Robust Cutpoints in the Logical Analysis of Numerical Data

Martin Anthony

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

Joel Ratsaby

Electrical and Electronics Engineering Department, Ariel University Center of Samaria, Ariel 40700, ISRAEL

Abstract

Techniques for the logical analysis of binary data have successfully been applied to non-binary data which has been 'binarized' by means of cutpoints; see [8, 9]. In this paper, we analyse the predictive performance of such techniques and, in particular, we derive generalization error bounds that depend on how 'robust' the cutpoints are.

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

Email addresses: m.anthony@lse.ac.uk (Martin Anthony), ratsaby@ariel.ac.il (Joel Ratsaby)

Appears in Discrete Applied Mathematics, 2012

1. Logical analysis of numerical data

1.1. LAD for binary data using positive patterns

Logical analysis of data (LAD) comprises a set of techniques to produce a data classifier. In classical LAD for binary data, we have a set $D \subseteq \{0,1\}^d \times \{0,1\}$ of labeled *observations* (or data-points) (x, y). Here x is an observation and y the corresponding binary label. The set of observations is partitioned into a set $D^+ = \{x : (x, 1) \in D\}$ of *positive* observations (labeled 1) and a set $D^- = \{x : (x, 0) \in D\}$ of *negative* observations (labeled 0). We shall sometimes abuse notation slightly and talk of $x \in D$ when we should say $x \in D^+ \cup D^-$ (because, technically, D is a set of labelled observations). The aim is to find a function h of a particular, simple, type (called a hypothesis) which fits the observations well. In a sense, such a hypothesis 'explains' the given data well and it is to be hoped that it generalises well to other data points, so far unseen. That is, we should like it to be the case that for most $x \in \{0,1\}^d$ which are not in D, h classifies y correctly, something we will make more precise shortly. Obvious candidates for such hypotheses are those that are *extensions* of D, meaning that, for all $(x, b) \in D$, we have $h(x) = b$.

In the standard LAD method for binary data described in [11], a disjunctive normal form Boolean function (a DNF) is produced. First, a *support set* of variables is found. By this is meant a set $S = \{i_1, i_2, \ldots, i_s\}$ such that no positive observation takes the same values as a negative observation on all the coordinates $i_1, i_2, \ldots, i_s$. (So, the positive and negative observations are distinguishable as sets when projected onto the coordinates given by S.) If S is a support set then there is some extension of D which depends only on the Boolean literals $u_i, \bar{u}_i$ for $i \in S$. In the technique described in [11], a small support set is found by solving a set-covering problem derived from the data set D. Once a support set has been found, one then looks for *positive patterns*. A (pure) positive pattern is a conjunction of literals which is satisfied by at least one positive observation in D (in which case we say that the observation is *covered* by the pattern) but by no negative observation. We then take as hypothesis h the disjunction of a set of positive patterns. If these patterns together cover all positive observations, then h is an extension of D.

1.2. Binarization of numerical data

The standard LAD techniques apply when the data is binary, $D \subseteq \{0,1\}^d \times \{0,1\}$. But in many applications, one deals with numerical data, in which $D \subseteq \mathbb{R}^d \times \{0,1\}$. Extending the methods of LAD to numerical data has been extensively investigated and used in many applications; see [9], for instance. A key initial step in using LAD methods when the set of observations has real-valued attributes is to *binarize* the data, so that $D \subseteq \mathbb{R}^d \times \{0,1\}$ is converted into a binary dataset $D^* \subseteq \{0,1\}^n \times \{0,1\}$, where, generally, $n \geq d$. The standard way to do so is to use *cutpoints* for each attribute.

We shall suppose from now on that $D \subseteq [0,1]^n \times \{0,1\}$, so all attribute values are between 0 and 1: clearly, this can be achieved simply by normalising the values. We shall also assume that D contains no contradictions: that is, we do not, for any $x \in \mathbb{R}^d$, have both $(x,0)$ and $(x,1)$ in D. For each attribute (or coordinate) $j = 1, 2, \ldots, d$, let

$$u_1^{(j)} < u_2^{(j)} < \cdots < u_{k_j}^{(j)}$$

be all the distinct values of the jth attribute of the members of D. The *candidate cutpoints* for attribute (or dimension) j are the values

$$\beta_i^{(j)} = \frac{u_i^{(j)} + u_{i+1}^{(j)}}{2}$$

for $i = 1, \ldots, k_j - 1$. These are the numbers halfway between successive values of the attribute. For each $j = 1, 2, \ldots d$ and $i = 1, 2, \ldots k_j - 1$, and for each $x \in D$, we define $b_i^{(j)}(x)$ to be 1 if and only if $x_j \geq \beta_i^{(j)}$. Let x^* be the resulting binary vector

$$x^* = (b_1^{(1)}(x), \ldots, b_{k_1}^{(1)}(x), b_1^{(2)}(x), \ldots, b_{k_2}^{(2)}(x), \ldots, b_1^{(d)}(x), \ldots, b_{k_d}^{(d)}(x)) \in \{0,1\}^n$$

where $n = \sum_{j=1}^d k_j$. Then we have a 'binarized' version $D^* = \{(x^*, b) : (x,b) \in D\}$ of the dataset D and we could apply LAD techniques to this binary dataset. Consider the dataset D^*. Because of the way in which the cutpoints are chosen in order to form D^*, there will be some set of positive patterns such that the disjunction of those patterns is an extension of D^*. This is simply because no two elements $x \in D^+$ and $y \in D^-$ are mapped

onto the same binarized observation in D^*, so the sets of members of D^* corresponding to positive and negative observations in D are disjoint.

There are a number of ways, however, in which this binarization is non-optimal. One immediate observation, as noted in [9], is that there is no need to use $\beta_i^{(j)}$ unless there exist $x \in D^+, y \in D^-$ with $x_j = u_i^{(j)}$ and $y_j = u_{i+1}^{(j)}$, or vice versa. So such 'non-essential' cutpoints can be eliminated, reducing the dimensionality of the binarized dataset. But there may be further redundancy in the use of the remaining cutpoints. In [8, 9], the authors consider the problem of finding a minimal set of cutpoints such that the corresponding binarized dataset will have an extension. This problem is phrased as a set-covering problem, which has an efficient greedy approximation algorithm, yielding a near-minimal number of cutpoints. We will comment further on this later, in discussing a variant of their approach. The outcome is that, in practice, the set of candidate cutpoints is reduced significantly to a set of used (or chosen) cutpoints.

1.3. The hypotheses

Let us denote by $\mathcal{A}^{(j)}$ the (reduced) set of cutpoints used for attribute j, and suppose the members of $\mathcal{A}^{(j)}$ are

$$a_1^{(j)} < a_2^{(j)} < \cdots < a_{l_j}^{(j)}.$$

A typical binarized $x \in [0,1]^d$ will be $x^* \in \{0,1\}^n$ where

$$x^* = (b_1^{(1)}(x), b_2^{(1)}(x), \ldots, b_{l_1}^{(1)}(x), b_1^{(2)}(x), \ldots, b_{l_2}^{(2)}(x), \ldots, b_1^{(d)}(x), \ldots, b_{l_d}^{(d)}(x)),$$

where $b_i^{(j)}(x) = 1$ if and only if $x_j \geq a_i^{(j)}$. Let the Boolean literal $u_i^{(j)}$ be given by $\mathbb{I}[x_j \geq a_i^{(j)}]$, where $\mathbb{I}[P]$ has logical value 1 if P is true and value 0 otherwise. Then a positive pattern is a conjunction of some of the Boolean variables $u_i^{(j)}$. Evidently, since (by definition of $u_i^{(j)}$) it is the case that $u_i^{(j)} = 1$ implies $u_{i'}^{(j)} = 1$ for $i > i'$, and any j, it follows that a typical positive pattern can be written in terms of these Boolean variables as

$$\bigwedge_{j=1}^{d} u_{r_j}^{(j)} \bar{u}_{s_j}^{(j)},$$

where $s_j > r_j$. (Here, $\wedge$ denotes the Boolean conjunction, the 'and' operator.) Geometrically, this positive pattern is the indicator function of the 'box'

$$[a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times [a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times [a_{r_d}^{(d)}, a_{s_d}^{(d)}).$$

1.4. Robustness

The paper [9] raises the issue of numerical attribute values being too close to cutpoints, something that could be problematic if there is a chance of measurement errors. There, the suggestion is that the corresponding binary variable could, due its potential unreliability, simply be considered missing and that one might seek to use 'robust' patterns, which, in a sense, are not dependent on the missing attribute value. In this paper, we also consider the distance between cutpoints and the (original, numerical) observations, but for a different reason. Rather than regard proximity of an attribute value to a cutpoint as a situation in which the corresponding binary attribute is deleted or considered unreliable, we continue to use it. But the smallest of the distances between cutpoints and the corresponding attribute values of the observations enters into our upper bound on the generalization accuracy of the resulting classifier. Moreover, it does so in such a way that the bound is worse (larger) if this minimal distance is small. This provides further motivation (other than considerations of measurement error or noise) to avoid the use of cutpoints which would be too close to the numerical attribute values of the observations, and it also leads to a variant of the cutpoint selection algorithm briefly referred to above.

2. Assessing the accuracy of learning

2.1. The probabilistic framework

To model the effectiveness of LAD for numerical data, we deploy a form of the popular 'PAC' model of computational learning theory (see [4, 17, 7]). This assumes that the labeled observations $(x, b) \in D$ (where $x \in [0, 1]^d$ and $b \in \{0, 1\}$) have been generated randomly (perhaps from some larger set of data) according to some fixed (but unknown) probability distribution P on

$Z = [0,1]^d \times \{0,1\}$. (This includes, as a special case, the situation in which x is drawn according to a fixed distribution on $[0,1]^d$ and the label b is then given deterministically by $b = t(x)$ where t is some fixed function.) Thus, if there are m data points in D, we may regard the data set as a *sample* $\mathbf{s} = ((x_1, b_1), \ldots, (x_m, b_m)) \in Z^m$, drawn randomly according to the product probability distribution P^m. In general terms, suppose that H is a set of functions from $X = [0,1]^d$ to $\{0,1\}$. Given any function $h \in H$, we can measure how well h matches the sample $\mathbf{s}$ through its *sample error*

$$\mathrm{er}_{\mathbf{s}}(h) = \frac{1}{m} |\{i : h(x_i) \neq b_i\}|$$

(the proportion of points in the sample incorrectly classified by h). In particular, we might be interested in *consistent* hypotheses: those for which the sample error is zero. An appropriate measure of how well h would perform on further examples is its *error*,

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

the probability that a further randomly drawn labeled data point would be incorrectly classified by h.

Much effort has gone into obtaining high-probability bounds on $\mathrm{er}_P(h)$. A typical result would state that, for all $\delta \in (0,1)$, with probability at least $1 - \delta$, for all $h \in H$ which are consistent with $\mathbf{s}$, we have $\mathrm{er}_P(h) < \epsilon(m, \delta)$, 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 [18, 14, 13], in which case $\epsilon(m, \delta)$ would typically involve the growth function [18, 7, 17, 2].

Much emphasis has been placed in practical machine learning techniques, such as Support Vector Machines [12], on 'learning with a large margin'. (See, for instance [16, 2, 3, 15].) Broadly speaking, the rationale behind margin-based generalization error bounds is that if a classifier has managed to achieve a 'wide' separation between 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. Related work involving 'width' rather than 'margin' has also been carried out [5] and, similarly, shows that 'definitive' classification is desirable. In a similar vein,

we will show that if the chosen cutpoints are 'robust' with respect to the datapoints (in that no attribute of a data point is too close to a cutpoint value), good generalization error bounds follow.

2.2. The hypotheses, and robustness

We start by clarifying what the hypothesis class is. As discussed above, the general hypotheses we consider can all be represented in the form

$$h = \mathbb{I}[B_1 \cup B_2 \cdots \cup B_r],$$

where:

- $r \in \mathbb{N}$;
- $\mathbb{I}[S]$ denotes the indicator function of a set S;
- there is, for each j, a set $\mathcal{A}^{(j)}$ of cutpoints such that each B_s is a box of the form $I_1 \times I_2 \times \cdots \times I_d$, where, for each j, $I_j = [\alpha, \alpha')$ with α, α' being cutpoints in $\mathcal{A}^{(j)}$ (with the convention that this includes also $[0, \alpha')$ and $[\alpha, 1)$).

It will, in the proofs, be convenient to define the following restricted class of such hypotheses: For $N \in \mathbb{N}$ and $\ell = (l_1, l_2, \ldots, l_d) \in \mathbb{N}^d$, we define $H_N(\ell) = H_N(l_1, l_2, \ldots, l_d)$ to be the set of such hypotheses in which $|\mathcal{A}^{(j)}| = l_j$, for $j = 1, 2, \ldots, d$ and in which the number r of boxes satisfies $r \leq N$. (Thus, it is the set of hypotheses in which we use l_j cutpoints for attribute j and, with respect to these cutpoints, take h to be the indicator function of the union of at most N boxes.) Then, the set H of all possible hypotheses that we could obtain by applying LAD to numerical data through binarization, and the subsequent construction of a classifier that is a disjunction of pure positive patterns, is $H = \bigcup\{H_N(\ell) : N \in \mathbb{N}, \ell \in \mathbb{N}^d\}$.

We make the following definition.

Definition 2.1. *Suppose that $h \in H$ and that the cutpoint sets used by h are $\mathcal{A}^{(1)}, \ldots, \mathcal{A}^{(d)}$. For $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_d) \in (0,1)^d$, we say that for $(x, b) \in [0,1]^d \times \{0,1\}$, h is Γ-robust on (x, b) if:*

- $h(x) = b$ *(so, h correctly classifies the sample point); and*
- *for all $j = 1, 2, \ldots, d$ and for all $a^{(j)} \in \mathcal{A}^{(j)}$, $|x_j - a^{(j)}| \geq \gamma_j$.*

Here, x_j denotes the jth coordinate of x.

So, h is Γ-robust on (x, b) if it correctly classifies x and, moreover, any attribute (coordinate) x_j of x is distance at least γ_j from any of the cutpoints in $\mathcal{A}^{(j)}$. Geometrically, this implies that, for each j, x lies distance at least γ_j in dimension j from any of boundary of any of the boxes used to define h. In fact, it means more than simply this: not only is it distance at least γ_j in dimension j from the boundaries of the boxes of which h is the union; but it is distance at least γ_j from the boundaries of any box defined by the cutpoints. So, to be explicit, Γ-robustness with respect to (x, b) is a property of both the cutpoint sets $\{\mathcal{A}^{(j)} : 1 \leq j \leq d\}$ and the hypothesis: for each j, x is distance at least γ_j in dimension j from any of the planes defined by the cutpoint values in Aj; and, also, h classifies x correctly.

We also can measure robustness on a sample:

Definition 2.2. *For a sample $\mathbf{s} = ((x_1, b_1), \ldots, (x_m, b_m))$, and for $h \in H$, the Γ-robustness error $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h)$ of h on $\mathbf{s}$ is the proportion of labeled observations (x_i, b_i) in $\mathbf{s}$ such that h is* not *Γ-robust on (x_i, b_i). That is,*

$$\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}[h \text{ is not } \Gamma-\text{robust on}(x_i, b_i)].$$

We say simply that h is Γ-robust on the sample $\mathbf{s}$ if $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$.

3. A generalization error bound

3.1. The bound

We denote by H the set of all possible hypotheses that we could obtain by applying LAD to numerical data through binarization, and the subsequent

construction of a classifier that is a disjunction of pure positive patterns. That is, in the earlier notation,

$$H = \bigcup \{H_N(\ell) : N \in \mathbb{N}, \ell \in \mathbb{N}^d\}.$$

Our main result is as follows. Note that, as explained earlier, we may regard a randomly-drawn training sample $\mathbf{s}$ as being distributed according to the product probability measure P^m.

Theorem 3.1. *Suppose that h is any hypothesis in H. Suppose $\delta \in (0,1)$. Then, with probability at least $1-\delta$ (with respect to the product measure P^m, where P is the underlying probability measure describing the generation of labeled examples), a sample $\mathbf{s}$ is such that:*

- *for all $\gamma \in (0,1)$;*
- *for all $l_1, l_2, \ldots, l_d \in \mathbb{N}$;*
- *for all $N \in \mathbb{N}$;*
- *if $h \in H$ is the indicator function of the union of at most N boxes based on cutpoint sets $\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(d)}$, where $|\mathcal{A}^{(i)}| = l_i$ for $i = 1, 2, \ldots, d$,*

then, if h is Γ-robust on $\mathbf{s}$, the error, $\mathrm{er}_P(h)$ of h is less than

$$\frac{2}{m}\left(\sum_{j=1}^{d}(l_j+2)\log_2\left(\frac{16}{\gamma_j}\right) + N + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Clearly, there will be some maximal values of γ_j such that the hypothesis we produce will be Γ-robust. One can determine this after the hypothesis is produced, and use this in the above bound.

We mention that the bound given in Theorem 3.1 is generally better than that obtainable using a standard approach (see [18, 7]) based on growth function and VC-dimension.

4. Cutpoint selection

The generalization error bounds obtained here suggest that there is potentially some advantage to be gained by choosing cutpoints in such a way that the hypotheses formed from these cutpoints are Γ-robust for large values of γ_j ($j = 1, 2, \ldots, d$). Suppose that for all j and for all chosen cutpoints $a_i^{(j)}$, the minimum distance of this cutpoint to the jth-coordinate of any datapoint is at least γ_j. It will then follow that any hypothesis constructed from these cutpoints will, if it is consistent with the sample be, furthermore, Γ-robust. (See Subsection 1.2 for a reminder of how the candidate cutpoints are determined.) We can therefore attach to each candidate cutpoint $\beta_i^{(j)}$ itself (rather than to a hypothesis) a measure of robustness: we define

$$\rho_D(\beta_i^{(j)}) = \min_{x \in D} |x_j - \beta_i^{(j)}|,$$

where x_j denotes the jth coordinate of x. It will be convenient to denote this by $\rho_i^{(j)}$.

Of course, since the bound of Theorem 3.1 depends not only on Γ but on the numbers, l_j, of cutpoints in each dimension, there would appear to be a need to have a trade-off between the robustness and the number of the selected cutpoints. Ideally, we would like to obtain a small number of highly-robust cutpoints in each dimension, but that may not be possible: we may be wise to choose a more than minimal number of cutpoints in order to avoid those with low robustness, for instance.

The paper [8] explains how the problem of finding a minimal sufficient number of cutpoints can be phrased as a set-covering problem. In [9], a greedy algorithm for choosing a reasonably small set of cutpoints is proposed. (By the performance guarantee for the greedy set-covering algorithm, the number of cutpoints is provably within a reasonable multiple of the minimum possible number.) In view of Theorem 3.1, which indicates that robust cutpoints are good, we can modify this algorithm to take this into account. As in [8], we start by using all possible cutpoints to binarize the data. That is, we start with the binarization obtained by using all the $\beta_i^{(j)}$, the midpoints in each dimension between the projections of the data points. Suppose that, with this particular binarization, as in Subsection 1.2, the binarized data set is

D^* (which has dimension n), and that $(D^+)^*$ and $(D^-)^*$ are, respectively, the binary vectors corresponding to positive and negative observations. For each pair (z, w) where $z \in (D^+)^*$ and $w \in (D^-)^*$, for $1 \leq j \leq d$, and, for each j, for i between 1 and k_j, let $a_{j,i}^{(z,w)}$ be the $\{0,1\}$-variable which is 1 if and only if z and w have a different value (0 or 1) in the coordinate that corresponds to the candidate cutpoint $\beta_i^{(j)}$ (which means that this cutpoint distinguishes between z and w). The problem of cutpoint minimization is to find the smallest set of cutpoints such that the corresponding binarization has the property that no positive and negative observation of the dataset have the same binarized vector. This can (as noted in [8]) be expressed as a set-covering problem. Expressed in terms of an integer program, the cutpoint minimization problem is:

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{d} \sum_{i=1}^{k_j} y_{j,i} \\ \text{subject to} \quad & \sum_{j=1}^{d} \sum_{i=1}^{k_j} a_{j,i}^{(z,w)} y_{j,i} \geq 1 \ \text{ for all } (z,w) \in (D^+)^* \times (D^-)^* \\ & y_{j,i} \in \{0,1\}, \ \ (1 \leq j \leq d, 1 \leq i \leq k_j). \end{aligned}$$

Here, $y_{j,i} = 1$ is to indicate that the corresponding cutpoint say $\beta_i^{(j)}$ is chosen. To formulate a variant of this, which takes into account the robustness of the cutpoints, let ϕ be a positive, decreasing, real function and consider the following *weighted* set-covering problem:

$$\begin{aligned} \text{minimize} \quad & \sum_{j=1}^{d} \sum_{i=1}^{k_j} \phi(\rho_i^{(j)}) y_{j,i} \\ \text{subject to} \quad & \sum_{j=1}^{d} \sum_{i=1}^{k_j} a_{j,i}^{(z,w)} y_{j,i} \geq 1 \ \text{ for all } (z,w) \in (D^+)^* \times (D^-)^* \\ & y_{j,i} \in \{0,1\}, \ \ (1 \leq j \leq d, 1 \leq i \leq k_j). \end{aligned}$$

For example, we could simply take $\phi(x) = 1/x$. Then the 'cost' of selecting cutpoint $\beta_i^{(j)}$ will be the inverse of its robustness. Or, motivated by the nature of the bound of Theorem 3.1, we might take $\phi(x) = \ln(1/x)$.

In [9], some variants of the standard set-covering formulation of cutpoint selection are suggested. One variant is to require that each binarized positive and negative observation be distinguished in more than one (say μ) coordinates, by requiring $\sum_{j=1}^{d}\sum_{i=1}^{k_j} a_{j,i}^{(z,w)} y_{j,i} \geq \mu$. Another is that the objective function $\sum_{j=1}^{d}\sum_{i=1}^{k_j} y_{j,i}$ be replaced by $\sum_{j=1}^{d}\sum_{i=1}^{k_j} u_{j,i} y_{j,i}$ where $u_{j,i}$ is some measure of the 'discriminating power' of the corresponding binary attribute. A third possible modification is to replace the constraints

$$\sum_{j=1}^{d}\sum_{i=1}^{k_j} a_{j,i}^{(z,w)} y_{j,i} \geq 1$$

by

$$\sum_{j=1}^{d}\sum_{i=1}^{k_j} a_{j,i}^{(z,w)} \lambda(z,w,i,j) y_{j,i} \geq 1,$$

where $\lambda(z,w,i,j)$ is some measure of how well the cutpoint $\beta_i^{(j)}$ separates the pair z, w. If we define $\lambda(z,w,i,j)$ to be simply $\rho_i^{(j)}$ (which does not explicitly depend on z and w, though it depends collectively on *set* of all (z,w)), then this third modification to the standard formulation is similar to that taken above where $\phi(x) = 1/x$ (but with the $y_{j,i}$ variables being 0 or $1/\rho_i^{(j)}$ rather than 0 or 1).

We can quickly obtain a good approximate solution to our modified cutpoint selection problem via the standard greedy algorithm for weighted set-covering. In this instance, the algorithm translates into the following cutpoint selection algorithm:

`Greedy Cutpoint Selection`

1. Initialize: $X := (D^+)^* \times (D^-)^*$, $\mathcal{A}^{(j)} := \emptyset$ $(j = 1, 2, \ldots, d)$.
2. Repeat steps (a)–(c) until $X = \emptyset$:
   (a) Choose (j,i) to maximize $\dfrac{|\{(z,w) \in X : a_{j,i}^{(z,w)} = 1\}|}{\phi(\rho_i^{(j)})}$
   (b) $\mathcal{A}^{(j)} := \mathcal{A}^{(j)} \cup \{\beta_i^{(j)}\}$
   (c) $X := X \setminus \{(z,w) : a_{j,i}^{(z,w)} = 1\}$
3. Output the cutpoint selection $\underline{\mathcal{A}} = (\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(d)})$.

5. Proof of the generalization error bound

We first derive a result in which the numbers of cutpoints in each dimension and the parameters γ_j are prescribed in advance, and in which the number of boxes involved in the hypothesis is bounded above by a prescribed number. We then remove the requirement that these parameters be fixed in advance, to obtain Theorem 3.1

For $\gamma \in (0,1)$, let $A_\gamma \subseteq [0,1]$ be the set of all integer multiples of $\gamma/2$ belonging to $[0,1]$, together with 1. So,

$$A_\gamma = \left\{0, \frac{\gamma}{2}, 2\frac{\gamma}{2}, 3\frac{\gamma}{2}, \ldots, \left\lfloor \frac{2}{\gamma} \right\rfloor \frac{\gamma}{2}, 1\right\}.$$

We have

$$|A_\gamma| \leq \left\lfloor \frac{2}{\gamma} \right\rfloor + 2 \leq \left\lfloor \frac{4}{\gamma} \right\rfloor.$$

For $\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(d)} \subseteq [0,1]$, let $\underline{\mathcal{A}} = (\mathcal{A}^{(1)}, \mathcal{A}^{(2)}, \ldots, \mathcal{A}^{(d)})$ and let $H_N(\underline{\mathcal{A}})$ be the set of hypotheses using cutpoints $\mathcal{A}^{(i)}$ for attribute i, and which are indicator functions of the unions of no more than N boxes defined by these cutpoints.

Let $N \in \mathbb{N}$ and $\ell = (l_1, l_2, \ldots, l_d) \in \mathbb{N}^d$ be fixed. Define $\hat{H}_N(\ell)$ as follows:

$$\hat{H}_N(\ell) = \bigcup \{H_N(\underline{\mathcal{A}}) : \mathcal{A}^{(j)} \subseteq A_{\gamma_j}, |\mathcal{A}^{(j)}| = l_j,\ j = 1, 2, \ldots, d\}.$$

So, $\hat{H}_N(\ell)$ is the set of hypotheses using, for attribute j, l_j cutpoints, all from A_{γ_j}, and involving no more than N boxes defined by these cutpoints.

We now bound the cardinality of the finite set $\hat{H}_N(\ell)$.

Lemma 5.1. *With the above notation, we have*

$$\log_2 \left|\hat{H}_N(\ell)\right| \leq \sum_{j=1}^{d} l_j \log_2 \left(\frac{4}{\gamma_j}\right) + 2N \sum_{j=1}^{d} \log_2(l_j + 2).$$

Proof: There are $\prod_{j=1}^{d} \binom{|A_{\gamma_j}|}{l_j}$ choices for the cutpoint sets $\mathcal{A}^{(j)}$. For each such choice, there are then no more than $\prod_{j=1}^{d} \binom{l_j+2}{2}$ possible ways of forming a box based on these cutpoints. (For, each side of the box is determined by an interval whose endpoints are either 0, 1, or cutpoints.) Thus the number of ways of choosing no more than N such boxes is bounded by $\prod_{j=1}^{d} \binom{l_j+2}{2}^N$. (This is slightly loose, since it counts the number of ordered selections, with repetition, of such boxes.)

It follows that

$$\begin{aligned}
\left|\hat{H}_N(\ell)\right| &\leq \prod_{j=1}^{d} \binom{l_j+2}{2}^N \prod_{j=1}^{d} \binom{|A_{\gamma_j}|}{l_j} \\
&\leq \prod_{j=1}^{d} (l_j+2)^{2N} \prod_{j=1}^{d} \binom{\lfloor 4/\gamma_j \rfloor}{l_j} \\
&\leq \prod_{j=1}^{d} (l_j+2)^{2N} \prod_{j=1}^{d} \left(\frac{4}{\gamma_j}\right)^{l_j}.
\end{aligned}$$

This gives the required result. □

Next, we use a 'symmetrization' technique (a general method that has its origins in the paper of Vapnik and Chervonenkis [18]).

Lemma 5.2. *If*

$$Q = \{\mathbf{s} \in Z^m : \exists h \in H_N(\ell) \text{ with } \mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0,\ \mathrm{er}_P(h) \geq \epsilon\}$$

and

$$T = \{(\mathbf{s}, \mathbf{s}') \in Z^m \times Z^m : \exists h \in H_N(\ell) \text{ with } \mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0,\ \mathrm{er}_{\mathbf{s}'}(h) \geq \epsilon/2\},$$

then, for $m \geq 8/\epsilon$, $P^m(Q) \leq 2\, P^{2m}(T)$.

Proof: We have

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

for $m \geq 8/\epsilon$. The final inequality follows from a well-known Chernoff bound [10], which is as follows: if $x_1, x_2, \ldots, x_m$ are independent $\{0,1\}$-valued random variables with $\mathrm{Prob}(x_i = 1) = p$, then the probability that $\sum_{i=1}^m x_i \leq (1-\alpha)pm$ is less than $\exp(-\alpha^2 pm/2)$. Now, suppose $\mathrm{er}_P(h) \geq \epsilon$. Then

$$\begin{aligned} \mathrm{Prob}(\mathrm{er}_{\mathbf{s}'}(h) < \epsilon/2) &\leq \mathrm{Prob}(m\mathrm{er}_{\mathbf{s}'}(h) < (1-1/2)\epsilon m) \\ &\leq \mathrm{Prob}(m\mathrm{er}_{\mathbf{s}'}(h) < (1-1/2)\mathrm{er}_P(h)m). \end{aligned}$$

Taking x_i to be 1 if h is erroneous on the ith entry of $\mathbf{s}'$, and 0 otherwise, we may apply the Chernoff bound to see that

$$\mathrm{Prob}(\mathrm{er}_{\mathbf{s}'}(h) < \epsilon/2) \leq \exp(-(1/2)^2 \mathrm{er}_P(h)m/2) \leq \exp(-\epsilon m/8),$$

and for $m \geq 8/\epsilon$, this is no more than $e^{-1} < 1/2$. So, for $m \geq 8/\epsilon$, $P^m(\mathrm{er}_{\mathbf{s}'}(h) \geq \epsilon/2) \geq 1/2$. □

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 $\sigma \in G$,

$$\sigma(z_1, z_2, \ldots, z_{2m}) = (z_{\sigma(1)}, \ldots, z_{\sigma(2m)}).$$

By invariance of P^{2m} under the action of G,

$$P^{2m}(T) \leq \max\{\mathbb{P}(\sigma\mathbf{z} \in T) : \mathbf{z} \in Z^{2m}\},$$

where $\mathbb{P}$ denotes the probability over uniform choice of σ from G. (See [18] and [2], for instance.)

Proposition 5.3. *Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, if $h \in H_N(\ell)$ and $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$, then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(\sum_{j=1}^{d} l_j \log_2\left(\frac{4}{\gamma_j}\right) + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Proof: Suppose $\sigma\mathbf{z} = (\mathbf{s}, \mathbf{s}') \in T$ and that $h \in H_N(\ell)$ is such that $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$ and $\mathrm{er}_{\mathbf{s}'}(h) \geq \epsilon/2$. The hypothesis h takes the form

$$h = \mathbb{I}[B_1 \cup B_2 \cdots \cup B_r],$$

where $r \leq N$ and each B_i is a box $I_1 \times I_2 \times \cdots \times I_d$. This means that, for each j, $I_j = [a_{r_j}^{(j)}, a_{s_j}^{(j)})$, for some $r_j, s_j \in \{0, 1, \ldots, l_j + 1\}$, where $\mathcal{A}^{(j)}$ comprises

$$a_1^{(j)} < a_2^{(j)} < \cdots < a_{l_j}^{(j)}$$

and where we interpret $a_0^{(j)} = 0$ and $a_{l_j+1}^{(j)} = 1$. (The indices r_j and s_j for box B_i also depend on i but we don't explicity denote this for the sake of notational ease.)

For every $a \in [0, 1]$, there is some $\hat{a} \in A_{\gamma_j}$ such that $|\hat{a} - a| < \gamma_j/2$. For a cutpoint $a^{(j)} \in \mathcal{A}^{(j)}$, let $\hat{a}^{(j)} \in A_{\gamma_j}$ be such that $|\hat{a}^{(j)} - a^{(j)}| < \gamma_j/2$. For each j, let $\hat{I}_j = [\hat{a}_{r_j}^{(j)}, \hat{a}_{s_j}^{(j)})$. (We define $\hat{I}_j = \emptyset$ if $\hat{a}_{r_j}^{(j)} \geq \hat{a}_{s_j}^{(j)}$.) Let $\hat{h} \in \hat{H}_N(\ell)$ be the hypothesis

$$\hat{h} = \mathbb{I}[\hat{B}_1 \cup \hat{B}_2 \cdots \cup \hat{B}_r],$$

where $\hat{B}_i = \hat{I}_1 \times \cdots \times \hat{I}_d$.

Suppose that h is Γ-robust on (x, b). Then $h(x) = b$ and, for all $j = 1, 2, \ldots, d$ and for all $a^{(j)} \in \mathcal{A}^{(j)}$, $|x_j - a^{(j)}| \geq \gamma_j$. It follows, since $|\hat{a}^{(j)} - a^{(j)}| < \gamma_j/2$, that $|x_j - \hat{a}^{(j)}| \geq \gamma_j/2$. So $\hat{h}$ is $\Gamma/2$-robust on (x, b). This argument shows that if $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$ then $\mathrm{er}_{\mathbf{s}}^{\Gamma/2}(\hat{h}) = 0$.

Suppose that $h(x) \neq b$. Then either $b = 0$ and a box containing x is included in the disjunction defining h; or $b = 1$ and no box included in h contains x. The hypothesis $\hat{h}$ is the disjunction of the boxes $\hat{B}_i$. Of course, if $\hat{h}(x) = h(x)$, then $\hat{h}(x) \neq b$ and $\hat{h}$ makes a misclassification. It could, however, be the case that $\hat{h}(x) = b$. For that to happen, it must be the case that either: (i) x belongs to some B_i but to no $\hat{B}_j$; or, (ii), x does not belong to any B_i but does belong to some $\hat{B}_j$. So, for some j, there is $a^{(j)} \in \mathcal{A}^{(j)}$ for which $a^{(j)} - x_j$ and $\hat{a}^{(j)} - x_j$ have opposite sign. This implies that x_j lies between the numbers $a^{(j)}$ and $\hat{a}^{(j)}$. But because $|\hat{a}^{(j)} - a^{(j)}| < \gamma_j/2$, we must therefore have $|x_j - \hat{a}^{(j)}| < \gamma_j/2$. So, although, in this situation, $\hat{h}$ would classify (x, b) correctly, it would not be $\Gamma/2$-robust on (x, b). This argument shows that if $\mathrm{er}_{\mathbf{s}'}(h) \geq \epsilon/2$ then $\mathrm{er}_{\mathbf{s}'}^{\Gamma/2}(\hat{h}) \geq \epsilon/2$.

So, we see that if $\sigma\mathbf{z} \in T$ then there is some $\hat{h} \in \hat{H}_N(\ell)$ for which $\sigma\mathbf{z} \in T(\hat{h})$, where

$$T(\hat{h}) = \{(\mathbf{s}, \mathbf{s}') \in Z^m \times Z^m : \mathrm{er}_{\mathbf{s}}^{\Gamma/2}(\hat{h}) = 0,\, \mathrm{er}_{\mathbf{s}'}^{\Gamma/2}(\hat{h}) \geq \epsilon/2\}.$$

Now, suppose $T(\hat{h}) \neq \emptyset$, so that for some $\tau \in G$, $\tau\mathbf{z} = (\mathbf{s}, \mathbf{s}') \in T(\hat{h})$, meaning that $\mathrm{er}_{\mathbf{s}}^{\Gamma/2}(\hat{h}) = 0$ and $\mathrm{er}_{\mathbf{s}'}^{\Gamma/2}(\hat{h}) \geq \epsilon/2$. Then, by symmetry, $\mathbb{P}\left(\sigma\mathbf{z} \in T(\hat{h})\right) = \mathbb{P}\left(\sigma(\tau\mathbf{z}) \in T(\hat{h})\right)$. Suppose that $\mathrm{er}_{\mathbf{s}'}^{\Gamma/2}(\hat{h}) = r/m$, where $r \geq \epsilon m/2$ is the number of x_i in $\mathbf{s}'$ on which $\hat{h}$ is not $\Gamma/2$-robust. Then those permutations σ such that $\sigma(\tau z) \in T(\hat{h})$ are precisely those that 'swap' elements other than these r, and there are $2^{m-r} \leq 2^{m-\epsilon m/2}$ such σ. It follows that, for each fixed $\hat{h} \in \hat{H}_N(\ell)$,

$$\mathbb{P}\left(\sigma\mathbf{z} \in T(\hat{h})\right) \leq \frac{2^{m(1-\epsilon/2)}}{|G|} = 2^{-\epsilon m/2}.$$

We therefore have

$$\mathbb{P}(\sigma\mathbf{z} \in T) \leq \mathbb{P}\left(\sigma\mathbf{z} \in \bigcup_{\hat{h} \in \hat{H}_N(\ell)} T(\hat{h})\right) \leq \sum_{\hat{h} \in \hat{H}_N(\ell)} \mathbb{P}(\sigma\mathbf{z} \in T(\hat{h})) \leq |\hat{H}_N(\ell)|\, 2^{-\epsilon m/2}.$$

So,

$$P^m(Q) \leq 2\, P^{2m}(T) \leq 2\, |\hat{H}_N(\ell)|\, 2^{-\epsilon m/2}.$$

This is at most δ when

$$\epsilon = \frac{2}{m}\left(\log_2 |\hat{H}_N(\ell)| + \log_2\left(\frac{2}{\delta}\right)\right).$$

Thus, with probability at least $1 - \delta$, if $h \in H_N(\ell)$ and $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$, then

$$\begin{aligned} \mathrm{er}_P(h) &< \frac{2}{m}\left(\log_2 |\hat{H}_N(\ell)| + \log_2\left(\frac{2}{\delta}\right)\right) \\ &= \frac{2}{m}\left(\sum_{j=1}^{d} l_j \log_2\left(\frac{4}{\gamma_j}\right) + 2N \sum_{j=1}^{d} \log_2(l_j + 2) + \log_2\left(\frac{2}{\delta}\right)\right). \end{aligned}$$

□

Next, we use this to obtain a result in which Γ and the l_j are not prescribed in advance. We use the following result, which generalizes one from [6].

Lemma 5.4. *Suppose P is any probability measure, $d \in \mathbb{N}$, and that*

$$\{E(\Gamma^{(1)}, \Gamma^{(2)}, \delta) : \Gamma^{(1)}, \Gamma^{(2)} \in (0,1]^d, \delta \leq 1\}$$

is a set of events such that:

- *for all $\Gamma \in (0,1]^d$, $P(E(\Gamma, \Gamma, \delta)) \leq \delta$,*
- *$\Gamma^{(1)} \leq \Gamma \leq \Gamma^{(2)}$ (component-wise) and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\Gamma^{(1)}, \Gamma^{(2)}, \delta_1) \subseteq E(\Gamma, \Gamma, \delta)$.*

Then

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

for $0 < \delta < 1$.

Proof: To prove this, we note that

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

□

We can now complete the proof of Theorem 3.1.

For $\Gamma \in (0,1)^d$, let $\epsilon(m, \delta, \ell, N, \Gamma)$ denote the quantity

$$\frac{2}{m}\left(\sum_{j=1}^{d} l_j \log_2\left(\frac{4}{\gamma_j}\right) + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Taking E to be set

$$E(\Gamma^{(1)}, \Gamma^{(2)}, \delta) = \{\mathbf{s} : \exists h \in H_N(\ell) \text{ s.t } \mathrm{er}_{\mathbf{s}}^{\Gamma^{(2)}}(h) = 0,\ \mathrm{er}_P(h) \geq \epsilon(m, \delta, \ell, N, \Gamma^{(1)})\},$$

we can apply Lemma 5.4 to obtain the following: with probability at least $1-\delta$, *for any* $\Gamma \in (0,1)^d$, if $h \in H_N(\ell)$ satisfies $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$, then $\mathrm{er}_P(h)$ is less than

$$\frac{2}{m}\left(\sum_{j=1}^{d} l_j \log_2\left(\frac{8}{\gamma_j}\right) + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2}{\delta}\right) + d + \sum_{j=1}^{d}\log_2\left(\frac{1}{\gamma_j}\right)\right).$$

This is bounded above by

$$\frac{2}{m}\left(\sum_{j=1}^{d} (l_j+2) \log_2\left(\frac{8}{\gamma_j}\right) + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2}{\delta}\right)\right).$$

It follows, by replacing δ by $\delta/2^{(l_1+l+2+\cdots+l_d+N)}$ that, for each $\ell \in \mathbb{N}^d$, with probability only at most $\delta/2^{(l_1+l+2+\cdots+l_d+N)}$ can there be some $\Gamma \in (0,1)^d$ and some $h \in H_N(\ell)$ such that $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$ and $\mathrm{er}_P(h) \geq \epsilon' = \epsilon'(m, \delta, \ell, N, \Gamma)$ where

$$\epsilon' = \frac{2}{m}\left(\sum_{j=1}^{d} (l_j+2) \log_2\left(\frac{8}{\gamma_j}\right) + 2N\sum_{j=1}^{d}\log_2(l_j+2) + \log_2\left(\frac{2\,2^{(l_1+l_2+\cdots+l_d+N)}}{\delta}\right)\right).$$

So, the probability that there is some Γ and some $h \in H = \bigcup\{H_N(\ell) : N \in \mathbb{N}, \ell \in \mathbb{N}^d\}$ with $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h) = 0$ and

$$\mathrm{er}_P(h) \geq \epsilon'(m, \delta, \ell, N, \Gamma)$$

is no more than

$$\begin{aligned}\sum_{N\in\mathbb{N}}\sum_{\ell\in\mathbb{N}^d}\frac{\delta}{2^{l_1+l_2+\cdots+l_d+N}} &= \sum_{N=1}^{\infty}\frac{\delta}{2^N}\sum_{l_1,l_2,\ldots,l_d=1}^{\infty}\prod_{j=1}^{d}\frac{1}{2^{l_j}}\\ &= \sum_{N=1}^{\infty}\frac{\delta}{2^N}\prod_{j=1}^{d}\sum_{l_j=1}^{\infty}\frac{1}{2^{l_j}}\\ &= \sum_{N=1}^{\infty}\frac{\delta}{2^N}\prod_{j=1}^{d}1=\delta.\end{aligned}$$

So, we have: with probability at least $1-\delta$, for all Γ, if $h\in H=\bigcup\{H_N(\ell):N\in\mathbb{N},\ell\in\mathbb{N}^d\}$ and $\mathrm{er}_{\mathbf{s}}^{\Gamma}(h)=0$, then

$$\begin{aligned}\mathrm{er}_P(h) &< \epsilon'=\frac{2}{m}\left(\sum_{j=1}^{d}(l_j+2)\log_2\left(\frac{8}{\gamma_j}\right)+2N\sum_{j=1}^{d}\log_2(l_j+2)+\log_2\left(\frac{2\,2^{(l_1+l_2+\cdots+l_d+N)}}{\delta}\right)\right)\\ &\leq \frac{2}{m}\left(\sum_{j=1}^{d}(l_j+2)\log_2\left(\frac{16}{\gamma_j}\right)+N+2N\sum_{j=1}^{d}\log_2(l_j+2)+\log_2\left(\frac{2}{\delta}\right)\right),\end{aligned}$$

and we obtain the result of the theorem.

6. Conclusions

In this paper, we have investigated some standard techniques for the logical analysis of numerical data, with the aim of quantifying the predictive accuracy of such methods in a probabilistic model of machine learning. In particular, we have obtained results which involve the 'robustness' of the cutpoints chosen to 'binarize' the data. These bounds suggest that it is advantageous to minimize a combination of the number of cutpoints and their robustness (and not simply the number of cutpoints). This, in turn, suggests a modification of the greedy cutpoint selection procedure to take robustness into account.

Acknowledgements

This work was supported by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. Thanks to Endre Boros and RUTCOR seminar participants for their helpful suggestions.

References

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

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

[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 (1992). *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science, 30. Cambridge University Press, Cambridge, UK.

[5] M. Anthony and J. Ratsaby. Maximal width learning of binary functions. *Theoretical Computer Science*, 411, 2010: 138–147.

[6] 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.

[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] Endre Boros, Peter L. Hammer, Toshihide Ibaraki and Alexander Kogan. Logical analysis of numerical data. *Mathematical Programming* 79 (1997): 163–190.

[9] Endre Boros, Peter L. Hammer, Toshihide Ibaraki, Alexander Kogan, Eddy Mayoraz and Ilya Muchnik. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering* 12 (2) (2000): 292–306.

[10] H. Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics* 23, 1952: 493–509.

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

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

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

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

[15] 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.

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

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

[18] 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.