the probability over uniform choice of σ from G. (See [22, 2], for instance.) Given a threshold decision list on $\mathbb{R}^n$, each test is of the form $f_i \leftarrow [w_i, \theta_i]$; that is, the test is passed if and only if $\langle w_i, x\rangle \geq \theta_i$. An equivalent functionality is obtained by using inputs in $\mathbb{R}^n$ augmented by -1, and using *homogeneous* threshold functions of $n+1$ variables; that is, ones with zero threshold. So any threshold decision list of length k on $\mathbb{R}^n$ can be realised as one on $\mathbb{R}^{n+1}$, defined on the subset $\mathbb{R}^n \times \{-1\}$, and with homogeneous threshold functions as its tests. Fix $z \in Z^{2m}$ and let $x = (x_1, x_2, \ldots, x_{2m}) \in X^{2m}$ be the corresponding vector of x_i, where $z_i = (x_i, b_i)$. For i between 1 and k, let C_i be a $\gamma_i/2$-cover of L with respect to the d_∞^x metric, where L is the set of linear functions $x \mapsto \langle w, x\rangle$ for $\|w\| = 1$, defined on the domain $D = \{(x, -1) : x \in \mathbb{R}^n, \|x\| \leq R\}$. Note that if $x \in \mathbb{R}^n$ satisfies $\|x\| \leq R$, then the corresponding $(x, -1)$ has length at most $\sqrt{R^2+1}$. So, by the covering number bound (1),

$$\log_2 |C_i| \leq \frac{144(R^2+1)}{\gamma_i^2} \log_2 \left(\left(\frac{32\sqrt{R^2+1}}{\gamma_i} + 14 \right) m \right) \leq \frac{288R^2}{\gamma_i^2} \log_2 \left(\frac{60Rm}{\gamma_i} \right). \qquad (2)$$

Suppose that h is a threshold decision list with k homogeneous threshold tests, defined on D. Denote the tests of the list by $f_1, f_2, \ldots, f_k$, where f_i corresponds to weight vector $w_i \in \mathbb{R}^{n+1}$. For each i, let $\hat{f}_i \in C_i$ satisfy $d_\infty^x(f_i, \hat{f}_i) < \gamma_i/2$, let $\hat{w}_i$ be the corresponding weight vector, and let $\hat{h}$ be the threshold decision list obtained from h by replacing each f_i by $\hat{f}_i$. The set $\hat{H}$ of possible such $\hat{h}$ is of cardinality $\prod_{i=1}^k |C_i|$. Suppose that $\sigma z = (s, s') \in R$ and that $\mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon/2$. Let $\Gamma/2 = (\gamma_1/2, \ldots, \gamma_k/2)$. Then, because for all $1 \leq j \leq 2m$ and all $1 \leq i \leq k$, $|\langle w_i, x_j\rangle - \langle \hat{w}_i, x_j\rangle| < \gamma_i/2$, it can be seen that $\mathrm{er}_s^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s(h)$ and $\mathrm{er}_{s'}^\Gamma(h) \geq \mathrm{er}_{s'}^{\Gamma/2}(\hat{h})$. Explicitly (denoting any given x_i by x), $\mathrm{er}_s^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s(h)$ follows from the observation that if $\langle w_i, x\rangle < 0$, then $\langle \hat{w}_i, x\rangle < \gamma_i/2$ and if $\langle w_i, x\rangle > 0$, then $\langle \hat{w}_i, x\rangle > -\gamma_i/2$; and $\mathrm{er}_{s'}^\Gamma(h) \geq \mathrm{er}_{s'}^{\Gamma/2}(\hat{h})$ follows from the facts that if $\langle \hat{w}_i, x\rangle < \gamma_i/2$ then $\langle w_i, x\rangle < \gamma_i$, and if

$\langle \hat{w}_i, x\rangle > -\gamma_i/2$ then $\langle w_i, x\rangle > -\gamma_i$. So, $\mathrm{er}_s^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s^{\Gamma/2}(\hat{h}) + \epsilon/2$, and therefore, for any $z \in Z^{2m}$,

$$\Pr\left(\sigma z \in R\right) \leq \Pr\left(\sigma z \in \bigcup_{\hat{h}\in\hat{H}} R(\hat{h})\right),$$

where

$$R(\hat{h}) = \{(s, s') \in Z^{2m} : \mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s^{\Gamma/2}(\hat{h}) + \epsilon/2\}.$$

Fix $\hat{h} \in \hat{H}$ and let $w_i = 0$ if $\hat{h}$ classifies z_i with margin at least $\Gamma/2$, and 1 otherwise. Then

$$\Pr\left(\sigma z \in R(\hat{h})\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^{m}(w_{m+i} - w_i) \geq \epsilon/2\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i|w_i - w_{m+i}| \geq \epsilon/2\right),$$

where the ε_i are independent (Rademacher) $\{-1, 1\}$ random variables, each taking value 1 with probability $1/2$, and where the last probability is over the joint distribution of the ε_i. Hoeffding's inequality bounds this probability by $\exp(-\epsilon^2 m/8)$. (See [2], for instance, for details.)

We therefore have

$$\Pr\left(\sigma z \in R\right) \leq \Pr\left(\sigma z \in \bigcup_{\hat{h}\in\hat{H}} R(\hat{h})\right) \leq |\hat{H}|\exp(-\epsilon^2 m/8),$$

which gives

$$P^m(Q) \leq 2\, P^{2m}(R) \leq 2\prod_{i=1}^{k}|C_i|\exp(-\epsilon^2 m/8).$$

Using the bound (2), we see that, provided

$$\epsilon \geq \epsilon_0 = \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k}\frac{288R^2}{\gamma_i^2}\log_2\left(\frac{60Rm}{\gamma_i}\right) + \ln\left(\frac{2}{\delta}\right)\right)},$$

then the probability of Q is at most δ. So, with probability at most $1-\delta$, for all $h \in H$, $\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \epsilon_0$. If, for each i, $m \geq R^2/\gamma_i^2$, then $\log_2(60Rm/\gamma_i) \leq 2\log_2(8m)$ and so, with probability at least $1-\delta$, for all $h \in H$,

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k}\frac{576R^2}{\gamma_i^2}\log_2(8m) + \ln\left(\frac{2}{\delta}\right)\right)}. \tag{3}$$

If, however, for some i, $m < R^2/\gamma_i^2$, then the bound (3) is trivially true (since the term under the square root is greater than 1). The result follows. □

Theorem 3.2 *Suppose $R \geq 1$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Fix $k \in \mathbb{N}$ and let H be the set of all threshold decision lists with k terms, defined on domain B_R. Let $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0,1]$ be given. Then, with probability at least $1-\delta$, the following holds for $s \in Z^m$: if h is any threshold decision list with k terms, and h classifies s with margin $\Gamma = (\gamma_1, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(576\, R^2 D(\Gamma) \log_2(8m) + \log_2\left(\frac{2}{\delta}\right)\right)$$

where $D(\Gamma) = \sum_{i=1}^{k}(1/\gamma_i^2)$.

Proof: This proof is similar to that of Theorem 3.1. It uses, first, the fact (see [22, 2]) that if

$$Q = \{s \in Z^m : \exists h \in H \text{ with } \mathrm{er}_s^{\gamma}(h) = 0,\ \mathrm{er}_P(h) \geq \epsilon\}$$

and

$$R = \{(s, s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_s^{\gamma}(h) = 0,\ \mathrm{er}_{s'}(h) \geq \epsilon/2\},$$

then $P^m(Q) \leq 2\, P^{2m}(R)$. As before, $P^{2m}(R) \leq \max_{z \in Z^{2m}} \Pr(\sigma z \in R)$, where $\Pr$ denotes the probability over uniform choice of σ from the 'swapping group' G. It can be shown that for any $z \in Z^{2m}$,

$$\Pr\left(\sigma z \in R\right) \leq \Pr\left(\sigma z \in \bigcup_{\hat{h} \in \hat{H}} R(\hat{h})\right),$$

where

$$R(\hat{h}) = \{(s, s') \in Z^{2m} : \exists h \in H \text{ with } \mathrm{er}_s^{\gamma}(h) = 0,\ \mathrm{er}_{s'}(h) \geq \epsilon/2\}.$$

It can then be seen that, for each fixed $\hat{h} \in \hat{H}$,

$$\Pr\left(\sigma z \in R(\hat{h})\right) \leq \frac{2^{m(1-\epsilon/2)}}{|\Gamma|} = 2^{-\epsilon m/2}.$$

The result then proceeds as does the proof of Theorem 3.1, using the bound (2). □

One difficulty with Theorems 3.1 and 3.2 is that the number, k, of terms, and the margins γ_i are specified *a priori*. A more useful generalization error bound would enable us to choose or tune (or observe) these parameters after learning. We now derive such a result. The approach we take to obtaining a result of this type differs from that taken in [17, 6], and gives a slightly better bound.

We first need a generalization of a result from [5], where the following is shown. Suppose $\mathbb{P}$ is any probability measure and that $\{E(\alpha_1, \alpha_2, \delta) : 0 < \alpha_1, \alpha_2, \delta \leq 1\}$ is a set of events such that:

- for all α, $\mathbb{P}(E(\alpha,\alpha,\delta)) \leq \delta$,
- if $0 < \alpha_1 \leq \alpha \leq \alpha_2 < 1$ and $0 < \delta_1 \leq \delta \leq 1$, then $E(\alpha_1,\alpha_2,\delta_1) \subseteq E(\alpha,\alpha,\delta)$.

Then

$$\mathbb{P}\left(\bigcup_{\alpha\in(0,1]} E(\alpha/2,\alpha,\delta\alpha/2)\right) \leq \delta$$

for $0 < \delta < 1$. We extend this result as follows.

Theorem 3.3 *Suppose $\mathbb{P}$ is any probability measure, $k \in \mathbb{N}$, and that*

$$\{E(\Gamma_1,\Gamma_2,\delta) : \Gamma_1,\Gamma_2 \in (0,1]^k, \delta \leq 1\}$$

is a set of events such that:

- *for all $\Gamma \in (0,1]^k$, $\mathbb{P}(E(\Gamma,\Gamma,\delta)) \leq \delta$,*
- *$\Gamma_1 \leq \Gamma \leq \Gamma_2$ (component-wise) and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\Gamma_1,\Gamma_2,\delta_1) \subseteq E(\Gamma,\Gamma,\delta)$.*

Then

$$\mathbb{P}\left(\bigcup_{\Gamma\in(0,1]^k} E\left((1/2)\Gamma,\Gamma,\delta\prod_{i=1}^{k}\frac{\gamma_i}{2^k}\right)\right) \leq \delta$$

for $0 < \delta < 1$.

Proof: To prove this, we note that

$$\begin{aligned}
&\mathbb{P}\left(\bigcup_{\Gamma\in(0,1]^k} E\left((1/2)\Gamma,\Gamma,\delta\prod_{i=1}^{k}\frac{\gamma_i}{2^k}\right)\right)\\
\leq\ &\mathbb{P}\left(\bigcup_{i_1,i_2,\ldots,i_k=0}^{\infty}\left\{E\left((1/2)\Gamma,\Gamma,\delta\prod_{i=1}^{k}\frac{\gamma_i}{2^k}\right) : \text{for } j=1,\ldots,k, \gamma_j \in \left(\left(\frac{1}{2}\right)^{i_j+1},\left(\frac{1}{2}\right)^{i_j}\right]\right\}\right)\\
\leq\ &\mathbb{P}\left(\bigcup_{i_1,i_2,\ldots,i_k=0}^{\infty} E\left(\left(\left(\frac{1}{2}\right)^{i_1+1},\ldots,\left(\frac{1}{2}\right)^{i_k+1}\right),\left(\left(\frac{1}{2}\right)^{i_1+1},\ldots,\left(\frac{1}{2}\right)^{i_k+1}\right),\delta\prod_{j=1}^{k}\left(\frac{1}{2}\right)^{i_j}\frac{1}{2^k}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \sum_{i_1,i_2,\ldots,i_k=0}^{\infty} \delta \prod_{j=1}^{k} \left(\frac{1}{2}\right)^{i_j+1} \\
&= \delta \prod_{j=1}^{k} \sum_{i_j=0}^{\infty} \left(\frac{1}{2}\right)^{i_j+1} \\
&= \delta \prod_{j=1}^{k} 1 = \delta.
\end{aligned}
$$

□

We now have the following result.

Theorem 3.4 *Suppose $R \geq 1$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Let H be the set of all threshold decision lists (with any number of terms) defined on domain B_R. With probability at least $1-\delta$, the following statements hold for $s \in Z^m$:*

1. for all $k \in \mathbb{N}$ *and* for all $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0,1]$, *if $h \in H$ has k terms, and* $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, *then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304\, R^2 D(\Gamma) \log_2(8m) + \ln\left(\frac{2}{\delta}\right) + 2k \ln 2 + \sum_{i=1}^{k} \ln\left(\frac{1}{\gamma_i}\right)\right)},$$

   where $D(\Gamma) = \sum_{i=1}^{k} (1/\gamma_i^2)$.

2. for all $k \in \mathbb{N}$, *and* for all $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0,1]$, *if $h \in H$ has k terms, and h classifies s with margin $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(2304\, R^2 D(\Gamma) \log_2(8m) + \log_2\left(\frac{2}{\delta}\right) + 2k + \sum_{i=1}^{k} \log_2\left(\frac{1}{\gamma_i}\right)\right),$$

   where $D(\Gamma) = \sum_{i=1}^{k} (1/\gamma_i^2)$.

Proof: Now, to prove the Theorem, fix $k \in \mathbb{N}$. If $\Gamma_1 = (\gamma_1^{(1)}, \ldots, \gamma_k^{(1)})$ and $\Gamma_2 = (\gamma_1^{(2)}, \ldots, \gamma_k^{(2)})$, let $E(\Gamma_1, \Gamma_2, \delta)$ be the event that there exists a threshold decision list h with k terms such that

$$\mathrm{er}_P(h) \geq \mathrm{er}_s^{\Gamma_2}(h) + \sqrt{\frac{8}{m}\left(576\, R^2 D(\Gamma_1) \log_2(8m) + \ln\left(\frac{2}{\delta}\right)\right)},$$

where $D(\Gamma_1) = \sum_{i=1}^{k}(1/\gamma_i^{(1)})^2$. Then, by Theorem 3.1, $P^m(E(\Gamma,\Gamma,\delta)) \leq \delta$, and it is easily seen that $\Gamma_1 \leq \Gamma \leq \Gamma_2$ and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\Gamma_1,\Gamma_2,\delta_1) \subseteq E(\Gamma,\Gamma,\delta)$. It follows that

$$P^m\left(\bigcup_{\Gamma\in(0,1]^k} E\left((1/2)\Gamma,\Gamma,\delta\prod_{i=1}^{k}\frac{\gamma_i}{2^k}\right)\right) \leq \delta.$$

So, with probability at least $1-\delta$, *for all* $\gamma_1,\gamma_2,\ldots,\gamma_k \in (0,1]$, if h is any threshold decision list with k terms, and $\Gamma = (\gamma_1,\gamma_2,\ldots,\gamma_k)$, then

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304\,R^2D(\Gamma)\log_2(8m) + \ln\left(\frac{2}{\delta}\right) + k\ln 2 + \sum_{i=1}^{k}\ln\left(\frac{1}{\gamma_i}\right)\right)},$$

where $D(\Gamma) = \sum_{i=1}^{k}(1/\gamma_i^2)$. This holds for any *fixed* k. Replacing δ by $\delta/2^k$, we see that, with probability at least $1-\delta/2^k$, for any h with k terms and any Γ,

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304\,R^2D(\Gamma)\log_2(8m) + \ln\left(\frac{2\,2^k}{\delta}\right) + k\ln 2 + \sum_{i=1}^{k}\ln\left(\frac{1}{\gamma_i}\right)\right)},$$

and so, with probability at least $1-\sum_{k=1}^{\infty}(\delta/2^k) = 1-\delta$, for all k, for all h of length k, and for all Γ,

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304R^2D(\Gamma)\log_2(8m) + \ln\left(\frac{2}{\delta}\right) + 2k\ln 2 + \sum_{i=1}^{k}\ln\left(\frac{1}{\gamma_i}\right)\right)}.$$

(Note that we could have replaced δ by $\delta\alpha_k$ where (α_k) is any sequence such that $\sum_{i=1}^{\infty}\alpha_k = 1$.) The second part of the Theorem is proved similarly, using Theorem 3.2. □

Shawe-Taylor and Cristianini [17] and Bennett *et al.* [6] proved a margin-based generalization result for the more general class of perceptron decision trees, in the case where there is zero Γ-margin error on the sample. The special case of their result that applies to threshold decision lists gives a bound (with probability at least $1-\delta$) of the form

$$\mathrm{er}_P(h) < O\left(\frac{1}{m}\left(R^2D(\Gamma)(\ln m)^2 + k\ln m + \ln\left(\frac{1}{\delta}\right)\right)\right). \tag{4}$$

(The O-notation indicates that constants have been suppressed.)

By comparison, the bound given in Theorem 3.2 is of order

$$\mathrm{er}_P(h) < O\left(\frac{1}{m}\left(R^2 D(\Gamma)\ln m + k + \sum_{i=1}^{k}\ln\left(\frac{1}{\gamma_i}\right) + \ln\left(\frac{1}{\delta}\right)\right)\right). \tag{5}$$

The first term of bound (5) is a $\ln m$ factor better than the corresponding term of (4). That this is so is because we have used Zhang's covering number bound, (1), and have not bounded the covering number by using results on fat-shattering dimension, coupled with the bound of Alon *et al.* [1] that gives a general bound on covering numbers in terms of fat-shattering dimension. Additionally, since all these probability bounds are trivial (greater than 1) unless $m > (R/\gamma_i)^2$ for all i, the remaining terms of the bound (5) are of order no more than $O(k \ln m)$, and are potentially much smaller. This improvement results from our development and use of Theorem 3.3. Theorem 3.4 is therefore an improvement over the results implied by [17, 6].

It is a simple matter to modify the above results so that they apply to perceptron decision trees. As in [17, 6], one only additionally has to take into account the different number of decision tree architectures on a given number of nodes. In this way, the techniques here can be used to improve the results in [17, 6], shaving a $\ln m$ factor from the leading term of the error bound, and replacing $k \ln m$ by a term of order $k + \sum_{i=1}^{k} \ln(1/\gamma_i)$. The details are omitted here, but are easily worked out.

4 Error bounds for multilevel threshold functions

Suppose that h is a k-level threshold function, represented by weight vector w and threshold vector $\theta = (\theta_1, \theta_2, \ldots, \theta_k)$ (where $\theta_1 \leq \theta_2 \cdots \leq \theta_k$). Regarded as a threshold decision list, the tests are the threshold functions t_i, where $t_i(y) = \mathrm{sgn}(\langle w, x\rangle - \theta_i)$. Recall that we say h classifies the labelled example (x, b) with margin $\gamma > 0$ if $h(x) = b$ and, for all $1 \leq i \leq k$, $|\langle w, x\rangle - \theta_i| \geq \gamma$. (In other words, h classifies x correctly, and x is distance at least γ from any of the hyperplanes defining the multilevel threshold function h.) As above, for a labelled sample s, $\mathrm{er}_s^\gamma(h)$, the sample error at margin γ, is the proportion of labelled examples in s that are *not* correctly classified with margin γ.

To bound generalization error in this special case, we take a slightly different approach to the one used above for general threshold decision lists. Rather than take a cover for each term of the decision list, a more 'global' approach can be taken, exploiting the fact that the planes are parallel. In taking this approach, however, the analysis allows only one margin parameter, γ,

rather than k possibly different margin parameters, one for each plane. (As before, for the sake of simplicity, we assume that $R \geq 1$ and $\gamma \leq 1$.)

Theorem 4.1 *Suppose $R \geq 1$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Fix $k \in \mathbb{N}$ and let H be the set of all k-level threshold functions defined on domain B_R. Let P be any probability distribution on Z, and suppose $\gamma \in (0,1]$ and $\delta \in (0,1)$. Then, with P^m-probability at least $1-\delta$, a sample s is such that if $h \in H$, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\gamma(h) + \sqrt{\frac{8}{m}\left(\frac{1152R^2}{\gamma^2}\log_2(9m) + k\ln\left(\frac{10R}{\gamma}\right) + \ln\left(\frac{2}{\delta}\right)\right)}.$$

Proof: Fix $\gamma \in (0,1]$. As earlier, with H the set of k-level threshold functions on B_R, if

$$Q = \{s \in Z^m : \exists h \in H \text{ with } \mathrm{er}_P(h) \geq \mathrm{er}_s^\gamma(h) + \epsilon\}$$

and

$$R = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\gamma(h) + \epsilon/2\},$$

then $P^m(Q) \leq 2\,P^{2m}(R)$. Also as before, $P^{2m}(R) \leq \max\{\Pr(\sigma z \in R) : z \in Z^{2m}\}$, where $\Pr$ denotes the probability over uniform choice of σ from the 'swapping group' G. Let L_R be the set of all functions of the form $x \mapsto \langle w, x\rangle$, where $w \in \mathbb{R}^n$ satisfies $\|w\| = 1$, and where the domains of the functions are B_R. Now fix $z \in Z^{2m}$, let $x \in X^{2m}$ be the corresponding x_i-vector, and let C be a $\gamma/4$-cover of L with respect to the d_∞^x metric. By (1),

$$\begin{aligned}
\log_2|C| &\leq \log_2 \mathcal{N}_\infty(L_R, \gamma/4, 2m) \\
&\leq \frac{576R^2}{\gamma^2}\log_2\left(2\left\lceil 16R/\gamma + 2\right\rceil 2m + 1\right) \\
&\leq \frac{576R^2}{\gamma^2}\log_2\left(\frac{80Rm}{\gamma}\right).
\end{aligned}$$

Each function in C is represented by a weight vector, and we shall denote the set of these weight vectors by $\hat{W}$. For each $w \in \mathbb{R}^n$, denote by $\hat{w}$ a member of $\hat{W}$ such that for $i = 1, 2, \ldots, 2m$, $|\langle w, x_i\rangle - \langle \hat{w}, x_i\rangle| < \gamma/4$. Let

$$D = \{\theta \in \mathbb{R} : \exists n \in \mathbb{Z} \cap [-(4R/\gamma) - 1, (4R/\gamma) + 1] \text{ s.t. } \theta = n(\gamma/4)\},$$

and let $\hat{\Theta} = D^k$. Then

$$|\hat{\Theta}| \leq \left(\frac{8R}{\gamma} + 2\right)^k \leq \left(\frac{10R}{\gamma}\right)^k.$$

Now, suppose h is a k-level threshold function defined on B_R. Then, of course, h is represented by a weight vector $w \in \mathbb{R}^n$ and a threshold vector $\theta \in \mathbb{R}^k$. Without loss of generality, $\|w\| = 1$, in which case, since, for all $x \in B_R$, $|\langle w, x\rangle| \leq R$, we can assume that each θ_i satisfies $|\theta_i| \leq R$. Then, denote by $\hat{\theta}$ a member of $\hat{\Theta}$ such that for $i = 1, 2, \ldots, k$, $|\theta_i - \hat{\theta}_i| \leq \gamma/4$. (Such a $\hat{\theta}$ exists by the way in which $\hat{\Theta}$ is defined.) Let $\hat{H}$ be the set of all k-level threshold functions representable by weight vectors $\hat{w} \in \hat{W}$ and threshold vectors $\hat{\theta} = (\theta_1, \ldots, \theta_k) \in \hat{\Theta}$. Then

$$|\hat{H}| \leq 2^{(576R^2/\gamma^2)\log_2(80Rm/\gamma)} \left(\frac{10R}{\gamma}\right)^k.$$

For each $h \in H$, let $\hat{h}$ be the k-level threshold vector with weight vector $\hat{w} \in \hat{W}$ and threshold vector $\hat{\theta} \in \theta$, where $\hat{w}$ and $\hat{\theta}$ satisfy the properties indicated above. For each $i = 1, 2, \ldots, 2m$, for each $j = 1, 2, \ldots, k$,

$$|(\langle w, x_i\rangle - \theta_j) - (\langle \hat{w}, x_i\rangle - \hat{\theta}_j)| \leq |\langle w, x_i\rangle - \langle \hat{w}, x_i\rangle| + |\theta_i - \hat{\theta}_i| \leq \gamma/4 + \gamma/4 = \gamma/2.$$

This means that, when x is any one of the x_i, and $1 \leq j \leq k$,

$$\begin{aligned}
\langle w, x\rangle < \theta_j &\implies \langle \hat{w}, x\rangle < \hat{\theta}_j + \gamma/2, \\
\langle w, x\rangle > \theta_j &\implies \langle \hat{w}, x\rangle > \hat{\theta}_j - \gamma/2, \\
\langle \hat{w}, x\rangle \leq \hat{\theta}_j + \gamma/2 &\implies \langle w, x\rangle < \theta_j + \gamma, \\
\langle \hat{w}, x\rangle \geq \hat{\theta}_j - \gamma/2 &\implies \langle w, x\rangle > \theta_j - \gamma.
\end{aligned}$$

It follows that, if $\sigma z = (s, s') \in R$ and $\mathrm{er}_{s'}(h) \geq \mathrm{er}^{\gamma}_{s}(h) + \epsilon/2$, then $\mathrm{er}^{\gamma/2}_{s}(\hat{h}) \geq \mathrm{er}_s(h)$ and $\mathrm{er}^{\gamma}_{s'}(h) \geq \mathrm{er}^{\gamma/2}_{s'}(\hat{h})$, and so

$$\mathrm{er}^{\gamma/2}_{s'}(\hat{h}) \geq \mathrm{er}^{\gamma/2}_{s}(\hat{h}) + \epsilon/2.$$

The proof now proceeds as the proof of Theorem 3.1. For any $z \in Z^{2m}$,

$$\Pr(\sigma z \in R) \leq \Pr\left(\sigma z \in \bigcup_{\hat{h} \in \hat{H}} R(\hat{h})\right),$$

where

$$R(\hat{h}) = \{(s, s') \in Z^{2m} : \mathrm{er}^{\gamma/2}_{s'}(\hat{h}) \geq \mathrm{er}^{\gamma/2}_{s}(\hat{h}) + \epsilon/2\}.$$

Fixing $\hat{h} \in \hat{H}$, we find that

$$\Pr\left(\sigma z \in R(\hat{h})\right) \leq \exp(-\epsilon^2 m/8).$$

Therefore,

$$P^m(Q) < 2\,|\hat{H}|\,\exp(-\epsilon^2 m/8) \leq 2\,2^{576R^2/\gamma^2\log_2(80Rm/\gamma)}\,\left(\frac{10R}{\gamma}\right)^k\,\exp(-\epsilon^2 m/8).$$

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

$$\mathrm{er}_P(h) < \mathrm{er}_s(h) + \sqrt{\frac{8}{m}\left(\left(\frac{576R^2}{\gamma^2}\right)\log_2\left(\frac{80Rm}{\gamma}\right) + k\ln\left(\frac{10R}{\gamma}\right) + \ln\left(\frac{2}{\delta}\right)\right)}.$$

The result follows on noting that the bound stated in the Theorem is trivially true if $m < R^2/\gamma^2$, and is implied by the bound just derived if $m \geq R^2/\gamma^2$. □

Theorem 4.2 *Suppose $R > 0$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Fix $k \in \mathbb{N}$ and let H be the set of all k-level threshold functions defined on domain B_R. Let P be any probability distribution on Z, and suppose $\gamma \in (0,1]$ and $\delta \in (0,1)$. Then, with P^m-probability at least $1-\delta$, a sample s is such that if $h \in H$ and $\mathrm{er}_s^\gamma(h) = 0$, then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(\frac{1152R^2}{\gamma^2}\log_2(9m) + k\log_2\left(\frac{10R}{\gamma}\right) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Proof: This result is obatined by modifying the proof of Theorem 4.1, just in the same way as Theorem 3.2 is obtained by modifying the proof of Theorem 3.1. One can show that the probability that there exists $h \in H$ such that $\mathrm{er}_s^\gamma(h) = 0$ and $\mathrm{er}_P(h) \geq \epsilon$ is at most

$$2\,2^{(576R^2/\gamma^2)\log_2(80Rm/\gamma)}\,\left(\frac{10R}{\gamma}\right)^k\,2^{-\epsilon m/2},$$

from which the result follows. □

The generalization error bound implied by Theorem 3.1 in the case in which $\gamma_i = \gamma$ for all i is, suppressing constants,

$$\mathrm{er}_P(h) < \mathrm{er}_s^\gamma(h) + O\left(\sqrt{\frac{1}{m}\left(\frac{R^2k}{\gamma^2}\ln m + \ln\left(\frac{1}{\delta}\right)\right)}\right)$$

(with probability at least $1-\delta$), whereas that of Theorem 4.1 is

$$\mathrm{er}_P(h) < \mathrm{er}_s^\gamma(h) + O\left(\sqrt{\frac{1}{m}\left(\frac{R^2}{\gamma^2}\ln m + k\ln\left(\frac{R}{\gamma}\right) + \ln\left(\frac{1}{\delta}\right)\right)}\right),$$

so there is some advantage in the more particular analysis that has been carried out for multi-level threshold functions.

It is straightforward to remove the *a priori* specification of γ and k, using the result from [5] mentioned before Theorem 3.3. The following bounds are obtained.

Theorem 4.3 *Suppose $R > 0$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$. Fix $k \in \mathbb{N}$ and let H be the set of all multilevel threshold functions defined on domain B_R. Let P be any probability distribution on Z. Then, with P^m-probability at least $1-\delta$, the following hold:*

1. for all $k \in \mathbb{N}$ *and* for all $\gamma \in (0,1]$, *if $h \in H$ is a k-level threshold function, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\gamma(h) + \sqrt{\frac{8}{m}\left(\frac{4608R^2}{\gamma^2}\log_2(9m) + k\ln 2 + k\ln\left(\frac{20R}{\gamma}\right) + \ln\left(\frac{4}{\delta\gamma}\right)\right)}.$$

2. for all $k \in \mathbb{N}$, *and* for all $\gamma \in (0,1]$, *if $h \in H$ is a k-level threshold function and h classifies s with margin γ, then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(\frac{4608R^2}{\gamma^2}\log_2(9m) + k + k\log_2\left(\frac{20R}{\gamma}\right) + \log_2\left(\frac{4}{\delta\gamma}\right)\right).$$

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM* 44(4), 1997: 615–631.

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

[3] M. Anthony and P. L. Bartlett. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9, 2000: 213–225.

[4] M. Anthony and N. L. Biggs. *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science, 30, 1992. Cambridge University Press, Cambridge, UK.

[5] P. L. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* 44(2), 1998: 525–536.

[6] K. Bennett, N. Cristianini, J. Shawe-Taylor and D. Wu. Enlarging the Margins in Perceptron Decision Trees. *Machine Learning*, 41, 2000: 295–313.

[7] 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.

[8] V. Bohossian and J. Bruck. Multiple threshold neural logic. In *Advances in Neural Information Processing, Volume 10: NIPS'1997*, Michael Jordan, Michael Kearns, Sara Solla (eds), MIT Press, 1998.

[9] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[10] R. M. Dudley. *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 63, Cambridge University Press, Cambridge, UK, 1999.

[11] R.G. Jeroslow. On defining sets of vertices of the hypercube by linear inequalities. *Discrete Mathematics*, 11, 1975: 119–124.

[12] M. Marchand and M. Golea. On learning simple neural concepts: from halfspace intersections to neural decision lists. *Network: Computation in Neural Systems*, 4, 1993: 67–85.

[13] S. Olafsson and Y. S. Abu-Mostafa. The capacity of multilevel threshold functions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 10 (2), 1988: 277–281.

[14] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[15] R. L. Rivest. Learning Decision Lists. *Machine Learning* 2 (3), 1987: 229–246.

[16] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1996: 1926–1940.

[17] J. Shawe-Taylor and N. Cristianini. Data-Dependent Structural Risk Minimisation for Perceptron Decision Trees. Neurocolt Technical Report NC2-TR-1998-003, May 1998.

[18] A. J. Smola, P. L. Bartlett, B. Schölkopf and D. Schuurmans (editors). *Advances in Large-Margin Classifiers (Neural Information Processing)*, MIT Press 2000.

[19] R. Takiyama. The separating capacity of a multi-threshold element. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7, 1985: 112–116.

[20] G. Turán and F. Vatan. Linear decision lists and partitioning algorithms for the construction of neural networks. Foundations of Computational Mathematics: selected papers of a conference held at Rio de Janeiro, Springer 1997, pp 414-423

[21] V. N. Vapnik: *Statistical Learning Theory*, Wiley, 1998.

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

[23] T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2, 2002: 527–550.