Journal of Machine Learning Research 5 (2004) 189–217 Submitted 8/03; Published 2/04

Generalization Error Bounds for Threshold Decision Lists

Martin Anthony M.ANTHONY@LSE.AC.UK
Department of Mathematics
London School of Economics
London WC2A 2AE
United Kingdom

Editor: Yoram Singer

Abstract

In this paper we consider the generalization accuracy of classification methods based on the iterative use of linear classifiers. The resulting classifiers, which we call *threshold decision lists* act as follows. Some points of the data set to be classified are given a particular classification according to a linear threshold function (or hyperplane). These are then removed from consideration, and the procedure is iterated until all points are classified. Geometrically, we can imagine that at each stage, points of the same classification are successively chopped off from the data set by a hyperplane. We analyse theoretically the generalization properties of data classification techniques that are based on the use of threshold decision lists and on the special subclass of *multilevel threshold functions*. We present bounds on the generalization error in a standard probabilistic learning framework. The primary focus in this paper is on obtaining generalization error bounds that depend on the levels of separation—or *margins*—achieved by the successive linear classifiers. We also improve and extend previously published theoretical bounds on the generalization ability of perceptron decision trees.

Keywords: Threshold decision lists, generalization error, large margin bounds, growth function, covering numbers, perceptron decision trees

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 or sequential 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*. This analysis is carried out within the framework of the probabilistic PAC model of learning and its variants (see Valiant, 1984; Vapnik, 1998; Anthony and Biggs, 1992; Anthony and Bartlett, 1999; Blumer et al., 1989).

1.1 Outline of the Paper

Probabilistic approaches to the theory of machine learning can provide bounds on the 'generalization error' of classifiers. Such results give probabilistic guarantees on the future performance of a classifier trained on a large random training set. This paper takes three main approaches to bounding the generalization error of threshold decision lists.

First, in the 'classical' approach to the PAC model, we present results on generalization that are obtained through bounding the growth function of these classes.

Secondly, we obtain bounds on the generalization error of threshold decision lists that depend on the levels of separation—or *margins*—achieved by the successive linear classifiers. We use techniques inspired by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000), and also give generalization bounds for *perceptron decision trees*, improving upon and extending previous such results from Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000).

Thirdly, we focus specifically on the special subclass comprising the multilevel threshold functions. Here, a different and more specialized analysis results in generalization error bounds that are better than those that follow from the general results on threshold decision lists.

The rest of this section introduces the classes of threshold decision lists and multilevel threshold functions, and discusses related work. Section 2 discusses the definitions of generalization error. In Section 3, we derive bounds on the growth function and use these to bound the generalization error. Section 4 discusses the important idea of large-margin classification as it applies to threshold decision lists. Here, we give generalization error bounds that depend on the sizes of the margins and we also indicate some improved bounds for perceptron decision trees. Section 5 discusses the more specific margin-based analysis for multilevel threshold functions. Section 6 concludes the paper and suggests some possible directions for future work.

1.2 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 $j = \min\{i : f_i(y) = 1\}$, or 0 if there are no j such that $f_j(y) = 1$. We call each f_i a *test*, and the pair (f_i, c_i) a *term* of the decision list. Decision lists were introduced by Rivest (1987), 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, $t(x) = \text{sgn}(\langle w, x\rangle - \theta)$, where $\text{sgn}(z) = 1$ if $z \geq 0$ and $\text{sgn}(z) = 0$ if $z < 0$. Given such w and θ, we say that t is represented by $[w, \theta]$ and we write $t \leftarrow [w, \theta]$. 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 by Marchand and Golea (1993) and *linear* decision lists by

Turan and Vatan (1997). Formally, a threshold decision list

$$f = (t_1, c_1), (t_2, c_2), \ldots, (t_r, c_r)$$

has each $t_i : \mathbb{R}^n \to \{0,1\}$ of the form $t_i(x) = \mathrm{sgn}(\langle w_i, x\rangle - \theta_i)$ 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 $j = \min\{i : t_i(y) = 1\}$ exists, or 0 otherwise (that is, if there are no j such that $t_j(y) = 1$).

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. It is unlikely that the positive and negative points can be separated by a hyperplane. However, we could use a hyperplane to separate off a set of points all of 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 (1975), but at each stage in his procedure, only positive examples may be ‘chopped off’ (not positive *or* negative).

If we consider threshold decision lists in which the hyperplanes are parallel, we obtain a special subclass, known as the *multilevel threshold functions*. 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 neighboring 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.

1.3 Related Work

The chopping procedure described above suggests that the use of threshold decision lists is fairly natural, if an iterative approach is to be taken to pattern classification. Other iterative approaches—which proceed by classifying some points, removing these from consideration, and proceeding recursively—have been taken, using different types of base classifier. For example, Magasarian’s multisurface method (Mangasarian, 1968) finds, at each stage, two parallel hyperplanes (as close together as possible) such that the points not enclosed between the two planes all have the same classification. It then removes these points and repeats. This method may be regarded as constructing a decision list in which the set of base functions F are the indicator functions of the complements of regions enclosed between two parallel hyperplanes.

The focus of this paper is generalization error rather than learning algorithms. The ‘chopping procedure’ as we have described it is a useful device to help us see that threshold decision lists have a fairly natural geometric interpretation. However, the algorithmic practicalities of implementing such a procedure have been investigated by Marchand and Golea (1993). They propose a method that relies on an incremental approximation algorithm for the NP-hard problem of finding at each stage a hyperplane that chops off as many remaining points as possible (the ‘densest hyperplane problem’). Reports on the experimental performance of their method can be found in Marchand and Golea (1993).

Threshold decision lists are special types of *perceptron decision trees*, decision trees in which the decision nodes compute threshold functions. Such trees have been studied by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000), where the importance of large margins was emphasised. The techniques used to derive the results of Section 4 extend those used to derive generalization error bounds in those papers. Bennett et al. (2000) consider learning algorithms for perceptron decision trees: specifically, they propose and test three variants of the OC1 perceptron decision tree learning algorithm (Murthy et al., 1994) that aim for a large margin of separation at each decision node. The theoretical generalization error bounds they derive apply to the case in which a perceptron decision tree is produced that is consistent with the training sample. The bounds for perceptron decision trees presented in this paper improve upon the bounds presented there and also apply, more generally, to the case in which some empirical error (measured with respect to the margins) is permitted.

The representational properties of threshold decision lists and multilevel threshold functions have been studied by a number of researchers, particularly in the context of Boolean functions. We mentioned above the paper of Jeroslow (1975). There, it is shown, essentially, that any Boolean function can be realized as a disjunction of threshold functions (and hence as a special type of threshold decision list). The general problem of decomposing a Boolean function into a disjunction of threshold functions has been considered independently of any machine learning considerations. Hammer et al. (1981) defined the *threshold number* of a Boolean function to be the minimum s such that f is a disjunction of s threshold functions; they and Zuev and Lipkin (1988) obtained results on the threshold numbers of increasing Boolean functions. Although any Boolean function can be expressed as a disjunction of threshold functions, threshold decision lists provide a more flexible representation. For instance, the parity function on n variables (in which the output is 1 precisely when the input to the function contains an odd number of entries equal to 1) can be represented by a threshold decision list with n terms; whereas, as observed by Jeroslow (1975), the shortest decomposition of parity into a disjunction of threshold functions involves 2^{n-1} threshold functions. Turan and Vatan (1997), by contrast, gave a specific example of a function with a necessarily long threshold decision list representation.

Decision lists in which the tests are defined with respect to points in the training sample have recently been investigated by Sokolova et al. (2003). They considered the case where the base class of tests consists of data-dependent balls (that is, the characteristic functions of balls centered on data points, and their complements). Additionally, Marchand et al. (2003) considered the use of disjunctions and conjunctions of functions constructed as threshold functions, possibly in some ‘feature space’. Here, examples $x \in X$ are transformed by a fixed function ϕ into points of the feature space $\phi(X)$, and the classifiers used are disjunctions or conjunctions of functions that, acting in feature space and on transformed examples $\phi(x)$, are threshold functions with weight vectors defined by three of the transformed examples. The problems studied in this paper are rather different: here, we consider general threshold decision lists (rather than just conjunctions or disjunctions), and the individual tests need not be data-dependent (or, at least, not in the explicit way that they are in Marchand et al., 2003).

Multilevel threshold functions have been studied in a number of papers (Bohossian and Bruck, 1998; Olafsson and Abu-Mostafa, 1988; Takiyama, 1985, for instance). They originally were of interest as the sets of functions computed by devices knows as *multilevel threshold elements* (Takiyama, 1985), generalizations of the linear threshold elements. The ‘capacity’ (in our terminology, the growth function) has been of particular interest. Olafsson and Abu-Mostafa (1988) gave an upper bound on the capacity, correcting a claimed upper bound of Takiyama (1985). Subsequently, Ngom et al. (2003) claimed to have improved this bound, but were mistaken (Anthony, 2002). A bound improving upon that of Olafsson and Abu-Mostafa (1988), and which is used in this paper, was given in Anthony (2002). Just as threshold decision lists (and disjunctions of threshold functions) are ‘universal’ for Boolean functions, so too are the multilevel threshold functions. Bohossian and

Bruck (1998) observed that any Boolean function can be realized as a multilevel threshold function. (Specifically, they showed that every Boolean function is a 2^n-level threshold function, an appropriate weight-vector being $w = (2^{n-1}, 2^{n-2}, \ldots, 2, 1)$. For that reason, they paid particular attention to the question of whether a function can be computed by a multilevel threshold function where the number of levels is polynomial.) Functions similar to multilevel threshold functions have also been of interest in multiple-valued logic (Obradović and Parberry, 1994; Ngom et al., 2003) where, instead of classification labels alternating between 0 and 1, a partition by k parallel planes defines a $(k+1)$-valued function.

2. Generalization Error

Following a form of the PAC model of computational learning theory (see Anthony and Biggs, 1992; Vapnik, 1998; Blumer et al., 1989), we assume that labeled 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 a set of functions from X to $\{0,1\}$. 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 labeled 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 $h \in H$, $\mathrm{er}_P(h) < \mathrm{er}_s(h) + \varepsilon(m,\delta)$, where $\varepsilon(m,\delta)$ (known as a *generalization error bound*) is decreasing in m and δ. Such results can be derived using uniform convergence theorems from probability theory (Vapnik and Chervonenkis, 1971; Pollard, 1984; Dudley, 1999), in which case $\varepsilon(m,\delta)$ would typically involve the growth function (see Vapnik and Chervonenkis, 1971; Blumer et al., 1989; Vapnik, 1998; Anthony and Bartlett, 1999). We indicate in the next section how this may be done for threshold decision lists.

Recently, emphasis has been placed in practical machine learning techniques such as Support Vector Machines (see Cristianini and Shawe-Taylor, 2000, for instance) on 'learning with a large margin'. (See Bartlett et al., 2000; Anthony and Bartlett, 1999, 2000; Shawe-Taylor et al., 1996, for example). 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. In Section 4, we apply such ideas to threshold decision lists.

3. Generalization Bounds Based on the Growth Function

In this section, we use some by-now classical techniques of computational or statistical learning theory to bound the generalization error.

3.1 Bounding the Error

The *growth function* of a set of functions H mapping from $X = \mathbb{R}^n$ to $\{0,1\}$ is defined as follow (Blumer et al., 1989; Vapnik and Chervonenkis, 1971). Let $\Pi_H : \mathbb{N} \to \mathbb{N}$ be given by

$$\Pi_H(m) = \max\{|H|_S| : S \subseteq X, |S| = m\},$$

where $H|_S$ denotes H restricted to domain S. Note that $\Pi_H(m) \leq 2^m$ for all m. The key probability results we employ are the following bounds, due respectively to Vapnik and Chervonenkis (1971) and Blumer et al. (1989) (see also Anthony and Bartlett, 1999): for any $\varepsilon \in (0,1)$,

$$P^m\left(\{s \in Z^m : \text{there exists } f \in H,\ \text{er}_P(f) \geq \text{er}_s(f) + \varepsilon\}\right) < 4\,\Pi_H(2m)\, e^{-m\varepsilon^2/8},$$

and, for $m \geq 8/\varepsilon$,

$$P^m\left(\{s \in Z^m : \text{there exists } f \in H,\ \text{er}_s(f) = 0 \text{ and } \text{er}_P(f) \geq \varepsilon\}\right) < 2\,\Pi_H(2m)\, 2^{-\varepsilon m/2}.$$

Thus, we can obtain (probabilistic) bounds on the error $\text{er}_P(f)$ of a function from a class H when we know something about the growth function of H.

3.2 Growth Function Bounds

We first consider the set of threshold decision lists on $\mathbb{R}^n$ with some number k of terms. (So, the length of the list is no more than k.)

Theorem 1 *Let H be the set of threshold decision lists on $\mathbb{R}^n$ with k terms, where $n,k \in \mathbb{N}$. Then*

$$\Pi_H(m) < 4^k \left(\sum_{i=0}^{n} \binom{m-1}{i}\right)^k.$$

Proof: Let S be any set of m points in $\mathbb{R}^n$. Suppose we have two decision lists

$$f = (f_1, c_1), \ldots, (f_k, c_k),\ \ g = (g_1, d_1), \ldots, (g_k, d_k)$$

in H, where the f_i and g_j are threshold functions on $\mathbb{R}^n$. Clearly, f and g will agree on all points of S if (i) $c_i = d_i$ for each i and (ii) $f_i(x) = g_i(x)$ for all $x \in S$. For fixed i, condition (ii) is an equivalence relation among threshold functions. The number of equivalence classes is $|K|_S|$ where K is the set of threshold functions. This is bounded by $\Pi_K(m)$, which, it is known (Cover, 1965; Blumer et al., 1989; Anthony and Bartlett, 1999), is bounded above as follows:

$$\Pi_K(m) \leq 2\sum_{i=0}^{n} \binom{m-1}{i}.$$

We can therefore upper bound $|H|_S|$ as follows:

$$|H|_S| \leq 2^k \left(2\sum_{i=0}^{n} \binom{m-1}{i}\right)^k.$$

Here, the first 2^k factor corresponds to the number of possible sequences of c_i and the remaining factor bounds the number of ways of choosing an equivalence class (with respect to S) of threshold functions, for each i from 1 to k. ■

There is a useful connection between certain types of decision list and threshold functions. We say that a decision list defined on $\{0,1\}^n$ is a 1-*decision list* if the Boolean function in each test is given by a formula that is a single literal. (So, for each i, there is some l_i such that *either* $f_i(y) = 1$ if and only if $y_{l_i} = 1$, *or* $f_i(y) = 1$ if and only if $y_{l_i} = 0$.) Then, it is known (Ehrenfeucht et al., 1989) (see also Anthony et al., 1995; Anthony, 2001) that any 1-decision list is a threshold function. In an easy analogue of this, any threshold decision list is a threshold function of threshold functions (Anthony, 2001). But a threshold function of threshold functions is nothing more than a two-layer threshold network, one of the simplest types of artificial neural network. (A similar observation was made by Marchand and Golea (1993) and Marchand et al. (1990), who construct a 'cascade' network from a threshold decision list.) So another way of bounding the growth function of threshold decision lists is to use this fact in combination with some known bounds (Baum and Haussler, 1989; Anthony and Bartlett, 1999) for the growth functions of linear threshold networks. This gives a similar, though slightly looser, upper bound.

To bound the growth function of the subclass consisting of k-level threshold functions, we use a result from (Anthony, 2002), which shows that the number of ways in which a set S of m points can be partitioned by k parallel hyperplanes is at most $\sum_{i=0}^{n+k-1} \binom{km}{i}$. (For fixed n and k, this bound is tight to within a constant, as a function of m.) Noting that we may assume adjacent regions to have different labels, there corresponds to each such partition at most two k-level threshold functions (defined on the domain restricted to S) and we therefore have the following bound.

Theorem 2 *Let H be the set of k-level threshold functions on $\mathbb{R}^n$. Then*

$$\Pi_H(m) \leq 2\sum_{i=0}^{n+k-1} \binom{km}{i}.$$

3.3 Generalization Error Bounds

Combining the results of the previous two subsections, we can obtain the following generalization error bounds.

Theorem 3 *Suppose that n and k are fixed positive integers and that s is a sample of m labeled points (x,b) of $Z = \mathbb{R}^n \times \{0,1\}$, each generated at random according to a fixed probability distribution P on Z. Let δ be any positive number less than one. Then the following hold with probability at least $1-\delta$:*

1. *If f is a threshold decision list with k terms, then the error $\mathrm{er}_P(f)$ of f and its sample error on s, $\mathrm{er}_s(f)$ are such that*

$$\mathrm{er}_P(f) < \mathrm{er}_s(f) + \sqrt{\frac{8}{m}\left(2k\ln 2 + nk\ln\left(\frac{e(2m-1)}{n}\right) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

2. *If f is a k-level threshold function, then*

$$\mathrm{er}_P(f) < \mathrm{er}_s(f) + \sqrt{\frac{8}{m}\left((n+k-1)\ln\left(\frac{2emk}{n+k-1}\right) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

Proof: We approximate the growth function of the class of k-term threshold decision lists by

$$\Pi_H(m) \leq 4^k\left(\sum_{i=0}^{n}\binom{m-1}{i}\right)^k < 4^k\left(\frac{e(m-1)}{n}\right)^{nk},$$

for $m > n$. Similarly, when H is the class of k-level threshold functions,

$$\Pi_H(m) \leq 2\sum_{i=0}^{n+k-1}\binom{km}{i} < 2\left(\frac{emk}{n+k-1}\right)^{n+k-1},$$

for $m \geq n+k$. The first part of the theorem is trivially true if $m \leq n$ (since then the stated upper bound on the error is at least 1). If $m > n$, then

$$\varepsilon_0 = \sqrt{\frac{8}{m}\left(2k\ln 2 + nk\ln\left(\frac{e(2m-1)}{n}\right) + \ln\left(\frac{4}{\delta}\right)\right)} \geq \sqrt{\frac{8}{m}\left(\ln\left(\frac{4\Pi_H(2m)}{\delta}\right)\right)},$$

and so

$$P^m\left(\{s \in Z^m : \text{there exists } f \in H,\ \mathrm{er}_P(f) \geq \mathrm{er}_s(f) + \varepsilon_0\}\right) < 4\Pi_H(2m)\, e^{-m\varepsilon_0^2/8} \leq \delta.$$

Thus, with probability at least $1-\delta$, for all $f \in H$, $\mathrm{er}_P(f) < \mathrm{er}_s(f) + \varepsilon_0$. The second part follows similarly. It is trivial for $m < n+k$ and it follows for $m \geq n+k$ on observing that, for H the class of k-level threshold functions,

$$\varepsilon_0' = \sqrt{\frac{8}{m}\left((n+k-1)\ln\left(\frac{2emk}{n+k-1}\right) + \ln\left(\frac{4}{\delta}\right)\right)} \geq \sqrt{\frac{8}{m}\left(\ln\left(\frac{4\Pi_H(2m)}{\delta}\right)\right)}.$$

■

For threshold decision lists that are consistent with a training sample, the following tighter bounds can be used.

Theorem 4 *Suppose that k and n are fixed positive integers and that s is a sample of m labeled points (x,b) of $Z = \mathbb{R}^n \times \{0,1\}$, each generated at random according to a fixed probability distribution P on Z. Let δ be any positive number less than one. Then the following hold with probability at least $1-\delta$:*

1. *If f is a threshold decision list with k terms and f is consistent with s (so that $\mathrm{er}_s(f) = 0$), then*
$$\mathrm{er}_P(f) < \frac{2}{m}\left(2k + nk\log_2\left(\frac{e(2m-1)}{n}\right) + \log_2\left(\frac{2}{\delta}\right)\right).$$

2. *If f is a k-level threshold function and f is consistent with s, then*
$$\mathrm{er}_P(f) < \frac{2}{m}\left((n+k-1)\log_2\left(\frac{2emk}{n+k-1}\right) + \log_2\left(\frac{2}{\delta}\right)\right),$$
for $n+k \geq 3$.

Proof: We use the growth function approximations of the proof of Theorem 3. For the class of threshold decision lists with k terms, and for $m > n$,

$$\varepsilon_0 = \frac{2}{m}\left(2k + nk\log_2\left(\frac{e(2m-1)}{n}\right) + \log_2\left(\frac{2}{\delta}\right)\right) \geq \frac{2}{m}\log_2\left(\frac{2\,\Pi_H(2m)}{\delta}\right),$$

and so

$$P^m\left(\{s \in Z^m : \text{there exists } f \in H,\ \mathrm{er}_s(f) = 0 \text{ and } \mathrm{er}_P(f) \geq \varepsilon_0\}\right) < 2\,\Pi_H(2m)\,2^{-\varepsilon_0 m/2} = \delta.$$

(Also, for $m \leq n$, the bound trivially holds.) The second part follows similarly on noting that, for the class of k-level threshold functions, if $m \geq n+k$, then

$$\varepsilon_0' = \frac{2}{m}\left((n+k-1)\log_2\left(\frac{2emk}{n+k-1}\right) + \log_2\left(\frac{2}{\delta}\right)\right) \geq \frac{2}{m}\log_2\left(\frac{2\,\Pi_H(2m)}{\delta}\right).$$

The bound is trivially true for $m < n+k$; and, for $m \geq n+k$, the condition $n+k \geq 3$ ensures that $m \geq 8/\varepsilon_0$, so that the bound of Blumer et al. (1989) applies. ■

The following variations of these results, in which k is not prescribed in advance, are perhaps more useful, since one does not necessarily know *a priori* how many terms a suitable threshold decision list will have.

Theorem 5 *With the notations as above, and for $n \geq 3$, the following holds with probability at least $1-\delta$:*

1. *If f is a threshold decision list, then*
$$\mathrm{er}_P(f) < \mathrm{er}_s(f) + \sqrt{\frac{8}{m}\left(2k\ln 2 + nk\ln\left(\frac{e(2m-1)}{n}\right) + \ln\left(\frac{14k^2}{\delta}\right)\right)},$$
where k is the number of terms of f.

2. *If f is a multilevel threshold function, then*

$$\mathrm{er}_P(f) < \mathrm{er}_s(f) + \sqrt{\frac{8}{m}\left((n+k-1)\ln\left(\frac{2emk}{n+k-1}\right) + \ln\left(\frac{14k^2}{\delta}\right)\right)},$$

   where k is the number of levels (terms) of f.

3. *If f is a threshold decision list and* $\mathrm{er}_s(f) = 0$, *then*

$$\mathrm{er}_P(f) < \frac{2}{m}\left(2k + nk\log_2\left(\frac{e(2m-1)}{n}\right) + \log_2\left(\frac{4k^2}{\delta}\right)\right)$$

   where k is the number of terms of f;

4. *If f is a multilevel threshold function and* $\mathrm{er}_s(f) = 0$, *then*

$$\mathrm{er}_P(f) < \frac{2}{m}\left((n+k-1)\log_2\left(\frac{2emk}{n+k-1}\right) + \ln\left(\frac{4k^2}{\delta}\right)\right),$$

   where k is the number of terms of f.

Proof: We prove the last part, the other three being very similar. We use a well-known technique often found in discussions of ‘structural risk minimisation’ and model selection (see Vapnik, 1982; Shawe-Taylor et al., 1996; Anthony and Bartlett, 1999, for instance). From Theorem 4, for any $\delta \in (0,1)$ and any $k \in \mathbb{N}$, if $n+k \geq 3$, then the probability p_k that there is a k-level threshold function f such that $\mathrm{er}_s(f) = 0$ and

$$\mathrm{er}_P(f) < \varepsilon'_k = \frac{2}{m}\left((n+k-1)\log_2\left(\frac{2emk}{n+k-1}\right) + \ln\left(\frac{2\pi^2k^2}{6\delta}\right)\right)$$

is less than $(\delta/k^2)(6/\pi^2)$. The fact that $n \geq 3$ ensures that $n+k \geq 3$. Hence the probability that, for some $k \in \mathbb{N}$, there is $f \in H$ with $\mathrm{er}_P(f) \geq \mathrm{er}_s(f) + \varepsilon_k$ is less than $\sum_{k=1}^{\infty} p_k < \delta(6/\pi^2)\sum_{k=1}^{\infty}(1/k^2) = \delta$. The result follows on noting that $2\pi^2/6 < 4$. ■

4. Margin-Based Error Bounds for Threshold Decision Lists

We now derive generalization error bounds dependent on the size of margins. The key qualitative difference between these bounds and those of Section 3 is that the margin-based bounds are dimension-independent, in that they do not depend on n.

4.1 Definition of Margin Error

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. Assume also, without any loss of generality, that $\|w_i\| = 1$ for each i. We say that h classifies the labeled 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.[1] 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 labeled 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 labeled 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.

Following the analysis of perceptron decision trees in Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000), we may want to consider separate margin parameters $\gamma_1,\gamma_2,\ldots,\gamma_k$ for each of the k terms of the decision list. We have the following definition.

Definition 6 *Suppose $h = (t_1,c_1),\ldots,(t_k,c_k)$ is a threshold decision list, where t_i is represented by weight vector w_i and threshold θ_i, where $\|w_i\| = 1$. Given $\Gamma = (\gamma_1,\gamma_2,\ldots,\gamma_k)$, we say that h classifies the labeled example (x,b) (correctly and) with margin Γ if $h(x) = b$ and, for all $1 \le i \le k$, $|\langle w_i,x\rangle - \theta_i| \ge \gamma_i$. We define $\mathrm{er}_s^{\Gamma}(h)$ to be the proportion of labeled examples in the sample s that are not classified with margin Γ.*

4.2 Covering Numbers

A useful 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 unlabeled sample of m points of X. Then, for $\varepsilon > 0$, $C \subseteq F$ 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 $d_\infty^x(f,\hat{f}) < \varepsilon$, where, for $f,g \in F$,

$$d_\infty^x(f,g) = \max_{1\le i\le m} |f(x_i) - g(x_i)|.$$

(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 $\varepsilon > 0$ and all $x \in X^m$ (for all m). In this case, given $x \in X^m$, we define the *d_∞^x-covering number* $\mathcal{N}_\infty(F,\varepsilon,x)$ to be the minimum cardinality of an ε-cover of F with respect to the d_∞^x-metric. We then define the (uniform) *d_∞-covering numbers* $\mathcal{N}_\infty(F,\varepsilon,m)$ by

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

Many bounds on covering numbers for specific classes have been obtained (see Anthony and Bartlett, 1999, 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 (Alon et al., 1997).

In this paper, we use a recent bound of Zhang (2002) 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

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

[1] The assumption that $\|w_i\| = 1$ ensures that the interpretation in terms of distance is valid; for, in this case the 'functional' and 'geometric' margins coincide. See the paper by Cristianini and Shawe-Taylor (2000).

Zhang (2002) has shown that

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

One thing of note is that this bound is dimension-independent: it does not depend on n. This bound differs from previous bounds (Bartlett, 1998; Anthony and Bartlett, 1999; Shawe-Taylor et al., 1996) for the logarithm of the d_∞-covering numbers in that it involves a factor of order $\ln m$ rather than $(\ln m)^2$.[2]

4.3 Margin-based Bounds

Following a method used by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000), together with the covering number bound of Zhang (2002), 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 7 *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*

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

where $D(\Gamma) = \sum_{i=1}^k (1/\gamma_i^2)$ and where the margin error $\mathrm{er}_s^\Gamma(h)$ is as in Definition 6.

Proof: The proof extends a technique of Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000) (where the case of zero margin error was the focus), and is motivated by proofs of Anthony and Bartlett (1999, 2000), Bartlett (1998), Shawe-Taylor et al. (1996), which in turn are based on the original work of Vapnik and Chervonenkis (1971).

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

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

and

$$T = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \varepsilon/2\},$$

then for $m \geq 2/\varepsilon^2$, $P^m(Q) \leq 2P^{2m}(T)$. For, we have

$$\begin{aligned} P^{2m}(T) &\geq P^{2m}\left(\exists h \in H : \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \varepsilon \text{ and } \mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \varepsilon/2\right) \\ &= \int_Q P^m\left(\left\{s' : \exists h \in H, \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \varepsilon \text{ and } \mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \varepsilon/2\right\}\right) dP^m(s) \\ &\geq \frac{1}{2}P^m(Q), \end{aligned}$$

[2] Previous approaches to bounding the d_∞-covering numbers first bounded the *fat-shattering dimension* and then used a result of Alon et al. (1997) that relates the covering numbers to the fat-shattering dimension. An additional $\ln m$ factor appears when this route is taken.

for $m \geq 2/\varepsilon^2$, where the final inequality follows from $P^m(\mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \varepsilon/2) \geq 1/2$, for any $h \in H$, true by Chebyshev's inequality.

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, $P^{2m}(T) \leq \max\{\Pr(\sigma z \in T) : z \in Z^{2m}\}$, where Pr denotes the probability over uniform choice of σ from G. (See Vapnik and Chervonenkis, 1971, and Anthony and Bartlett, 1999, for instance.)

Given a threshold decision list on $B_R \subseteq \mathbb{R}^n$, 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 B_R 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 B_R can be realized as one on $\mathbb{R}^{n+1}$, defined on the subset $B_R \times \{-1\}$, 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 minimum-sized $\gamma_i/2$-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),

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

Suppose that $h = (f_1, c_1), \ldots, (f_k, c_k)$ 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 $\hat{f}_i$, leaving the c_i unchanged. The set $\hat{H}$ of all possible such $\hat{h}$ is of cardinality at most $2^k \prod_{i=1}^k |C_i|$ (where the 2^k factor corresponds to the choices of the values c_i). Suppose that $\sigma z = (s, s') \in T$ and that $\mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \varepsilon/2$. 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, $\mathrm{er}_s^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s^{\Gamma/2}(\hat{h}) + \varepsilon/2$, and therefore, for any $z \in Z^{2m}$,

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

where $S(\hat{h}) = \{(s, s') \in Z^{2m} : \mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s^{\Gamma/2}(\hat{h}) + \varepsilon/2\}$. Fix $\hat{h} \in \hat{H}$ and let $v_i = 0$ if $\hat{h}$ classifies z_i with margin at least $\Gamma/2$, and 1 otherwise. Then

$$\Pr\left(\sigma z \in S(\hat{h})\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^m (v_{m+i} - v_i) \geq \varepsilon/2\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^m \varepsilon_i |v_i - v_{m+i}| \geq \varepsilon/2\right),$$

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 $\exp(-\varepsilon^2 m/8)$. (See Anthony and Bartlett, 1999, for instance, for details.)

We therefore have

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

which gives

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

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

$$\varepsilon \geq \varepsilon_0 = \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \frac{288R^2}{\gamma_i^2} \ln\left(\frac{60Rm}{\gamma_i}\right) + \ln\left(\frac{2}{\delta}\right) + k\right)},$$

(in which case we certainly also have $m \geq 2/\varepsilon^2$) then the probability of Q is at most δ. So, with probability at most $1-\delta$, $\text{er}_P(h) < \text{er}_s^\Gamma(h) + \varepsilon_0$ for all $h \in H$. If, for each i, $m \geq R^2/\gamma_i^2$, then $\ln(60Rm/\gamma_i) \leq 2\ln(8m)$ and so, with probability at least $1-\delta$, for all $h \in H$,

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

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. ■

A tighter bound can be given when the margin error is zero, as follows. (The bound involves $1/m$ rather than $1/\sqrt{m}$.)

Theorem 8 *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*

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

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

Proof: This proof is similar to that of Theorem 7. It uses, first, the fact[3] that if

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

and

$$T = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \text{er}_s^\Gamma(h) = 0, \text{er}_{s'}(h) \geq \varepsilon/2\},$$

[3] For similar results, see Vapnik and Chervonenkis (1971); Blumer et al. (1989); and Anthony and Bartlett (1999).

then, for $m \geq 8/\varepsilon$, $P^m(Q) \leq 2P^{2m}(T)$. This is so, because

$$\begin{aligned} P^{2m}(T) &\geq P^{2m}\left(\exists h \in H : \mathrm{er}_s^{\Gamma}(h) = 0,\ \mathrm{er}_P(h) \geq \varepsilon \text{ and } \mathrm{er}_{s'}(h) \geq \varepsilon/2\right) \\ &= \int_Q P^m\left(\left\{s' : \exists h \in H,\ \mathrm{er}_s^{\Gamma}(h) = 0,\ \mathrm{er}_P(h) \geq \varepsilon \text{ and } \mathrm{er}_{s'}(h) \geq \varepsilon/2\right\}\right) dP^m(s) \\ &\geq \frac{1}{2}P^m(Q), \end{aligned}$$

for $m \geq 8/\varepsilon$. The final inequality follows from the fact that if $\mathrm{er}_P(h) = 0$, then for $m \geq 8/\varepsilon$, $P^m(\mathrm{er}_{s'}(h) \geq \varepsilon/2) \geq 1/2$, for any $h \in H$, something that follows for $m \geq 8/\varepsilon$ by Chebyshev's inequality or a Chernoff bound (Anthony and Biggs, 1992, for instance). As before, $P^{2m}(T) \leq \max_{z \in Z^{2m}} \Pr(\sigma z \in T)$, where Pr denotes the probability over uniform choice of σ from the 'swapping group' G. A very similar argument to that given in the proof of Theorem 3 establishes that for any $z \in Z^{2m}$,

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

where $S(\hat{h}) = \{(s,s') \in Z^{2m} : \mathrm{er}_{\hat{s}}^{\Gamma/2}(h) = 0, \mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) \geq \varepsilon/2\}$. Now, suppose $S(\hat{h}) \neq \emptyset$, so that for some $\tau \in G$, $\tau z = (s,s') \in S(\hat{h})$, meaning that $\mathrm{er}_s^{\Gamma/2}(\hat{h}) = 0$ and $\mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) \geq \varepsilon/2$. Then, by symmetry, $\Pr\left(\sigma z \in S(\hat{h})\right) = \Pr\left(\sigma(\tau z) \in S(\hat{h})\right)$. Suppose that $\mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) = r/m$, where $r \geq \varepsilon m/2$ is the number of x_i in s' not classified with margin $\Gamma/2$ by $\hat{h}$. Then those permutations σ such that $\sigma(\tau z) \in S(\hat{h})$ are precisely those that 'swap' elements other than these r, and there are $2^{m-r} \leq 2^{m-\varepsilon m/2}$ such σ. It follows that, for each fixed $\hat{h} \in \hat{H}$,

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

The proof then proceeds as does the proof of Theorem 7, using the bound (2). ∎

4.4 Uniform Margin-based Bounds

One difficulty with Theorems 7 and 8 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, tune, or observe these parameters after learning. We now derive such a result. The approach we take to obtaining a 'uniform' result of this type differs from that taken by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000), and gives a slightly better bound.

We first need a generalization of a result from (Bartlett, 1998), 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 modify and extend this result as follows.

Theorem 9 *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:

(a) for all $\Gamma \in (0,1]^k$, $\mathbb{P}(E(\Gamma, \Gamma, \delta)) \leq \delta$,

(b) $\Gamma_1 \leq \Gamma \leq \Gamma_2$ (component-wise) and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\Gamma_1, \Gamma_2, \delta_1) \subseteq E(\Gamma, \Gamma, \delta)$.

Then

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

for $0 < \delta < 1$, where

$$c(\Gamma) = \left\{ \prod_{i=1}^{k} \log_2\left(\frac{4}{\gamma_i}\right) \right\}^{-2}.$$

Proof: Denoting by $\mathbf{u} = (1, 1, \ldots, 1)$ the all-1 vector of length k, we have

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

Here, we have used property (b) of the events $E(\Gamma_1, \Gamma_2, \delta)$, together with the following two observations: if $\gamma_j \in ((1/2)^{i_j+1}, (1/2)^{i_j}]$, then $(1/2)\Gamma \leq (1/2)^{i_j+1}\mathbf{u}$ and $\Gamma \geq (1/2)^{i_j+1}\mathbf{u}$; and $\gamma_i \in ((1/2)^{i_j+1}, (1/2)^{i_j}]$ implies

$$\left(\log_2\left(\frac{4}{\gamma_j}\right) \right)^2 \geq (i_j+2)^2 \geq (i_j+1)(i_j+2),$$

so that

$$c(\Gamma) \leq \prod_{j=1}^{k} \frac{1}{(i_j+1)(i_j+2)}.$$

Now, by property (a),

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

■

We can now obtain the following ‘uniform’ result.

Theorem 10 *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*

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304R^2D(\Gamma)\ln(8m)+\ln\left(\frac{2}{\delta}\right)+2k+2\sum_{i=1}^{k}\ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right)},$$

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*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(2304R^2D(\Gamma)\log_2(8m)+\log_2\left(\frac{2}{\delta}\right)+2k+2\sum_{i=1}^{k}\ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right),$$

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

Proof: 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

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

where $D(\Gamma_1) = \sum_{i=1}^{k}(1/\gamma_i^{(1)})^2$. Then, by Theorem 7, $P^m(E(\Gamma,\Gamma,\delta)) \leq \delta$, 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

$$P^m\left(\bigcup_{\Gamma\in(0,1]^k} E\left((1/2)\Gamma,\Gamma,\delta c(\Gamma)\right)\right) \leq \delta,$$

where

$$c(\Gamma) = \left\{\prod_{i=1}^{k}\log_2\left(\frac{4}{\gamma_i}\right)\right\}^{-2}.$$

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

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

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,

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

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,

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(2304R^2D(\Gamma)\ln(8m) + \ln\left(\frac{2}{\delta}\right) + 2k + 2\sum_{i=1}^{k}\ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right)}.$$

(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 8. ■

4.5 Comparison with Related Results

Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000) 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

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

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

By comparison, the bound given in Theorem 10 is of order

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

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), rather than bounding the covering number by using results on fat-shattering dimension, coupled with the bound ofAlon et al. (1997). 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 $O(k + \ln\ln m)$ rather than the $O(k \ln m)$ of (4), and they are potentially much smaller. This improvement results from the use of Theorem 9. Theorem 10 is therefore an improvement over the results implied by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000).

4.6 Bounds for Perceptron Decision Trees

Although the focus of this paper is threshold decision lists, we now show how the analysis here can be used to improve and extend results on perceptron decision trees given by Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000). Recall that these are decision trees in which the decision nodes compute threshold functions. The definition of margin error $\mathrm{er}^{\Gamma}(h)$ for a perceptron decision tree classifier h is defined in a straightforward way by extending Definition 6. Suppose the threshold functions computed at the decision nodes are $t_1, \ldots, t_k$, where k is the number of decision nodes, and suppose that t_i is represented by weight vector w_i and threshold θ_i, where $\|w_i\| = 1$. Given $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, we say that the tree h classifies the labeled example (x, b) with margin Γ if $h(x) = b$ and, for all $1 \le i \le k$, $|\langle w_i, x\rangle - \theta_i| \ge \gamma_i$. Then, for a labeled sample s, $\mathrm{er}_s^{\Gamma}(h)$ is the proportion of labeled examples in s that are not classified with margin Γ.

Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000) obtain a generalization error bound for the special case in which the margin error is zero. The following theorem improves that bound and provides a bound applicable in the case of non-zero margin error. (We have stated only a 'uniform' result; that is, one in which the margin parameters and tree size are not fixed *a priori*. However, embedded in the proof are the corresponding non-uniform results.) The proof is a modification of the proofs of the theorems for threshold decision lists, in which we make use of the fact that the number of binary trees with k vertices—and hence the number of decision tree skeletons with k decision nodes (as noted in Quinlan and Rivest, 1989; Bennett et al., 2000)—is given by the Catalan number $N_k = \frac{1}{k+1}\binom{2k}{k}$.

Theorem 11 *For $k \in \mathbb{N}$, let $N_k = \dfrac{1}{k+1}\dbinom{2k}{k}$. Suppose $R \ge 1$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \le R\}$. Let H be the set of all perceptron decision trees (of any size and structure) 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 decision nodes, and $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304 R^2 D(\Gamma)\ln(8m) + \ln\left(\frac{2^{k+2}N_k}{\delta}\right) + k + 2\sum_{i=1}^{k}\ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right)},$$

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 *decision nodes, and* h *classifies* s *with margin* $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, *then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(2304R^2 D(\Gamma)\log_2(8m) + \log_2\left(\frac{2^{k+2}N_k}{\delta}\right) + k + 2\sum_{i=1}^{k}\ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right),$$

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

Proof: The proof is similar to that of Theorems 7, 8 and 10, so we will omit some of the detail. As in the proof of Theorem 7, for any $k \in \mathbb{N}$, for H the class of perceptron decision trees (or, rather, the functions represented by such trees) with k decision nodes, for any Γ, if

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

and

$$T = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^{\Gamma}(h) + \epsilon/2\},$$

then for $m \geq 2/\epsilon^2$, $P^m(Q) \leq 2P^{2m}(T)$. With G the swapping permutation group, we have, as before, $P^{2m}(T) \leq \max\{\Pr(\sigma z \in T) : z \in Z^{2m}\}$, where Pr denotes the probability over uniform choice of σ from G. Given a perceptron decision tree on B_R, we may (as discussed in the proof of Theorem 7) realize the tree as one defined on $D = \{(x,-1) : x \in \mathbb{R}^n, \|x\| \leq R\}$, in which the decision nodes compute homogeneous threshold functions. Fixing $z \in Z^{2m}$, and arguing as in the proof of Theorem 7, for i between 1 and k, let C_i be a minimal-cardinality $\gamma_i/2$-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 D. Then $|C_i|$ is bounded as in (2). Suppose that, in a given perceptron decision tree h, and at a given decision node, the test is given by the threshold function f_i, represented by weight vector w_i and let $\hat{w}_i$ be an element of the cover C_i which is distance less than $\gamma_i/2$ from w_i. Then a very similar analysis to that in Theorem 7 establishes that if $\hat{h}$ is the tree obtained by replacing each f_i by $\hat{f}_i$, we have

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

where $S(\hat{h}) = \{(s,s') \in Z^{2m} : \mathrm{er}_{s'}^{\Gamma/2}(\hat{h}) \geq \mathrm{er}_s^{\Gamma/2}(\hat{h}) + \epsilon/2\}$. Now, the set $\hat{H}$ of all such $\hat{h}$ will have cardinality bounded as follows:

$$|\hat{H}| \leq 2^{k+1}N_k\prod_{i=1}^{k}|C_i| \leq 2^{k+1}N_k\prod_{i=1}^{k}2^{(288R^2/\gamma_i^2)\log_2(60Rm/\gamma_i)},$$

where the factor of 2^{k+1} accounts for the possible binary values at the $k+1$ leaves of the tree, and N_k accounts for the number of skeletons of trees with k decision nodes. By arguing precisely as in Theorem 7, we can then establish that for $\delta \in (0,1)$, for fixed k and fixed Γ, with probability at least $1-\delta$, for all perceptron decision trees with k decision nodes, $\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \epsilon(\Gamma, \delta, k, m)$, where

$$\epsilon(\Gamma,\delta,k,m) = \sqrt{\frac{8}{m}\left(576R^2D(\Gamma)\ln(8m) + \ln\left(\frac{2^{k+2}N_k}{\delta}\right)\right)},$$

Next, we apply Theorem 9. Fixing k and taking $E(\Gamma_1, \Gamma_2, \delta)$ to be the event that there exists a perceptron decision tree h with k decision nodes such that $\text{er}_P(h) \geq \text{er}_s^{\Gamma_2}(h) + \varepsilon(\Gamma_1, \delta, k, m)$, we establish that with probability at least $1-\delta$, for *all* Γ,

$$\text{er}_P(h) < \text{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(2304 R^2 D(\Gamma) \ln(8m) + \ln\left(\frac{2^{k+2} N_k}{\delta}\right) + 2\sum_{i=1}^{k} \ln\left(\log_2\left(\frac{4}{\gamma_i}\right)\right)\right)}.$$

Finally, replacing δ by $\delta/2^k$ and proceeding as in the final part of the proof of Theorem 9, we obtain the desired result. The proof of the second part of the theorem is similar. ■

The result given in Shawe-Taylor and Cristianini (1998) and Bennett et al. (2000) corresponds to the second case given in Theorem 11 and takes the form: with probability at least $1-\delta$, for all Γ and for any perceptron decision tree such that $\text{er}^{\Gamma}(h) = 0$,

$$\text{er}_P(h) < O\left(\frac{1}{m}\left(D(\Gamma)(\ln m)^2 + k \ln m + \ln N_k + \ln\left(\frac{1}{\delta}\right)\right)\right),$$

where we have suppressed the constants. Theorem 11 improves upon this, as can be seen by similar considerations to those made in comparing bounds (4) and (5) above. In particular, an expression of order $D(\Gamma)(\ln m)^2 + k \ln m$ is replaced by one of order $D(\Gamma) \ln m + k + \ln\ln m$.

5. Margin-Based Error Bounds for Multilevel Threshold Functions

Suppose that h is a k-level threshold function, represented by weight vector w with $\|w\| = 1$ 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 $t_i(y) = \text{sgn}(\langle w, x\rangle - \theta_i)$. Recall that we say h classifies the labeled 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 labeled sample s, $\text{er}_s^{\gamma}(h)$, the sample error at margin γ, is the proportion of labeled 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 considers 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 $\gamma \leq 1$.)

5.1 Generalization Error Bounds for k-level Threshold Functions

We have the following result.

Theorem 12 *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*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\gamma(h) + \sqrt{\frac{8}{m}\left(\frac{1152R^2}{\gamma^2}\ln(9m) + k\ln\left(\frac{10R}{\gamma}\right) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

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

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

and

$$T = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\gamma(h) + \varepsilon/2\},$$

then $P^m(Q) \leq 2P^{2m}(T)$. 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 minimum size of L with respect to the d_∞^x metric. By (1),

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

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

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

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 $w \in \mathbb{R}^n$ with $\|w\| = 1$ and a threshold vector $\theta \in \mathbb{R}^k$. Since, for all $x \in B_R$, $|\langle w,x\rangle| \leq \|w\|\|x\| = \|x\| \leq R$, 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

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

(Here, the first factor of 2 accounts for the two different ways in which the classifications can alternate as we traverse the planes is a normal direction.) 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 $1 \le j \le k$,

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

It follows that $\mathrm{er}_{s'}^{\gamma/2}(\hat{h}) \ge \mathrm{er}_{s'}(h)$ and $\mathrm{er}_s^{\gamma}(h) \ge \mathrm{er}_s^{\gamma/2}(\hat{h})$. So, if we have $\sigma z = (s, s') \in T$ and $\mathrm{er}_{s'}(h) \ge \mathrm{er}_s^{\gamma}(h) + \varepsilon/2$, then

$$\mathrm{er}_{s'}^{\gamma/2}(\hat{h}) \ge \mathrm{er}_{s'}(h) \ge \mathrm{er}_s^{\gamma}(h) + \varepsilon/2 \ge \mathrm{er}_s^{\gamma/2}(\hat{h}) + \varepsilon/2.$$

The proof now proceeds as the proof of Theorem 7. For any $z \in Z^{2m}$,

$$\Pr(\sigma z \in T) \le \Pr\left(\sigma z \in \bigcup_{\hat{h} \in \hat{H}} S(\hat{h})\right),$$

where

$$S(\hat{h}) = \{(s, s') \in Z^{2m} : \mathrm{er}_{s'}^{\gamma/2}(\hat{h}) \ge \mathrm{er}_s^{\gamma/2}(\hat{h}) + \varepsilon/2\}.$$

Fixing $\hat{h} \in \hat{H}$, we find that, by Hoeffding's inequality,

$$\Pr\left(\sigma z \in S(\hat{h})\right) \le \exp(-\varepsilon^2 m/8).$$

Therefore,

$$P^m(Q) < 2\,|\hat{H}|\,\exp(-\varepsilon^2 m/8) \le 4\,2^{576R^2/\gamma^2 \log_2(80Rm/\gamma)} \left(\frac{10R}{\gamma}\right)^k \exp(-\varepsilon^2 m/8).$$

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

$$\mathrm{er}_P(h) < \mathrm{er}_s(h) + \sqrt{\frac{8}{m}\left(\left(\frac{576R^2}{\gamma^2}\right)\ln\left(\frac{80Rm}{\gamma}\right) + k\ln\left(\frac{10R}{\gamma}\right) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

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 $m \ge R^2/\gamma^2$. ■

For the case in which the margin error is zero, a better bound can be derived.

Theorem 13 *Suppose $R > 0$ and $Z = B_R \times \{0,1\}$, where $B_R = \{x \in \mathbb{R}^n : \|x\| \le R\}$. 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 obtained by modifying the proof of Theorem 12, just in the same way as Theorem 8 is obtained by modifying the proof of Theorem 7. First, one uses the fact that if

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

and

$$T = \{(s,s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_s^{\gamma}(h) = 0, \mathrm{er}_{s'}(h) \geq \varepsilon/2\},$$

then, for $m \geq 8/\varepsilon$, $P^m(Q) \leq 2P^{2m}(T)$. As before, $P^{2m}(T) \leq \max_{z \in Z^{2m}} \Pr(\sigma z \in T)$, where Pr denotes the probability over uniform choice of σ from the 'swapping group' G. Then, it can be seen that for any $z \in Z^{2m}$,

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

where $S(\hat{h}) = \{(s,s') \in Z^{2m} : \mathrm{er}_{\hat{s}}^{\gamma/2}(h) = 0, \mathrm{er}_{s'}^{\gamma/2}(\hat{h}) \geq \varepsilon/2\}$ and where $\hat{H}$ is as in the proof of Theorem 12. Arguing as in the proof of Theorem 8, if $S(\hat{h}) \neq \emptyset$, so that for some $\tau \in G$, $\tau z = (s,s') \in S(\hat{h})$, then $\Pr\left(\sigma z \in S(\hat{h})\right) = \Pr\left(\sigma(\tau z) \in S(\hat{h})\right)$. Supposing that $\mathrm{er}_{s'}^{\gamma/2}(\hat{h}) = r/m$, where $r \geq \varepsilon m/2$ is the number of x_i in s' not classified with margin $\gamma/2$ by $\hat{h}$, we see that there are at most $2^{m-r} \leq 2^{m-\varepsilon m/2}$ σ such that $\sigma(\tau z) \in S(\hat{h})$. Hence, for each $\hat{h} \in \hat{H}$,

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

The proof then proceeds as does the proof of Theorem 12. ■

5.2 Uniform Margin-based Bounds for Multilevel Threshold Functions

It is straightforward to remove the *a priori* specification of γ and k, using Theorem 9. The following bounds are obtained.

Theorem 14 *Suppose $R > 0$ 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 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*

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\gamma}(h) + \varepsilon(\gamma, \delta, k, m)$$

where $\varepsilon = \varepsilon(\gamma, \delta, k, m)$ is given by

$$\varepsilon = \sqrt{\frac{8}{m}\left(\frac{4608R^2}{\gamma^2}\ln(9m) + k + k\ln\left(\frac{20R}{\gamma}\right) + \ln\left(\frac{4}{\delta}\right) + 2\ln\left(\log_2\left(\frac{4}{\gamma}\right)\right)\right)}.$$

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*

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

Proof: Let $E(\gamma_1,\gamma_2,\delta) \subseteq Z^m$ be the event that there exists $h \in H$ with k terms such that

$$\mathrm{er}_P(h) \geq \mathrm{er}_s^{\gamma_2}(h) + \varepsilon'(\gamma_1,\delta,k,m),$$

where

$$\varepsilon'(\gamma,\delta,k,m) = \sqrt{\frac{8}{m}\left(\frac{1152R^2}{\gamma^2}\ln(9m) + k\ln\left(\frac{10R}{\gamma}\right) + \ln\left(\frac{4}{\delta}\right)\right)}.$$

Then, by Theorem 12, $P^{2m}(E(\gamma,\gamma,\delta)) \leq \delta$. It is also clear that $0 < \gamma_1 \leq \gamma \leq \gamma_2 < 1$ and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\gamma_1,\gamma_2,\delta_1) \subseteq E(\gamma,\gamma,\delta)$. By Theorem 9, with $\delta/2^k$ in place of δ, we therefore have that, for any fixed $k \in \mathbb{N}$, with probability at least $1-\delta/2^k$, for *all* $\gamma \in (0,1]$, every k-level threshold function h satisfies

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\gamma}(h) + \varepsilon'(\gamma/2, \delta c(\gamma)/2^k, k, m),$$

where $c(\gamma) = 1/(\log_2(4/\gamma))^2$. Thus, with probability at least $1-\delta$, *for all* $\gamma \in (0,1]$ and *all* $k \in \mathbb{N}$, every k-level threshold function has

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\gamma}(h) + \varepsilon'(\gamma/2, \delta c(\gamma)/2^k, k, m) \leq \mathrm{er}_s^{\gamma}(h) + \varepsilon(\gamma,\delta,k,m).$$

The first part of the result now follows, and the second is proved similarly, using Theorem 13. ■

5.3 Comparison with the Bounds for General Threshold Decision Lists

The generalization error bound implied by Theorem 7 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 12 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. Similar comments apply to the respective ‘uniform’ bounds of Theorem 10 and Theorem 14.

6. Conclusions and Further Work

This paper has derived different types of theoretical bounds on the generalization error of threshold decision lists. Applying the standard PAC model, by bounding the growth functions, we have given bounds for threshold decision lists and multilevel threshold functions. We then derived generalization error bounds that involve the margins by which successive planes in the threshold decision list ‘clear’ the training examples. These bounds improve upon those that follow (for the special case in which the margin error is zero) from earlier results of Bennett et al. (2000) and Shawe-Taylor and Cristianini (1998). Although threshold decision lists have been the focus of this paper, we have also presented generalization error bounds for perceptron decision trees that improve and extend (to the case in which margin error need not be zero) previous such bounds from Bennett et al. (2000) and Shawe-Taylor and Cristianini (1998). For the subclass of multilevel threshold functions (those threshold decision lists in which the defining hyperplanes may be taken to be parallel), a different approach to constructing empirical covers has been shown to lead to better margin-based bounds than those that would follow from the general bounds obtained for threshold decision lists.

There are several possible directions for further investigation.

We used upper bounds on the growth functions of threshold decision lists and multilevel threshold functions to upper bound generalization error. An interesting combinatorial question concerns the VC-dimension of these classes. Lower bounds on the VC-dimension would provide worst-case lower bounds on generalization error (see Ehrenfeucht et al., 1989; Anthony and Biggs, 1992; Anthony and Bartlett, 1999; Blumer et al., 1989). Certainly, upper bounds on the VC-dimensions follow from the bounds we obtained on the growth functions, but these are quite likely to be loose and a more direct attempt might be productive in obtaining not only better upper bounds, but also lower bounds, on the VC-dimension.

There are other approaches to deriving generalization error bounds. Of particular importance recently have been methods using Rademacher complexity and local Rademacher complexity, together with concentration-of-measure results (Bartlett and Mendelson, 2001; Mendelson, 2003; Bartlett et al., 2002; Bousquet et al., 2002; Bousquet, 2003). It would be interesting to investigate such approaches for threshold decision lists.

The margin-based results obtained here for multilevel threshold functions only involve a single margin parameter rather than separate ones for each plane, and it is possible that a different approach might permit such added flexibility.

We have not considered in this paper the algorithmics of learning threshold decision lists. As mentioned, heuristics for learning threshold decision lists were studied by Marchand and Golea (1993), and although no theoretical generalization error bounds were derived there, the techniques appeared to perform well in experiments. Furthermore, the perceptron decision tree algorithms FAT, MOC1, and MOC2 due to Bennett et al. (2000) are variants of the OC1 algorithm (Murthy et al., 1994) that are explicitly driven by the aim of maximising the margins at the decision nodes. It would be interesting to modify the techniques of Marchand and Golea (1993) with a view to obtaining large margins, and to modify the algorithms of Bennett et al. (2000) so as to learn a threshold decision list rather than a perceptron decision tree.

Acknowledgements

I am grateful to the referees for a number of helpful suggestions and, in particular, for comments leading to an improvement of Theorem 9.

References

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

M. Anthony (2001). *Discrete Mathematics of Neural Networks: Selected Topics*. SIAM Monographs on Discrete Mathematics and Applications. Society for Industrial and Applied Mathematics, Philadelphia, PA.

M. Anthony (2002). Partitioning points by parallel planes. RUTCOR Research Report RRR-39-2002, Rutgers Center for Operations Research. (Also, CDAM research report LSE-CDAM-2002-10, Centre for Discrete and Applicable Mathematics, London School of Economics.) To appear, *Discrete Mathematics*.

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

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

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

M. Anthony, G. Brightwell and J. Shawe-Taylor (1995). On specifying Boolean functions by labelled examples. *Discrete Applied Mathematics*, 61: 1–25.

P. L. Bartlett (1998). 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): 525–536.

P. L. Bartlett, O. Bousquet and S. Mendelson (2002), Localized Rademacher complexities. *Proceedings of the* 15*th Annual Conference on Computational Learning Theory*, ed. J. Kivinen and R. H. Sloan. Springer Lecture Notes in Artificial Intelligence 2375.

P. Bartlett and S. Mendelson (2001), Rademacher and Guassian complexities: risk bounds and structural results. In *Proceedings of the* 14*th Annual Conference on Computational Learning Theory*, Lecture Notes in Artificial Intelligence, Springer, 224-240.

E. Baum and D. Haussler (1989). What size net gives valid generalization? *Neural Computation*, 1(1): 151–160.

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

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

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

O. Bousquet (2003). New Approaches to Statistical Learning Theory. Annals of the Institute of Statistical Mathematics 55 (2): 371-389.

S. Boucheron, G. Lugosi and P. Massart (2000). A sharp concentration inequality with applications. *Random Structures and Algorithms*, 16: 277–292.

O. Bousquet, V. Koltchinskii and D. Panchenko (2002). Some local measures of complexity on convex hulls and generalization bounds. *Proceedings of the* 15*th Annual Conference on Computational Learning Theory*, ed. J. Kivinen and R. H. Sloan. Springer Lecture Notes in Artificial Intelligence 2375.

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

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

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

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

P. L Hammer, T. Ibaraki and U. N. Peled (1981). Threshold numbers and threshold completions. *Annals of Discrete Mathematics* 11: 125–145.

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

O. L. Mangasarian (1968). Multisurface method of pattern separation. *IEEE Transactions on Information Theory* IT-14 (6): 801–807.

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

M. Marchand, M. Golea and P. Ruján (1990). A convergence theorem for sequential learning in two-layer perceptrons. *Europhys. Lett.* 11: 487–492.

M. Marchand, M. Shah, J. Shawe-Taylor and M. Sokolova (2003). The Set Covering Machine with Data-Dependent Half-Spaces. Proceedings of the Twentieth International Conference on Machine Learning (ICML'2003), 520–527, Morgan Kaufmann, San Francisco CA.

S. Mendelson (2003). A few notes on Statistical Learning Theory. In *Advanced Lectures in Machine Learning,* (S. Mendelson, A. J. Smola Eds), LNCS 2600, 1-40, Springer.

S. K. Murthy, S.Kasif and S. Salzberg (1994). A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research* 2: 1–32.

A. Ngom, I. Stojmenović and J. Žunić (2003). On the number of multilinear partitions and the computing capacity of multiple-valued multiple-threshold perceptrons, IEEE Transactions on Neural Networks 14(3): 469–477.

Z. Obradović and I. Parberry (1994). Learning with discrete multivalued neurons. *Journal of Computer and System Sciences* 49: 375–390.

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

D. Pollard (1984). *Convergence of Stochastic Processes*. Springer-Verlag.

J. R. Quinlan and R. Rivest (1989). Inferring decision trees using the minimum description length principle. *Information and Computation* 80: 227–248.

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

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

J. Shawe-Taylor and N. Cristianini (1998). Data-Dependent Structural Risk Minimisation for Perceptron Decision Trees. Neurocolt Technical Report NC2-TR-1998-003.

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

M. Sokolova, M. Marchand, N. Japkowicz, and J. Shawe-Taylor (2003). The Decision List Machine. Advances in Neural Information Processing Systems 15, 921–928, MIT-Press, Cambridge, MA, USA.

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

G. Turán and F. Vatan (1997). 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, 414-423.

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

V. N. Vapnik (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York..

V. N. Vapnik (1998). *Statistical Learning Theory*, Wiley.

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

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

A. Zuev and L. I. Lipkin (1988). Estimating the efficiency of threshold representations of Boolean functions. *Cybernetics* 24: 713–723. (Translated from Kibernetika (Kiev), 6, 1988: 29–37.)