The Performance of a New Hybrid Classifier Based on Boxes and Nearest Neighbors

Martin Anthony
Department of Mathematics
London School of Economics and Political Science
Houghton Street, London WC2A2AE
UK

Joel Ratsaby
Electrical and Electronics Engineering Department
Ariel University Center of Samaria
Ariel 40700
Israel

Abstract

In this paper we present a new type of binary classifier defined on the unit cube. This classifier combines some of the aspects of the standard methods that have been used in the logical analysis of data (LAD) and geometric classifiers, with a nearest-neighbor paradigm. We assess the predictive performance of the new classifier in learning from a sample, obtaining generalization error bounds that improve as the 'sample width' of the classifier increases.

1 Introduction

In this paper we introduce a new method of classifying points of $[0,1]^n$ into two classes. The classifiers we use combine the use of 'boxes' with a nearest-neighbor approach and for this reason we describe it as a *hybrid* classifier. Both classification by boxes and nearest-neighbor classification have been widely used. For instance, the use of boxes is integral to many of the standard methods used in the logical analysis of data (LAD); see (Boros et al. 1997) and (Boros et al. 2000), for instance.

The primary purpose of this paper is to quantify the performance of the hybrid classifiers by bounding their generalization error. In doing so, we obtain bounds that depend on a measure of how 'robust' the classification is on the training sample.

In order to describe the hybrid classifiers, and to place them in context, we first review the use of boxes in some of the standard LAD methods. (It should be noted that classification by unions of boxes have been more widely studied, not just in the context of LAD; see (Bshouty et al. 1998) for instance.) We then describe the new hybrid classifiers of interest here. Then, we describe how predictive performance of classifiers can be quantified in a probabilistic model of machine learning, and we derive generalization error bounds for the hybrid classifiers.

2 Classification using unions of boxes

In standard logical analysis of date (LAD) for binary data, we have a set $D \subseteq \{0,1\}^n \times \{0,1\}$ of labeled *observations* (or data-points, or training examples) (x,b). Here x is an observation and b 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). The aim is to find a classifier $h : \{0,1\}^n \to \{0,1\}$ which fits the observations well, and, it is hoped, generalizes well to other data points, so far unseen. In the standard LAD method for binary data described in (Crama, Hammer, and Ibaraki 1988), a disjunctive normal form Boolean function (a DNF) is produced. The terms of this DNF are termed *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. A more general technique combines the use of positive patterns with *negative* patterns, conjunctions which cover some negative observations. A point of $\{0,1\}^n$ is labeled as positive (that is, assigned value 1) if it is covered by at least one positive pattern, but no negative patterns; and it is labeled as negative (that is, assigned value 0) if it is covered by at least one negative pattern, but no positive patterns. If a point is covered by both types of pattern, then its classification is often determined by using a *discriminant*, which takes into account (perhaps in a weighted way) the number of positive and the number of negative patterns covering it.

These standard LAD techniques apply when the data is binary. However, many applications involve numerical data, in which $D \subseteq [0,1]^n \times \{0,1\}$. Extending the methods of LAD to numerical data has been investigated and used in many applications; see (Boros et al. 2000), for instance. A key initial step is to *binarize* the data, so that $D \subseteq [0,1]^n \times \{0,1\}$ is converted into a binary dataset $D^* \subseteq \{0,1\}^d \times \{0,1\}$, where, generally, $d \geq n$. The standard way to do so is to use *cutpoints* for each attribute (that is, for each of the n geometrical dimensions). For each attribute (or dimension) $j = 1, 2, \ldots, n$, 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. For each j, let

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

for $i = 1, \ldots, k_j - 1$. These are the numbers halfway between successive values of the attribute. For each $j =$

$1, 2, \ldots n$ 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

$$x^* = (b_1^{(1)}(x), \ldots, b_{k_1}^{(1)}(x), \ldots, b_1^{(n)}(x), \ldots, b_{k_n}^{(n)}(x)),$$

where $d = \sum_{j=1}^n k_j$. Then we have a 'binarized' version $D^* = \{(x^*, b) : (x, b) \in D\}$ of the dataset D and we could apply standard LAD techniques to this binary dataset. There are a number of ways, however, in which this binarization might be non-optimal and, usually, some cutpoints can be eliminated, reducing the dimensionality of the binarized dataset. In (Boros et al. 1997) and (Boros et al. 2000), the authors consider the problem of finding a minimal suitable set of cutpoints. This problem is phrased as a set-covering problem, which has an efficient greedy approximation algorithm, yielding a near-minimal number of cutpoints. In (Anthony and Ratsaby), variants on these approaches are discussed, the aim being to find 'robust' cutpoints; that is, cutpoints which define hyperplanes geometrically at least a certain distance from the data points.

Suppose, then, that a set $C^{(j)}$ of K_j cutpoints is selected for attribute j, and suppose the members of $C^{(j)}$ are

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

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

$$(b_1^{(1)}(x), \ldots, b_{K_1}^{(1)}(x), \ldots, b_1^{(n)}(x), \ldots, b_{K_n}^{(n)}(x)),$$

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 $u_i^{(j)}$) $u_i^{(j)} = 1$ 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

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

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'

$$[a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times [a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times [a_{r_n}^{(n)}, a_{s_n}^{(n)}).$$

With this approach, then, the simplest LAD classifier, which corresponds to a disjunctive normal form, is the indicator function of a union of boxes of this type; and all other regions of $[0, 1]^d$ are classified as negative. With the use also of negative patterns, we then have two separate unions of boxes: one labeled as positive, and the other negative. The other regions of $[0, 1]^d$ must also be classified and, as mentioned above, this is often done by a discriminator, the simplest approach being to classify a point as positive if and only if it is covered by at least as many positive patterns as negative patterns. This would, by default, classify as positive a point or a region which is covered by no patterns at all, of either type.

3 A hybrid classifier based on boxes and distance

3.1 Definition of the classifiers

The classifiers we introduce here are in many ways similar to those that result, as just described, from the use of positive and negative patterns in the logical analysis of numerical data. (But their use is not confined to LAD.) However, we combine the use of boxes with the use of a nearest-neighbor paradigm. Explicitly, if a point of $[0, 1]^n$ is not in the union of 'positive' boxes (the region covered by positive patterns) or in the union of 'negative' boxes, then it is not simply classified as positive; instead, we take into account the distance of the point from these two unions of boxes. If it is 'closer' to the positive boxes than the negative ones, we classify it as positive. We now describe the classifiers.

For each j between 1 and d, suppose there is a set $C^{(j)} = \{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\} \subseteq [0, 1]$. We call these the cutpoint sets. An open box (defined with respect to the cutpoint sets) is a set of the form

$$(a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)}),$$

where $0 \leq r_j < s_j \leq K_j + 1$ (and where $a_0^{(j)}$ is interpreted as 0 and $a_{K_j+1}^{(j)}$ as 1). Note that the 'sides' of the box in each dimension, j, are defined by two cutpoints from $C^{(j)}$ (or the end-points 0, 1) and that these need not be adjacent cutpoints. The cutpoint sets $C^{(j)}$ define $\prod_{j=1}^n \binom{K_j+2}{2}$ open boxes. Now take S_+ and S_- to be unions of some such boxes, in such a way that S_+ and S_- are disjoint. The boxes in S_+ are *positive* (labeled 1) and those in S_- negative (labeled 0); and, generally, there are unlabeled boxes, not in S_+ or S_-.

For a point $x \in X = [0, 1]^n$ let

$$\|x\| := \|x\|_\infty = \max_{1 \leq j \leq n} |x_j|$$

denote the max-norm of x. For two points $x, x' \in X$ the distance between them is $\|x - x'\|$ and for a set $S \subseteq X$ we define the distance from x to S to be $\operatorname{dist}(x, S) = \inf_{x' \in S} \|x - x'\|$. Clearly, for $x \in S$, $\operatorname{dist}(x, S) = 0$.

Given the pair S_+ and S^- of unions of open boxes, denote their closures by $\overline{S_+}$ and $\overline{S_-}$. (So, these are just the same unions of boxes, but with the boundaries included.) We define

$$f_+(x) = \operatorname{dist}\left(x, \overline{S_+}\right), \quad f_-(x) = \operatorname{dist}\left(x, \overline{S_-}\right)$$

and we let

$$f(x) = \frac{f_-(x) - f_+(x)}{2}. \tag{1}$$

So each pair (S_+, S_-) has a unique f associated with it.

Our classifiers will be the binary functions of the form

$$h(x) = \operatorname{sgn}(f(x)),$$

where $\operatorname{sgn}(z) = 1$ if $z \geq 0$ and $\operatorname{sgn}(z) = 0$ if $z < 0$. So, if F is the set of all real-valued functions of the form

$$f = \frac{f_- - f_+}{2},$$

where f_+ and f_- correspond to unions of boxes S_+ and S_-, then the set of classifiers is

$$H = \{\operatorname{sgn}(f) \mid f \in F\}$$

3.2 Example

Figure 1 shows a 2-dimensional example. We have five cutpoints in each dimension. The white boxes form S_+ and the black boxes form S_-. The grey region is the region which will be classified, in our method, using the distance to the boxes of each type (the 'nearest-neighbor' paradigm). When the whole domain is classified in the way described above, we obtain the partition indicated in Figure 2: the white region is labeled 1 and the black region 0.

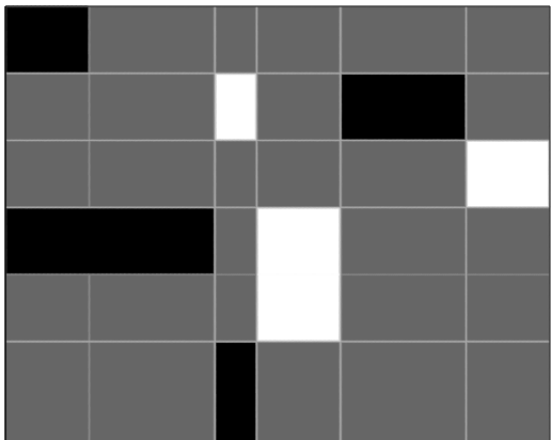

Figure 1: 'Before classification': the labeled boxes.

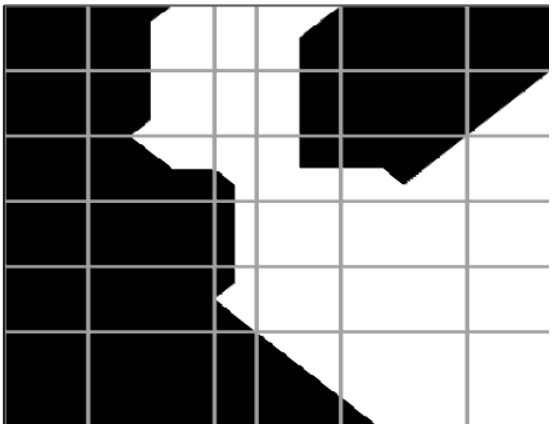

Figure 2: 'After classification': the classification of the whole domain.

3.3 Rationale

We are interested in this particular kind of classifier for several reasons. A special case of it corresponds quite naturally to a very simple and intuitive learning algorithm. Assume that the cutpoints have the property that we can find boxes defined by them, each of which contains only positive or only negative observations from the known data set. (The standard LAD algorithms for cutpoint selection guarantee this.) Then we could simply take S_+ to be the union of all boxes containing positive observations and S_- the union of those containing negative observations. Any other point x of the domain is then classified according to whether it is closer to the positive boxes or the negative ones.

Furthermore, these classifiers can be used in conjunction with LAD-type methods. One could run an LAD-type algorithm to produce positive and negative patterns. Each pattern corresponds to a box. Let R_+ be the union of the boxes defined by positive patterns and R_- the union of the boxes defined by negative patterns. The fact that some points (and indeed some box regions) could be covered by both positive and negative patterns means that R_+ and R_- need not be disjoint. The intersection $R_+ \cap R_-$ would itself be a union of boxes, and these could be classified, as in standard LAD procedures, using a discriminant. This would assign a box in the intersection a positive classification if and only if the number of positive patterns covering it (that is, boxes of R_+ containing it) is at least the number of negative patterns covering it (boxes of R_- containing it). The classification of these (sub-) boxes would then be resolved, and we could then form S_+ and S_- as the unions of the boxes now labeled 1 and 0, respectively. Then, any point not falling into $S_+ \cup S_-$ (that is, any point not covered by any pattern, positive or negative) is not simply classified as positive by default, but is classified according to whether it is closer to the positive region or the negative region.

Another attractive feature of the classifier produced is that it has a representation which, unlike 'black-box' classification schemes (for instance, based on neural networks), can be described and understood: there are box-shaped regions where we assert a known classification, and the classification anywhere else is determined by an arguably fairly sensible nearest-neighbor approach.

It is also useful that there is an underlying *real-valued* function f. This, as we will see, is useful in analyzing the performance of the classifier. Moreover, the *value* of f (not just its sign) has some geometrical significance. In particular, if $f(x)$ is relatively large, it means that x is quite far from boxes of the opposite classification: it is *not* the case that x is very near the boundary of a box which has the opposite classification. If all points of the data set satisfy this, then it means that the classification is, in a sense, 'robust'. (In related work in (Anthony and Ratsaby), a similar notion of robustness of the cutpoints for standard LAD methods is investigated and algorithms for selecting robust cutpoints are discussed.) We could interpret the value of the function f as an indicator of how confident we might be about the classification of a point: a point in the domain with a large value of f will be classified as positive, and more 'definitely' so than one with a smaller, but still positive, value of f. We might think that the classification of the first point is more reliable than that of the second, because the large value of f, indicating that the point is far from negative boxes, provides strong justification for a positive classification. For instance, consider again the example we've been studying. A contour plot of the function f is shown in Figure 3. The darkest regions are those with lowest (that is, most negative) values of f and the lightest are those with highest value of f. The very dark or very light regions are, arguably, those for which we

can most confidently classify the points. Figure 4 has some contour lines indicated, and Figure 5 is a colored version of Figure 3.

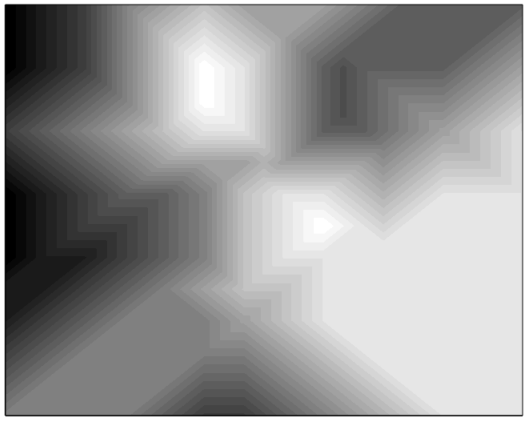

Figure 3: Contour plot of the underlying real-valued function f.

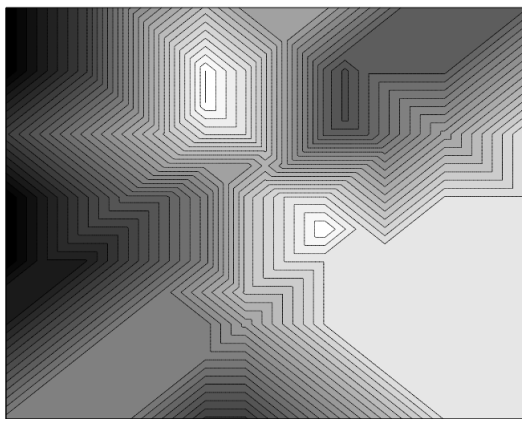

Figure 4: A contour plot of f, with contour lines idicated.

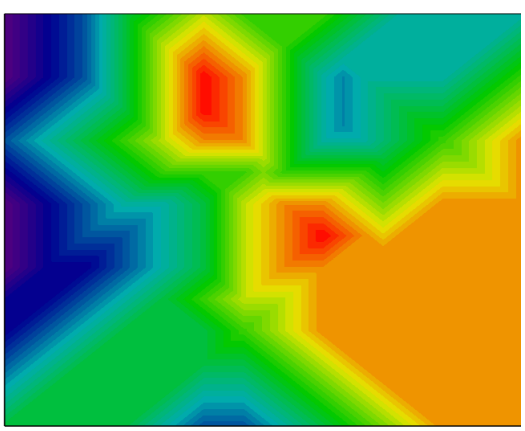

Figure 5: Colored contour plot of f.

4 Predictive performance of the classifier

4.1 Probabilistic modeling of learning

To quantify the performance of a classifier after training, we use a form of the popular 'PAC' model of computational learning theory (see (Anthony and Biggs 1992), (Vapnik 1998), (Blumer et al. 1989)). This assumes that we have some training examples $z_i = (x_i, b_i) \in Z = [0,1]^n \times \{0,1\}$, each of which has been generated randomly according to some fixed probability measure P on Z. These training examples are, in the LAD terminology, the positive and negative observations we are given at the outset. Then, we can regard a training sample of length m, which is an element of Z^m, as being randomly generated according to the product probability measure P^m. Suppose that F is the set of functions we are using to classify. (So, recall that F is a set of real-valued functions and that the corresponding binary classification functions are the functions $h = \mathrm{sgn}(f)$ for $f \in F$.

The natural way to measure the predictive accuracy of $f \in F$ in this context is by the probability that the sign of f correctly classifies future randomly drawn examples. We therefore use the following error measure of the classifier $h = \mathrm{sgn}(f)$:

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

Of course, we do not know this error: we only know how well the classifier performs on the training sample. We could quantify how well $f \in F$ matches the training sample by using the *sample error* of $h = \mathrm{sgn}(f)$:

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

(the proportion of points in the sample incorrectly classified by the sign of f). But we will find it more useful to use a variant of this, involving a 'width' or 'margin' parameter γ. Much emphasis has been placed in practical machine learning techniques, such as Support Vector Machines (Cristianini and Shawe-Taylor 2000), on 'learning with a large margin'. (See, for instance (Smola et al. 2000), (Anthony and Bartlett 1999), (Anthony and Bartlett 2000) and(Shawe-Taylor et al. 1996).) Related work involving 'width' (applicable to binary-valued rather than real-valued functions) rather than 'margin' has also been carried out (Anthony and Ratsaby 2010) and, similarly, shows that 'definitive' classification is desirable. If $h = \mathrm{sgn}(f)$, we define

$$\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = \frac{1}{m}|\{i : f(x_i)b_i < \gamma\}|.$$

This is the proportion of $z_i = (x_i, b_i)$ in the sample for which *either* $\mathrm{sgn}(f(x_i)) \neq b_i$, *or* $\mathrm{sgn}(f(x_i)) = b_i$ *but* $|f(x_i)| < \gamma$. So it is the fraction of the sample that is either misclassified by the classifier, or is correctly classified but *not definitively so*, in the sense that the value of $f(x_i)$ is *only just* of the right sign (and not correct 'with a margin' of at least γ).

Much effort has gone into obtaining high-probability bounds on $\mathrm{er}_P(h)$ in terms of $\mathrm{er}^{\gamma}_{\mathbf{z}}(f)$. A typical result would

be of the following form: for all $\delta \in (0,1)$, with probability at least $1-\delta$, for all $f \in F$,

$$\mathrm{er}_P(\mathrm{sgn}(f)) < \mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon(m,\gamma,\delta),$$

where ϵ decreases with m and δ. We obtain a bound of a similar, but slightly different form, in this paper for the set of hybrid classifiers we are considering.

4.2 Covering the set of classifiers

We now consider *covering numbers*, in order to deploy some results on probabilistic learning. Suppose that F is a set of functions from a domain X to some bounded subset Y of $\mathbb{R}$. For a finite subset S of X, the $l_\infty(S)$-norm is defined by $\|f\|_{l_\infty(S)} = \max_{x\in S}|f(x)|$. For $\gamma > 0$, a γ-cover of F with respect to $l_\infty(S)$ is a subset $\hat{F}$ of F with the property that for each $f \in F$ there exists $\hat{f} \in \hat{F}$ with the property that for all $x \in S$, $|f(x) - \hat{f}(x)| < \gamma$. The *covering number* $\mathcal{N}(F,\gamma,l_\infty(S))$ is the smallest cardinality of a covering for F with respect to $l_\infty(S)$ and the *uniform covering number* $\mathcal{N}_\infty(F,\gamma,m)$ is the maximum of $\mathcal{N}(F,\gamma,l_\infty(S))$, over all S with $S \subseteq X$ and $|S| = m$.

We will make use of the following result from (Anthony and Bartlett 1999). (It is a consequence of Theorem 13.7 there, which is a slight improvement of a result from (Bartlett 1998). More standard bounds do not have a factor of 2 in front of the $\mathrm{er}_{\mathbf{z}}^{\gamma}(f)$ but involve ϵ^2 rather than ϵ in the negative exponential. This type of bound is therefore potentially more useful when $\mathrm{er}_{\mathbf{z}}(f)$ is small.)

Theorem 4.1 *Suppose that F is a set of real-valued functions defined on a domain X and that P is any probability measure on $Z = X \times \{0,1\}$. Then, for any $\epsilon \in (0,1)$, any $\gamma > 0$ and any positive integer m,*

$$P^m\left(\{\mathbf{z} \in Z^m : \mathrm{er}_P(\mathrm{sgn}(f)) \geq 2\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon \text{ for some } f \in F\}\right)$$

is at most

$$\leq 4\mathcal{N}_\infty(F,\gamma/2,2m)e^{-\epsilon m/8}.$$

One approach to bounding the covering number of a function class F with respect to the $l_\infty(S)$-norm is to construct and bound the size of a covering with respect to the sup-norm $\|f\|_\infty$ on X. This clearly also serves as a covering with respect to $l_\infty(S)$, for any S, since if $\|f - \hat{f}\|_\infty < \gamma$ then, by definition of the sup-norm, $\sup_{x\in X}|f(x) - \hat{f}(x)| < \gamma$ and, hence, for all $x \in X$ (and, therefore, for all $x \in S$ where S is some subset of X), $|f(x) - \hat{f}(x)| < \gamma$. This is the approach we now take.

The following result will be useful to us.

Lemma 4.2 *Suppose f_+ is defined as above, and that it corresponds to the set S_+, a union of boxes based on cutpoints $a_i^{(j)}$ (for $1 \leq j \leq n$ and $1 \leq i \leq K_j$). Then, for any $x \in [0,1]^n$, there exists a pair of indices $1 \leq q \leq n$, $1 \leq p \leq K_q$ such that the distance between x and $\overline{S_+}$ satisfies* $\mathrm{dist}\left(x, \overline{S_+}\right) = \left|x_q - a_p^{(q)}\right|$.

Proof: We have

$$\mathrm{dist}\left(x, \overline{S_+}\right) = \inf_{x' \in \overline{S_+}} \phi\left(x, x'\right)$$

where, for each fixed x, $\phi(x,x') = \max_{1\leq j\leq n}|x_j - x'_j|$ is continuous in x'. Since the set $\overline{S_+}$ is closed, by the extreme-value theorem $\phi(x,x')$ attains its greatest lower bound on $\overline{S_+}$. If $x \in S_+$ then it is attained at $x' = x$. If $x \notin S_+$ then it is attained at some point on the boundary of S_+. This boundary consists of a union of 'sides', each side being a set of the form

$$V_i^{(j)} = \left\{ z \,\middle|\, z_j = a_i^{(j)},\ a_{r_k}^{(k)} \leq z_k \leq a_{s_k}^{(k)}, k \neq j \right\}.$$

A closest point z^* to x on $V_i^{(j)}$ is one each of whose coordinate values (on dimension k where $k \neq j$) either equals the corresponding coordinate value of x or equals a cutpoint value $a_{r_k}^{(k)}$ or $a_{s_k}^{(k)}$. Thus the distance $\|x - z^*\|$ either equals $\left|x_j - a_i^{(j)}\right|$; or $\left|x_k - a_k^r(k)\right|$ or $\left|x_k - a_{s_k}^{(k)}\right|$, for some $k \neq j$. The distance between x and $\overline{S_+}$ equals the minimal distance between x and any of the sides $V_i^{(j)}$. It follows that this distance equals $\left|x_s - a_p^{(q)}\right|$ for some $1 \leq q \leq n$, $1 \leq p \leq K_s$. □

We now bound the covering number of the set of classifiers that derive from at most some fixed number, B, of boxes, each of which is defined with respect to some set of K_j cutpoints in dimension j. To be specific, suppose $B \in \mathbb{N}$ and that, for each j between 1 and n, $K_j \in \mathbb{N}$. Let $\mathbf{k} = (K_1, K_2, \ldots, K_n) \in \mathbb{N}^n$ and let $F(\mathbf{k}, B)$ be the set of all the classifiers obtained as follows: (i) for each j, there is a set $C^{(j)} = \{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\} \subseteq [0,1]$; and, (ii) the boxes taken to form $S_+ \cup S_-$ are at most B in number, and each of the form

$$(a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)}).$$

where $0 \leq r_j < s_j \leq K_j + 1$ (and where $a_0^{(j)}$ is interpreted as 0 and $a_{K_j+1}^{(j)}$ as 1). (Note: we specify here the *numbers* K_j of cutpoints in each dimension, but we do not fix the *sets* of cutpoints. Note also that the boxes need not be disjoint.)

We have the following bound.

Theorem 4.3 *Let $B \in \mathbb{N}$ and $\mathbf{k} = (K_1, K_2, \ldots, K_n) \in \mathbb{N}^n$. Then, if $F(\mathbf{k}, B)$ is as just defined, we have*

$$\begin{aligned} &\ln \mathcal{N}_\infty(F(\mathbf{k},B),\gamma,m) \\ \leq\ & \sum_{j=1}^{n} K_j \ln\left(\frac{3}{\gamma}\right) + 2B\sum_{j=1}^{n}\ln(K_j+2) + B, \end{aligned}$$

for all $m \in \mathbb{N}$ and all $\gamma \in (0,1)$.

Proof: As indicated in the preceding discussion, we construct a covering of $F(\mathbf{k}, B)$ with respect to the sup-norm on $F(\mathbf{k}, B)$. Let $N = \lceil 1/\gamma \rceil$ and let

$$G_\gamma = \left\{0, \gamma, 2\gamma, \ldots, \left\lfloor \frac{1}{\gamma} \right\rfloor \gamma, 1\right\} \subseteq [0,1].$$

Note that $|G_\gamma| \leq \lfloor 1/\gamma \rfloor + 2 \leq \lfloor 3/\gamma \rfloor$. Let us define the class $\hat{F}(\mathbf{k}, B)$ of classifiers to be those satisfying: (i) for each j,

there are K_j cutpoints in dimension j, and each belongs to G_γ; and, (ii) the boxes taken to form $S_+ \cup S_-$ are at most B in number. Then we claim that $\hat{F}(\mathbf{k}, B)$ is a γ-covering of $F(\mathbf{k}, B)$ with respect to the sup-norm.

Given any $f \in F(\mathbf{k}, B)$ let $C^{(j)}$ be the set of cutpoints $\{a_1^{(j)}, a_2^{(j)}, \ldots, a_{K_j}^{(j)}\}$, for $1 \leq j \leq n$. For each $a_i^{(j)}$ there exists a corresponding $\hat{a}_i^{(j)} \in G_\gamma$ that satisfies $\left|a_i^{(j)} - \hat{a}_i^{(j)}\right| \leq \gamma$. For each box Q in S_+, where

$$Q = (a_{r_1}^{(1)}, a_{s_1}^{(1)}) \times (a_{r_2}^{(2)}, a_{s_2}^{(2)}) \times \cdots \times (a_{r_n}^{(n)}, a_{s_n}^{(n)}),$$

let $\hat{Q}$ be the box

$$\hat{Q} = (\hat{a}_{r_1}^{(1)}, \hat{a}_{s_1}^{(1)}) \times (\hat{a}_{r_2}^{(2)}, \hat{a}_{s_2}^{(2)}) \times \cdots \times (\hat{a}_{r_n}^{(n)}, \hat{a}_{s_n}^{(n)}).$$

Let $\hat{S}_+$ be the union of the boxes $\hat{Q}$ corresponding to the boxes Q forming S_+. In an analogous way, define $\hat{S}_-$. The function class $\hat{F}(\mathbf{k}, B)$ is precisely the set of all functions $\hat{f}$, defined by

$$\hat{f}(x) = \frac{\hat{f}_-(x) - \hat{f}_+(x)}{2},$$

where

$$\hat{f}_-(x) = \mathrm{dist}(x, \overline{\hat{S}_-}), \quad \hat{f}_+(x) = \mathrm{dist}(x, \overline{\hat{S}_+}).$$

We now show that $\|f - \hat{f}\|_\infty \leq \gamma$.

Fix any $x \in X$. Let us compute the values of $f_+(x)$ and $\hat{f}_+(x)$. From Lemma 4.2, there exist indices r, s such that $f_+(x) = \mathrm{dist}\left(x, \overline{S_+}\right) = \left|x_s - a_r^{(s)}\right|$. Denote by $\hat{a}_r^{(s)} \in G_s$ the cutpoint in G_γ that satisfies $\left|\hat{a}_r^{(s)} - a_r^{(s)}\right| \leq \gamma$. Then we have,

$$\begin{aligned} f_+(x) &= \left|x_s - a_r^{(s)}\right| \\ &\geq \left|x_s - \hat{a}_r^{(s)}\right| - \gamma \\ &\geq \inf\{\|x - z\| : z \in \overline{\hat{S}_+}\} - \gamma \\ &= \mathrm{dist}\left(x, \overline{\hat{S}_+}\right) - \gamma \\ &= \hat{f}_+(x) - \gamma. \end{aligned}$$

Also, from Lemma 4.2, there exist indices p, q such that $\hat{f}_+(x) = \mathrm{dist}\left(x, \overline{\hat{S}_+}\right) = \left|x_p - \hat{a}_q^{(p)}\right|$. Hence we have,

$$\begin{aligned} \hat{f}_+(x) &= \left|x_p - \hat{a}_q^{(p)}\right| \\ &\geq \left|x_p - a_q^{(p)}\right| - \gamma \\ &\geq \inf\{\|x - z\| : z \in \overline{S_+}\} - \gamma \\ &= \mathrm{dist}\left(x, \overline{S_+}\right) - \gamma \\ &= f_+(x) - \gamma. \end{aligned}$$

It follows that $\|f_+ - \hat{f}_+\| \leq \gamma$. The same argument holds for the pair f_- and $\hat{f}_-$, and so it follows that

$$\begin{aligned} &\|f - \hat{f}\|_\infty \\ &= \sup_{x \in X} \left|f(x) - \hat{f}(x)\right| \\ &= \frac{1}{2} \sup_{x \in X} \left|f_+(x) + f_-(x) - \hat{f}_+(x) - \hat{f}_-(x)\right| \\ &\leq \frac{1}{2} \sup_{x \in X} \left|f_+(x) - \hat{f}_+(x)\right| + \frac{1}{2} \sup_{x \in X} \left|f_-(x) - \hat{f}_-(x)\right| \\ &\leq \gamma. \end{aligned}$$

Thus for each $f \in F(\mathbf{k}, B)$ there exists $\hat{f} \in \hat{F}(\mathbf{k}, B)$ such that $\|f - \hat{f}\|_\infty \leq \gamma$, and $\hat{F}(\mathbf{k}, B)$ is therefore a γ-covering of $F(\mathbf{k}, B)$ in the sup-norm.

We now bound the cardinality of $\hat{F}(\mathbf{k}, B)$. Note that since there are K_j cutpoints in each dimension j, and each of these is from G_γ, a set of cardinality at most $\lfloor 3/\gamma \rfloor$, it follows that there are at most $\prod_{j=1}^n \binom{\lfloor 3/\gamma \rfloor}{K_j}$ possible ways of choosing he cutpoints for a function in $\hat{F}(\mathbf{k}, B)$. A box is defined by choosing a pair of cutpoints in each dimension (allowing also for the possibility that one end of the interval defining the box in any given dimension can be 0 or 1). We then choose B boxes, and, next, each box is assigned either a 0 label or a 1 label (that is, it is chosen to be part of $\hat{S}_-$ or $\hat{S}_+$). Thus, we have

$$\begin{aligned} |\hat{F}(\mathbf{k}, B)| &\leq \prod_{j=1}^n \binom{\lfloor 3/\gamma \rfloor}{K_j} \binom{\prod_{j=1}^n \binom{K_j+2}{2}}{B} 2^B \\ &\leq \prod_{j=1}^n \left\lfloor \frac{3}{\gamma} \right\rfloor^{K_j} \prod_{j=1}^n (K_j + 2)^{2B} 2^B. \end{aligned}$$

It follows, therefore, that $\ln \mathcal{N}_\infty(F, \gamma, m)$ is at most

$$|\hat{F}(\mathbf{k}, B)| \leq \sum_{j=1}^n K_j \ln\left(\frac{3}{\gamma}\right) + 2B \sum_{j=1}^n \ln(K_j + 2) + B.$$

□

4.3 A generalization error bound

Theorem 4.1, together with Theorem 4.3, could now be used to bound the generalization error of a classifier when B and $K_1, K_2, \ldots, K_n$ are prescribed in advance. However, the following more useful result does not require these to be known or prescribed.

Theorem 4.4 *Suppose $\delta \in (0, 1)$, and suppose P is any probability measure on $X = [0, 1]^n$. Then, with P^m-probability at least $1 - \delta$, a sample $\mathbf{z}$ is such that:*

- *for all $\gamma \in (0, 1)$;*
- *for all $K_1, K_2, \ldots, K_n \in \mathbb{N}$;*
- *for all $B \in \mathbb{N}$;*
- *if $f \in F(\mathbf{k}, B)$, then*

$$\mathrm{er}_P(\mathrm{sgn}(f)) \leq 2\,\mathrm{er}_{\mathbf{z}}^\gamma(f) + \epsilon(m, \gamma, \delta, \mathbf{k}, B),$$

where $\epsilon(m, \gamma, \delta, \mathbf{k}, B)$ is

$$\frac{8}{m}\left(\ln\left(\frac{12}{\gamma\delta}\right) + \sum_{j=1}^n K_j \ln\left(\frac{24}{\gamma}\right) + 2B + 2B\sum_{j=1}^n \ln(K_j + 2)\right).$$

Proof: Theorems 4.1 and 4.3 have the following immediate consequence: for $\mathbf{k} \in \mathbb{N}^n$ and $B \in \mathbb{N}$, with probability at least $1-\delta$, for all $f \in F(\mathbf{k}, B)$,

$$\mathrm{er}_P(f) < 2\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon_1(m, \gamma, \delta, \mathbf{k}, B)$$

where $\epsilon_1(m, \gamma, \delta, \mathbf{k}, B)$ is

$$\frac{8}{m}\left(\ln\left(\frac{4}{\delta}\right) + \sum_{j=1}^{n} K_j \ln\left(\frac{6}{\delta}\right) + B + 2B\sum_{j=1}^{n}\ln(K_j+2)\right).$$

For $\alpha_1, \alpha_2, \delta \in (0,1)$, let $E(\alpha_1, \alpha_2, \delta)$ be the set of $\mathbf{z} \in Z^m$ for which there exists some $f \in F(\mathbf{k}, B)$ with $\mathrm{er}_P(f) \geq 2\mathrm{er}_{\mathbf{z}}^{\alpha_2}(f) + \epsilon_1(m, \alpha_1, \delta, \mathbf{k}, B)$. Then, as just noted, $P^m(E(\alpha, \alpha, \delta)) \leq \delta$ and, also, if $\alpha_1 \leq \alpha \leq \alpha_2$ and $\delta_1 \leq \delta$, then $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$. It follows, from (Bartlett 1998) and (Anthony and Bartlett 1999), that

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

In other words, for fixed $\mathbf{k}$ and B, with probability at least $1-\delta$, *for all* $\gamma \in (0,1]$, we have

$$\mathrm{er}_P(f) < 2\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon_2(m, \gamma, \delta, \mathbf{k}, B),$$

where $\epsilon_2(m, \gamma, \delta, \mathbf{k}, B)$ is

$$\frac{8}{m}\left(\ln\left(\frac{4}{\delta}\right) + \sum_{j=1}^{n} K_j \ln\left(\frac{12}{\delta\gamma}\right) + B + 2B\sum_{j=1}^{n}\ln(K_j+2)\right).$$

(Note that γ now need not be prescribed in advance.) It now follows that the probability that for *some* $\mathbf{k}$ and for *some* B, we have

$$\mathrm{er}_P(f) \geq 2\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon_2\left(m, \gamma, \frac{\delta}{2^{(B+\sum_{j=1}^{n} K_j)}}, \mathbf{k}, B\right)$$

for *some* $\gamma \in (0,1)$ is at most

$$\begin{aligned}
\sum_{B,K_1,\ldots K_n=1}^{\infty} \frac{\delta}{2^{(B+\sum_{j=1}^{n} K_j)}} &= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\sum_{K_1,\ldots,K_n=1}^{\infty}\prod_{j=1}^{n}\frac{1}{2^{K_j}} \\
&= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\prod_{j=1}^{n}\sum_{K_j=1}^{\infty}\frac{1}{2^{K_j}} \\
&= \sum_{B=1}^{\infty}\frac{\delta}{2^B}\prod_{j=1}^{n} 1 = \delta.
\end{aligned}$$

The result now follows.

□

For any classifier of the type considered, there will be some maximal value of γ such that $\mathrm{er}_{\mathbf{z}}^{\gamma}(f) = 0$. We call this value of γ the *width* of f on $\mathbf{z}$. This terminology is motivated by the earlier observation that the value of $f(x)$ measures the distance from x to the nearest box with the opposite classification. (The term 'margin' might be more standard in the general context of using real-valued functions for classification, but 'width' seems more geometrically appropriate here.) Theorem 4.4 does not specify γ in advance, so we have the following immediate corollary.

Theorem 4.5 *With the same notation as above, with P^m-probability at least $1-\delta$, a sample $\mathbf{z}$ is such that for any $f \in F$, $\mathrm{er}_P(\mathrm{sgn}(f))$ is at most*

$$\frac{4}{m}\left(\ln\left(\frac{12}{\gamma(f,\mathbf{z})\delta}\right) + \sum_{j=1}^{n} K_j \ln\left(\frac{24}{\gamma(f,\mathbf{z})}\right) + 2B(f) + B(f)\sum_{j=1}^{n}\ln(K_j$$

where $\gamma(f, \mathbf{z})$ is the width of f on $\mathbf{z}$, and f involves $B(f)$ boxes, defined with respect to some set of K_j cutpoints in dimension j (for $1 \leq j \leq d$).

We could also use Theorem 4.4 as a guide to 'model selection'. The theorem states that, with probability at least $1-\delta$,

$$\begin{aligned}
\mathrm{er}_P(\mathrm{sgn}(f)) &< E(m, \gamma, \delta, \mathbf{k}, B) \\
&= 2\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon(m, \gamma, \delta, \mathbf{k}, B).
\end{aligned}$$

For fixed m and δ, $\epsilon(m, \gamma, \delta, \mathbf{k}, B)$ decreases as γ increases, and $\mathrm{er}_{\mathbf{z}}^{\gamma}(f)$ increases as γ increases. Therefore $E(m, \gamma, \delta, \mathbf{k}, B)$ is the sum of two quantities, one of which increases and one of which decreases as γ increases, and there is hence a trade-off between the two quantities. Clearly, also, the parameters $\mathbf{k}$ and B can be varied. This motivates the use of a learning algorithm that returns a classifier which minimizes the combination $E(m, \gamma, \delta, \mathbf{k}, B)$. The (high-probability) generalization error bound for such an algorithm take the form

$$\mathrm{er}_P(\mathrm{sgn}(f))$$

$$\leq \inf_{\gamma,\mathbf{k},B} \inf\{2\,\mathrm{er}_{\mathbf{z}}^{\gamma}(f) + \epsilon(m, \gamma, \delta, \mathbf{k}, B) : f \in F(\mathbf{k}, B)\}.$$

5 Conclusions and further work

This paper has introduced a new classifier, which is a hybrid between classical LAD (or unions of boxes) approaches and nearest-neighbor. We analyzed its generalization error, obtaining, in particular, error bounds that depend on a measure of the classifier's robustness, which we term its 'width'. In related ongoing work, we are addressing some of the algorithmic issues associated with this means of classification.

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886.

References

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

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

Anthony, M., and Biggs, N. 1992. *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press. Reprinted 1997.

Anthony, M., and Ratsaby, J. Robust cutpoints in the logical analysis of numerical data. *Discrete Applied Mathematics* to appear.

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

Bartlett, P. L. 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.

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

Boros, E.; Hammer, P. L.; Ibaraki, T.; and Kogan, A. 1997. Logical analysis of numerical data. *Mathematical Programming* 79:163–190.

Boros, E.; Hammer, P. L.; Ibaraki, T.; Kogan, A.; Mayoraz, E.; and Muchnik, I. 2000. An implementation of logical analysis of data. *IEEE Transactions on Knowledge and Data Engineering* 12(2):292–306.

Bshouty, N. H.; Goldberg, P.; Goldman, S.; and Mathias, H. 1998. Exact learning of discretized geometric concepts. *SIAM journal on Computing* 28(2):674–699.

Crama, Y.; Hammer, P.; and Ibaraki, T. 1988. Cause-effect relationships and partially defined boolean functions. *Annals of Operations Research* 16:299–325.

Cristianini, N., and Shawe-Taylor, J. 2000. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge: Cambridge University Press.

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

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

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