Analysis of a multi-category classifier

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

The use of boxes for pattern classification has been widespread and is a fairly natural way in which to partition data into different classes or categories. In this paper we consider multi-category classifiers which are based on unions of boxes. The classification method studied may be described as follows: find boxes such that all points in the region enclosed by each box are assumed to belong to the same category, and then classify remaining points by considering their distances to these boxes, assigning to a point the category of the nearest box. This extends the simple method of classifying by unions of boxes by incorporating a natural way (based on proximity) of classifying points outside the boxes. We analyse the generalization accuracy of such classifiers and we obtain generalization error bounds that depend on a measure of how definitive is the classification of training points.

Keywords: Multi-category classification, box clustering, generalization error.

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

Appears in Discrete Applied Mathematics, 2012

1. Box-based multi-category classifiers

Classification in which each category or class is a union of boxes is a long-studied and natural method for pattern classification. It is central, for instance, to the methods used for logical analysis of data (see, for example [9, 10, 15, 6]) and has been more widely studied as a geometrical classifier (see [11], for instance). More recently, unions of boxes have been used in combination with a nearest-neighbor (or proximity) paradigm for binary classification [5] and multi-category classification [13], enabling meaningful classification for points of the domain that lie outside any of the boxes.

In this paper, we analyse multi-category classifiers of the type decribed by Felici *et al.* [13]. In that paper, they describe a set of classifiers based on boxes and nearest-neighbor, where the metric used for the nearest-neighbor measure is the Manhattan (or taxicab) metric. (We give explicit details shortly.) They use an agglomerative box-clustering method to produce a set of candidate classifiers of this type. They then select from these one that is, in a sense they define, optimal. First they focus on the classifiers which are, with respect to the two dimensions of error on the sample, E, and complexity (number of boxes), B, Pareto-optimal. Among these they then select a classifier that minimizes an objective function of the form $(E - E_0)^2 + (B - B_0)^2$ (effecting a tradeoff between error and complexity) and, if there is more that one such classifier, they choose that which minimizes E. They provide some experimental evidence that this approach works. Here, we obtain generalization error bounds for the box-based classifiers of the type considered in [13], within a version of the standard PAC model of probabilistic learning. The bounds we obtain depend on the error and complexity and they improve (that is, they decrease) the more 'definite' is the classification of the sample points.

Suppose points of $[0, 1]^n$ are to be classified into C classes, which we will assume are labeled $1, 2, \ldots, C$. We let $[C]$ denote the set $\{1, 2, \ldots, C\}$.

A *box* (or, more exactly, an axis-parallel box) in $\mathbb{R}^n$ is a set of the form

$$\mathbf{I}(u, v) = \{x \in \mathbb{R}^n : u_i \le x_i \le v_i,\ 1 \le i \le n\},$$

where $u, v \in [0, 1]^n$ and $u \le v$, meaning that $u_i \le v_i$ for each i. We consider multi-category classifiers which are based on C unions of boxes, as we now describe. For $k = 1, \ldots, C$, suppose that S_k is a union of some number, B_k, of boxes:

$$S_k = \bigcup_{j=1}^{B_k} \mathbf{I}(u(k,j), v(k,j)).$$

Here, the jth box is defined by $u(k,j), v(k,j)$ where $u(k,j), v(k,j) \in [0,1]^n$ and $u(k,j) \leq v(k,j)$ (so, for each i between 1 and n, $u(k,j)_i \leq v(k,j)_i$). We assume, further, that for $k \neq l$, $S_k \cap S_l = \emptyset$. We think of S_k as being a region of the domain all of whose points we assume to belong to class k. So, as in [13] and [15], for instance, the boxes in S_k might be constructed by 'agglomerative' box-clustering methods.

To define our classifiers, we will make use of a metric on $[0,1]^n$. To be specific, as in [13], d will be the d_1 (or 'Manhattan' or 'taxicab') metric: for $x, y \in [0,1]^n$,

$$d(x,y) = \sum_{i=1}^{n} |x_i - y_i|.$$

We could equally well (as in [5], where the two-class case is the focus) use the supremum or d_∞ metric, defined by

$$d_\infty(x,y) = \max\{|x_i - y_i| : 1 \leq i \leq n\}$$

and similar results would be obtained. For $x \in [0,1]^n$ and $S \subseteq [0,1]^n$, the distance from x to S is

$$d(x,S) = \inf_{y \in S} d(x,y).$$

Let $\mathcal{S} = (S_1, S_2, \ldots, S_C)$ and denote by $h_{\mathcal{S}}$ the classifier from $[0,1]^n$ into $[C]$ defined as follows: for $x \in [0,1]^n$,

$$h_{\mathcal{S}}(x) = \text{argmin}_{1 \leq k \leq C} d(x, S_k),$$

where if $d(x, S_k)$ is minimized for more than one value of k, one of these is chosen randomly as the value of $h_{\mathcal{S}}$. So, in other words, the class label assigned to x is k where S_k is the closest to x of the regions $S_1, S_2, \ldots, S_C$. We refer to $B = B_1 + \cdots + B_C$ as the number of boxes in $\mathcal{S}$ and in $h_{\mathcal{S}}$. We will denote by H_B the set of all such classifiers where the number of boxes is B. The set of all possible classifiers we consider is then $H = \bigcup_{B=1}^{\infty} H_B$.

These classifiers, therefore, are based, as a starting point, on regions assumed to be of particular categories. These regions are each unions of boxes, and the regions do not overlap. (In practice, these boxes and the corresponding regions will likely have been constructed directly from a training sample by finding boxes containing sample points of a particular class, and merging, or agglomerating these; see [13].) See, for example, Figure 1. The three types of boxes are indicated, and the pale gray region is the region not covered by any of the boxes.

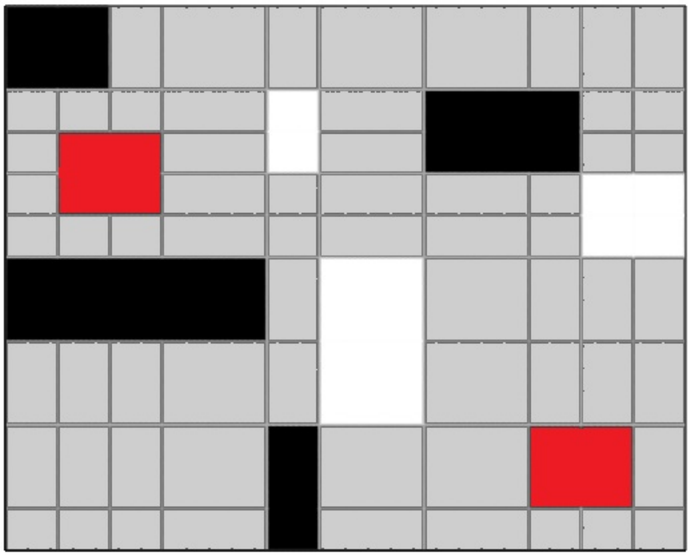

Figure 1: Boxes of three categories.

Then, for all other points of the domain, the classification of a point is given by the class of the region to which it is closest (in the d_1 metric). For the initial configuration of boxes indicated in Figure 1, the final classification of the whole domain is as indicated in Figure 2. Bounding lines for the boxes have been inserted in these figures to make it easier to see the correspondence between them.

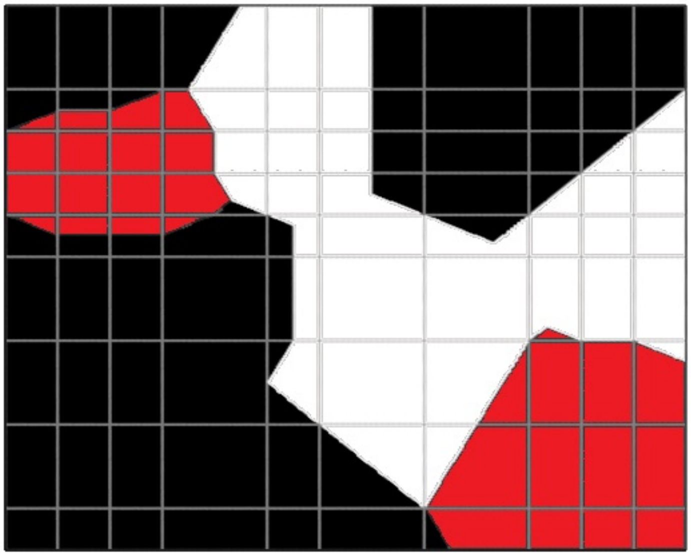

Figure 2: Classification of the whole domain resulting from the boxes of Figure 1.
,

These classifiers seem quite natural, from a geometrical point of view; and unlike 'black-box' classifiers (such as neural networks), can be described and understood: there are box-shaped regions where we assert a known classification, and the classification elsewhere is determined by an arguably fairly sensible nearest-neighbor approach.

It is also potentially useful that the classification is explicitly based on distance, and so there is a *real-valued* function f underlying the classifier: if the classification of x is k, then we can consider how much further x is from the next-nearest category of box. This real number, $f(x) = \min_{l \neq k} d(x, S_l) - d(x, S_k)$, quantifies, in a sense, how sure or definitive the classification is. In particular, if $f(x)$ is relatively large, it means that x is quite far from boxes of the other categories. 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 more 'definitely' than one with a small value of f and we might think that the classification of the first point is more reliable than that of the second, because the larger value of f indicates that the first point is further from boxes of other categories than is the second point. Furthermore, by considering the value of f *on the sample points*, we have some measure of how ‘robust’ the classifier is on the training sample. This measure of robustness plays a role in our generalization error bounds. In earlier work [6], we considered classical LAD methods, and analysed their performance in terms of a related measure of robustness. But that paper analysed only the standard LAD methods, in which (to describe it in terms of boxes) the classification of any point that was not contained in a box of some category would be determined randomly, rather than by means of a nearest neighbor paradigm.

2. Generalization error bounds for definitive classification

Guermeur [14] describes a fairly general framework for PAC-analysis of multi-category classifiers and we will show how we can use his results to bound the generalization error for our classifiers. This involves defining a certain function class and then bounding the covering numbers of that class. First, however, we obtain a tighter bound for the case in which the ‘margin error’ is zero. What this means is that we bound the error of the classifier in terms of how definitive the classification of the training sample is.

2.1. Classifiers with definitive classification on all sample points

The following definition describes what we mean by definitive classification on the sample.

Definition 2.1. *Let $Z = X \times Y$ where $X = [0,1]^n$ and $Y = [C] = \{1, 2, \ldots, C\}$. For $\gamma \in (0,1)$, and for $\mathbf{z} \in Z^m$, we say that $h_{\mathcal{S}} \in H$ achieves margin γ on the sample point $z = (x, y)$ in $\mathbf{z}$ if for all $l \neq y$, $d(x, S_l) > d(x, S_y) + \gamma$. We say that $h_{\mathcal{S}}$ achieves margin γ on the sample $\mathbf{z}$ if it achieves margin γ on each of the (x_i, y_i) in $\mathbf{z}$.*

So, $h_{\mathcal{S}}$ achieves margin γ on a sample point if S_y is the closest of the S_k to x (so that the class label assigned to x will be y) and, also, every other S_l

has distance from x that is at least γ greater than the distance from x to S_y. Note that we need only consider values of γ in the range $(0, n]$, since the maximum value of the Manhattan metric on $[0, 1]^n$ is n.

To quantify the performance of a classifier after training, we use a form of the 'PAC' model of computational learning theory. (See, for instance [8, 4, 20].)This assumes that we have training examples $z_i = (x_i, y_i) \in Z = [0, 1]^n \times [C]$, each of which has been generated randomly according to some fixed probability measure P on Z. (The sequence of z_i is i.i.d according to P.) 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.

The error of a classifier $h_{\mathcal{S}}$ is the probability that it does not definitely assign the correct class to a subsequent randomly-drawn instance (and we include in this the cases in which there are more than one equally close S_k, for the random choice then made might be incorrect). So, the error is the probability that it is *not* true that for $(x, y) \in Z$ we have $d(x, S_y) < d(x, S_k)$ for all $k \neq y$; that is,

$$\text{er}_P(h_{\mathcal{S}}) = P\left(\{(x, y) : d(x, S_y) \geq \min_{k \neq y} d(x, S_k)\}\right).$$

What we would hope is that, if a classifier performs well on a large enough sample, then its error is likely to be low. The following result is of this type (where good performance on the sample means correct, and definitively correct, classification on the sample).

Theorem 2.2. *Let $\delta \in (0, 1)$ and suppose P is a probability measure on $Z = [0, 1]^n \times [C]$. With P^m-probability at least $1 - \delta$, $\mathbf{z} \in Z^m$ will be such that we have the following: for all B and for all $\gamma \in (0, n]$, for all $h_{\mathcal{S}} \in H_B$, if $h_{\mathcal{S}}$ achieves margin γ on $\mathbf{z}$, then*

$$\text{er}_P(h_{\mathcal{S}}) < \frac{2}{m}\left(2nB \log_2\left(\frac{12n}{\gamma}\right) + 2B \log_2 C + \log_2\left(\frac{4n}{\delta\gamma}\right)\right).$$

In particular, Theorem 2.2 provides a high-probability guarantee on the real error of a classifier once training has taken place, based on the observed margin that has been obtained. (By the observed margin, we mean the largest value of γ such that a margin of γ has been achieved.) For this bound to be of use, the total number of boxes must, as a function of m, be sublinear, which has implications for the control of B during training.

2.2. Proof of Theorem 2.2

We first derive a result in which B and γ are fixed in advance. We then remove the requirement that these parameters be fixed, to obtain Theorem 2.2.

For $\gamma \in (0,1)$, let $A_\gamma \subseteq [0,1]$ be the set of all integer multiples of $\gamma/(4n)$ belonging to $[0,1]$, together with 1. So,

$$A_\gamma = \left\{0, \frac{\gamma}{4n}, 2\frac{\gamma}{4n}, 3\frac{\gamma}{4n}, \ldots, \left\lfloor \frac{4n}{\gamma} \right\rfloor \frac{\gamma}{4n}, 1\right\}.$$

We have

$$|A_\gamma| \leq \left\lfloor \frac{4n}{\gamma} \right\rfloor + 2 \leq \left\lfloor \frac{6n}{\gamma} \right\rfloor.$$

Let $\hat{H}_B \subseteq H_B$ be the set of all classifiers of the form $h_{\mathcal{S}}$ where, for each k, $\mathcal{S}_k$ is a union of boxes of the form $\mathbf{I}(u,v)$ where $u, v \in A_\gamma^n$.

Lemma 2.3. *With the above notation, we have*

$$\left|\hat{H}_B\right| \leq \left(\frac{6n}{\gamma}\right)^{2nB} C^B.$$

Proof: The number of possible boxes $\mathbf{I}(u,v)$ with $u, v \in A_\gamma^n$ is

$$\binom{|A_\gamma|}{2}^n \leq \left(\left\lfloor \frac{6n}{\gamma} \right\rfloor^2\right)^n \leq \left(\frac{6n}{\gamma}\right)^{2n}.$$

A classifier in $\hat{H}_B$ is obtained by choosing B such boxes and labeling each with a category between 1 and C. So,

$$|\hat{H}_B| \leq \binom{(6n/\gamma)^{2n}}{B} C^B \leq \left(\frac{6n}{\gamma}\right)^{2nB} C^B,$$

as required. □

For $\gamma \geq 0$, we define the γ-margin error of $h_{\mathcal{S}}$ on a sample $\mathbf{z}$. Denoted by $E^{\gamma}_{\mathbf{z}}(h_{\mathcal{S}})$, this is simply the proportion of (x, y) in $\mathbf{z}$ in which $h_{\mathcal{S}}$ does not achieve a margin of γ. In other words,

$$E^{\gamma}_{\mathbf{z}}(h_{\mathcal{S}}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I} \left\{ \exists l \neq y_i : d(x_i, S_l) \leq d(x, S_{y_i}) + \gamma \right\}.$$

Here, $\mathbb{I}A$ denotes the indicator function of a set (or event) A. Evidently, to say that $h_{\mathcal{S}}$ achieves margin γ on $\mathbf{z}$ is to say that $E^{\gamma}_{\mathbf{z}}(h_{\mathcal{S}}) = 0$.

The next part of the proof uses a 'symmetrization' technique similar to those first used in [21, 19, 12, 18] and in subsequent work extending those techniques to learning with real-valued functions, such as [16, 1, 3, 7].

Lemma 2.4. *If*

$$Q = \{\mathbf{s} \in Z^m : \exists h_{\mathcal{S}} \in H_B \text{ with } E^{\gamma}_{\mathbf{s}}(h_{\mathcal{S}}) = 0, \, \mathrm{er}_P(h_{\mathcal{S}}) \geq \epsilon\}$$

and

$$T = \{(\mathbf{s}, \mathbf{s}') \in Z^m \times Z^m : \exists h_{\mathcal{S}} \in H_B \text{ with } E^{\gamma}_{\mathbf{s}}(h_{\mathcal{S}}) = 0, \, E^{0}_{\mathbf{s}'}(h_{\mathcal{S}}) \geq \epsilon/2\},$$

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

Proof: We have

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

for $m \geq 8/\epsilon$. The final inequality follows from the fact that if $\mathrm{er}_P(h_{\mathcal{S}}) \geq \epsilon$, then for $m \geq 8/\epsilon$, $P^m(E^0_{\mathbf{s}'}(h_{\mathcal{S}}) \geq \epsilon/2) \geq 1/2$, for any $h_{\mathcal{S}} \in H_B$, something that follows by a Chernoff bound. □

We next bound the probability of T and use this to obtain a generalization error bound for fixed γ and B.

Proposition 2.5. *Let $B \in \mathbb{N}$, $\gamma \in (0, n]$ and $\delta \in (0, 1)$. Then, with probability at least $1 - \delta$, if $h_{\mathcal{S}} \in H_B$ and $E^\gamma_{\mathbf{s}}(h_{\mathcal{S}}) = 0$, then*

$$\mathrm{er}_P(h_{\mathcal{S}}) < \frac{2}{m}\left(2nB\log_2\left(\frac{6n}{\gamma}\right) + B\log_2 C + \log_2\left(\frac{2}{\delta}\right)\right).$$

Proof: 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 $\sigma \in G$,

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

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, for instance, [18, 2].)

Fix $\mathbf{z} \in Z^{2m}$. Suppose $\tau\mathbf{z} = (\mathbf{s}, \mathbf{s}') \in T$ and that $h_{\mathcal{S}} \in H_B$ is such that $E^\gamma_{\mathbf{s}}(h_{\mathcal{S}}) = 0$ and $E^0_{\mathbf{s}'}(h_{\mathcal{S}}) \geq \epsilon/2$. Suppose that $\mathcal{S} = (S_1, S_2, \ldots, S_C)$ where, for each k,

$$S_k = \bigcup_{j=1}^{B_k} \mathbf{I}(u(k,j), v(k,j)).$$

Let $\hat{u}(k,j), \hat{v}(j,k) \in A^n_\gamma$ be such that, for all r,

$$|\hat{u}(k,j)_r - u(k,j)_r| \leq \frac{\gamma}{4n}, \ \ |\hat{v}(k,j)_r - v(k,j)_r| \leq \frac{\gamma}{4n}.$$

These exist by definition of A_γ. Let

$$\hat{S}_k = \bigcup_{j=1}^{B_k} \mathbf{I}(\hat{u}(k,j), \hat{v}(k,j))$$

and $\hat{\mathcal{S}} = (\hat{S}_1, \ldots, \hat{S}_C)$. Let $\hat{h}_{\mathcal{S}}$ be the corresponding classifier in $\hat{H}_B$; that is, $\hat{h}_{\mathcal{S}} = h_{\hat{\mathcal{S}}}$. (Recall that $\hat{H}_B$ is the set of classifiers of the form $h_{\mathcal{S}}$ where, for each k, $\mathcal{S}_k$ is a union of boxes of type $\mathbf{I}(u,v)$ where $u, v \in A_\gamma^n$.)

The following geometrical fact (easily seen) will be useful: for any k, for any $a \in S_k$, there exists $\hat{a} \in \hat{S}_k$ such that $d(a, \hat{a}) \le \gamma/4$; and, conversely, for any $\hat{a} \in \hat{S}_k$, there exists $a \in S_k$ such that $d(a, \hat{a}) \le \gamma/4$. For any k, there is some $a \in S_k$ such that $d(x, S_k) = d(x, a)$. If $\hat{a} \in \hat{S}_k$ is such that $|a_r - \hat{a}_r| \le \gamma/(4n)$ for all r, then it follows that

$$d(a, \hat{a}) = \sum_{r=1}^{n} |a_r - \hat{a}_r| \le \gamma/4.$$

So,

$$d(x, \hat{S}_k) \le d(x, \hat{a}) \le d(x, a) + d(a, \hat{a}) \le d(x, S_k) + \gamma/4.$$

A similar argument shows that, for each k,

$$d(x, S_k) \le d(x, \hat{S}_k) + \gamma/4.$$

Suppose that, for all $l \ne y$, $d(x, S_l) \ge d(x, S_y) + \gamma$. Then, if $l \ne y$, we have

$$d(x, \hat{S}_l) \ge d(x, S_l) - \frac{\gamma}{4} \ge d(x, S_y) + \gamma - \frac{\gamma}{4} \ge d(x, \hat{S}_y) - \frac{\gamma}{4} + \gamma - \frac{\gamma}{4} = d(x, \hat{S}_y) + \frac{\gamma}{2}.$$

So, if $h_{\mathcal{S}}$ achieves margin γ on (x, y), then $\hat{h}_{\mathcal{S}}$ achieves margin $\gamma/2$. It follows that if $E_{\mathbf{s}}^{\gamma}(h_{\mathcal{S}}) = 0$ then $E_{\mathbf{s}}^{\gamma/2}(\hat{h}_{\mathcal{S}}) = 0$. Now suppose that there is $l \ne y$ such that $d(x, S_l) < d(x, S_y)$. Then

$$d(x, \hat{S}_l) \le d(x, S_l) + \frac{\gamma}{4} < d(x, S_y) + \frac{\gamma}{4} \le d(x, \hat{S}_y) + \frac{\gamma}{4} + \frac{\gamma}{4} = d(x, \hat{S}_y) + \frac{\gamma}{2}.$$

This argument shows that if $E_{\mathbf{s}'}^{0}(h_{\mathcal{S}}) \ge \epsilon/2$, then $E_{\mathbf{s}'}^{\gamma/2}(h_{\mathcal{S}}) \ge \epsilon/2$.

It now follows that if $\tau\mathbf{z} \in T$, then, for some $\hat{h}_{\mathcal{S}} \in H_B$, $\tau\mathbf{z} \in R(\hat{h}_{\mathcal{S}})$, where

$$R(\hat{h}_{\mathcal{S}}) = \{(\mathbf{s}, \mathbf{s}') \in Z^m \times Z^m : E_{\mathbf{s}}^{\gamma/2}(\hat{h}_{\mathcal{S}}) = 0,\ E_{\mathbf{s}'}^{\gamma/2}(\hat{h}_{\mathcal{S}}) \geq \epsilon/2\}.$$

By symmetry, $\mathbb{P}\left(\sigma\mathbf{z} \in R(\hat{h}_{\mathcal{S}})\right) = \mathbb{P}\left(\sigma(\tau\mathbf{z}) \in R(\hat{h}_{\mathcal{S}})\right)$. Suppose that $E_{\mathbf{s}'}^{\gamma/2}(\hat{h}_{\mathcal{S}}) = r/m$, where $r \geq \epsilon m/2$ is the number of (x_i, y_i) in $\mathbf{s}'$ on which $\hat{h}_{\mathcal{S}}$ does not achieve margin $\gamma/2$. Then those permutations σ such that $\sigma(\tau\mathbf{z}) \in R(\hat{h}_{\mathcal{S}})$ are precisely those that do not transpose these r coordinates, and there are $2^{m-r} \leq 2^{m-\epsilon m/2}$ such σ. It follows that, for each fixed $\hat{h}_{\mathcal{S}} \in \hat{H}_B$,

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

We therefore have

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

So,

$$P^m(Q) \leq 2\, P^{2m}(T) \leq 2\, |\hat{H}_B|\, 2^{-\epsilon m/2} \leq 2\left(\frac{6n}{\gamma}\right)^{2nB} C^B 2^{-\epsilon m/2}$$

This is at most δ when

$$\epsilon = \frac{2}{m}\left(2nB\log_2\left(\frac{6n}{\gamma}\right) + B\log_2 C + \log_2\left(\frac{2}{\delta}\right)\right),$$

as stated. □

Next, we use this to obtain a result in which γ and B are not prescribed in advance. For $\alpha_1, \alpha_2 \in (0, n]$ and $\delta \in (0, 1)$, let $E(\alpha_1, \alpha_2, \delta)$ be the set of $\mathbf{z} \in Z^m$ for which there exists some $h_{\mathcal{S}} \in H_B$ which achieves margin α_2 on $\mathbf{z}$ and which has $\text{er}_P(f) \geq \epsilon_1(m, \alpha_1, \delta, B)$, where

$$\epsilon_1(m, \alpha_1, \delta, B) = \frac{2}{m}\left(2nB\log_2\left(\frac{6n}{\alpha_1}\right) + B\log_2 C + \log_2\left(\frac{2}{\delta}\right)\right).$$

Proposition 2.5 tells us that $P^m(E(\alpha, \alpha, \delta)) \leq \delta$. It is also clear that 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 a slightly modified version of a result from [7], that

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

In other words, for fixed B, with probability at least $1-\delta$, *for all* $\gamma \in (0,1]$, we have that if $h_{\mathcal{S}} \in H_B$ achieves margin γ on the sample, then $\text{er}_P(h_{\mathcal{S}}) < \epsilon_2(m, \gamma, \delta, B)$, where

$$\epsilon_2(m, \gamma, \delta, B) = \frac{2}{m}\left(2nB\log_2\left(\frac{12n}{\gamma}\right) + B\log_2 C + \log_2\left(\frac{4n}{\delta\gamma}\right)\right).$$

Note that γ now need not be prescribed in advance. It now follows that with probability at least $1 - \delta/2^B$, for any $\gamma \in (0, n]$, if $h_{\mathcal{S}} \in H_B$ achieves margin γ, then $\text{er}_P(h_{\mathcal{S}}) \leq \epsilon_2(m, \gamma, \delta/2^B, B)$. So, with probability at least $1 - \sum_{B=1}^{\infty}(\delta/2^B) = 1 - \delta$, we have: for all B, for all $\gamma \in (0, 1)$, if $h_{\mathcal{S}} \in H_B$ achieves margin γ, then

$$\text{er}_P(h_{\mathcal{S}}) \leq \epsilon_2(m, \gamma, \delta/2^B, B) = \frac{2}{m}\left(2nB\log_2\left(\frac{12n}{\gamma}\right) + B\log_2 C + \log_2\left(\frac{4n}{\delta\gamma}\right) + B\right).$$

Theorem 2.2 now follows on noting that $\log_2 C \geq 1$.

2.3. Allowing less definitive classification on some sample points

As mentioned, Guermeur [14] has developed a fairly general framework in which to analyse multi-category classification, and we can apply one of his results to obtain a generalization error bound applicable to the case in which a margin of γ is not obtained on all the training examples. We describe his result and explain how to formulate our problem in his framework. This requires us to define a certain real-valued class of functions, the values of which may themselves impart some useful information as to how 'confident' one might be about classification. To use his result to obtain a generalization error bound, we then bound the covering numbers of this class of functions.

What we obtain is a high probability bound that takes the following form: for all B, for all $\gamma \in (0,1)$, if $h_{\mathcal{S}} \in H_B$, then

$$\mathrm{er}_P(h_{\mathcal{S}}) \leq E^{\gamma}_{\mathbf{z}}(h_{\mathcal{S}}) + \epsilon(m, \gamma, \delta, B),$$

where ϵ tends to 0 as $m \to \infty$ and ϵ decreases as γ increases. (Recall that $E^{\gamma}_{\mathbf{z}}(h_{\mathcal{S}})$ is the γ-margin error of $h_{\mathcal{S}}$ on the sample.) The rationale for seeking such a bound is that there is likely to be a trade-off between margin error on the sample and the value of ϵ: taking γ small so that the margin error term is zero might entail a large value of ϵ; and, conversely, choosing γ large will make ϵ relatively small, but lead to a large margin error term. So, in principle, since the value γ is free to be chosen, one could optimize the choice of γ on the right-hand side of the bound to minimize it.

We now describe Guermeur's framework (with some adjustments to the notation to make it consistent with the notation used here). There is a set $\mathcal{G}$ of functions from $X = [0,1]^n$ into $\mathbb{R}^C$, and a typical $g \in \mathcal{G}$ is represented by its component functions $g = (g_k)_{k=1}^C$. Each $g \in \mathcal{G}$ satisfies the constraint

$$\sum_{k=1}^{C} g_k(x) = 0, \quad \forall x \in X.$$

A function of this type acts as a classifier as follows: it assigns category $l \in [C]$ to $x \in X$ if and only if $g_l(x) > \max_{k \neq l} g_k(x)$. (If more than one value of k maximizes $g_k(x)$, then the classification is left undefined, assigned some value $*$ not in $[C]$.) The *risk* of $g \in \mathcal{G}$, when the underlying probability measure on $X \times Y$ is P, is defined to be

$$R(g) = P\left(\{(x,y) \in X \times [C] : g_y(x) \leq \max_{k \neq y} g_k(x)\}\right).$$

For $(v,k) \in \mathbb{R}^C \times [C]$, let $M(v,k) = \dfrac{1}{2}\left(v_k - \max\limits_{l \neq k} v_l\right)$ and, for $g \in \mathcal{G}$, let Δg be the function $X \to \mathbb{R}^C$ given by

$$\Delta g(x) = (\Delta g_k(x))_{k=1}^C = (M(g(x),k))_{k=1}^C.$$

Given a sample $\mathbf{z} \in (X \times [C])^m$, let

$$R_{\gamma,z}(g) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left\{\Delta g_{y_i}(x_i) < \gamma\right\}.$$

To describe Guermeur's result, we need covering numbers. Suppose $\mathcal{H}$ is a set of functions from X to $\mathbb{R}^C$. For $\mathbf{x} \in X^m$, define the metric $d_{\mathbf{x}}$ on $\mathcal{H}$ by

$$d_{\mathbf{x}}(h, h') = \max_{1 \le i \le m} \|h(x_i) - h'(x_i)\|_\infty = \max_{1 \le i \le m} \max_{1 \le k \le C} |h_k(x_i) - h'_k(x_i)|.$$

For $\alpha > 0$, a finite subset $\hat{\mathcal{H}}$ of $\mathcal{H}$, is said to be a (proper) α-cover of $\mathcal{H}$ (with respect to metric $d_{\mathbf{x}}$) if for each $h \in \mathcal{H}$ there exists $\hat{h} \in \hat{\mathcal{H}}$ such that $d_{\mathbf{x}}(h, \hat{h}) \le \alpha$. The class $\mathcal{H}$ is totally bounded if for each $\alpha > 0$, for each m, and for each $\mathbf{x} \in X^m$, there is a finite α-cover of $\mathcal{H}$ with respect to $d_{\mathbf{x}}$. The smallest cardinality of an α-cover of $\mathcal{H}$ with respect to $d_{\mathbf{x}}$ is denoted $\mathcal{N}(\alpha, \mathcal{H}, d_{\mathbf{x}})$. The *covering number* $\mathcal{N}(\alpha, \mathcal{H}, m)$ is defined by

$$\mathcal{N}(\alpha, \mathcal{H}, m) = \max_{\mathbf{x} \in X^m} \mathcal{N}(\alpha, \mathcal{H}, d_{\mathbf{x}}).$$

A result following from [14] is (in the above notation) as follows:

Theorem 2.6. *Let $\delta \in (0,1)$ and suppose P is a probability measure on $Z = [0,1]^n \times [C]$. With P^m-probability at least $1 - \delta$, $\mathbf{z} \in Z^m$ will be such that we have the following: for all $\gamma \in (0, n]$ and for all $g \in \mathcal{G}$,*

$$R(g) \le R_{\gamma,m}(g) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\gamma/4, \Delta\mathcal{G}, 2m) + \ln\left(\frac{2n}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

We now use Theorem 2.6, with an appropriate choice of function space $\mathcal{G}$. Let us fix $B \in \mathbb{N}$ and let $\mathcal{S} = (S_1, \ldots, S_C)$ be, as before, C unions of boxes, with B boxes in total. Let $g^{\mathcal{S}}$ be the function $X = [0,1]^n \to \mathbb{R}^C$ defined by $g^{\mathcal{S}} = (g_k^{\mathcal{S}})_{k=1}^C$, where

$$g_k^{\mathcal{S}}(x) = \frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_k).$$

Let $\mathcal{G}_B = \{g^{\mathcal{S}} : h_{\mathcal{S}} \in H_B\}$. Then these functions satisfy the constraint that their coordinate functions sum to the zero function, since

$$\sum_{k=1}^{C} g_k(x) = \sum_{k=1}^{C} \frac{1}{C} \sum_{i=1}^{C} d(x, S_i) - \sum_{k=1}^{C} d(x, S_k) = \sum_{k=1}^{C} d(x, S_k) - \sum_{k=1}^{C} d(x, S_k) = 0.$$

For each k,

$$\begin{aligned}
\Delta g_k^{\mathcal{S}}(x) &= M(g^{\mathcal{S}}(x), k) \\
&= \frac{1}{2}\left(g_k^{\mathcal{S}}(x) - \max_{l \neq k} g_l^{\mathcal{S}}(x)\right) \\
&= \frac{1}{2}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_k) - \max_{l \neq k}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_l)\right)\right) \\
&= \frac{1}{2}\left(-d(x, S_k) - \max_{l \neq k}(-d(x, S_l))\right) \\
&= \frac{1}{2}\left(\min_{l \neq k} d(x, S_l) - d(x, S_k)\right).
\end{aligned}$$

From the definition of g,

$$\begin{aligned}
g_y^{\mathcal{S}}(x) \leq \max_{k \neq y} g_k(x) \iff & \frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_y) \leq \max_{k \neq y}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_k)\right) \\
\iff & \min_{k \neq y} d(x, S_k) \leq d(x, S_y).
\end{aligned}$$

So it follows that $R(g^{\mathcal{S}}) = \mathrm{er}_P(h_{\mathcal{S}})$. Similarly, $R_{\gamma,\mathbf{z}}(g^{\mathcal{S}}) = E_{\mathbf{z}}^{\gamma}(h_{\mathcal{S}})$.

By bounding the covering numbers of our class $\Delta\mathcal{G}_B$ of functions, and then by removing the restriction that B be specified in advance, we obtain the following result.

Theorem 2.7. *Let $\delta \in (0, 1)$ and suppose P is a probability measure on $Z = [0, 1]^n \times [C]$. With P^m-probability at least $1 - \delta$, $\mathbf{z} \in Z^m$ will be such that we have the following: for all B and for all $\gamma \in (0, 1)$, for all $h_{\mathcal{S}} \in H_B$,*

$$\mathrm{er}_P(h_{\mathcal{S}}) \leq E_{\mathbf{z}}^{\gamma}(h_{\mathcal{S}}) + \sqrt{\frac{2}{m}\left(2nB\ln\left(\frac{6n}{\gamma}\right) + 2B\ln C + \ln\left(\frac{2n}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

2.4. Proof of Theorem 2.7

As the preceding discussion makes clear, the functions in $\Delta\mathcal{G}_B$ are all of the form $(\Delta g_1^{\mathcal{S}}, \ldots, \Delta g_C^{\mathcal{S}})$, where

$$\Delta g_k^{\mathcal{S}} = \frac{1}{2}\left(\min_{l \neq k} d(x, S_l) - d(x, S_k)\right).$$

Here, $\mathcal{S}$, as before, involves B boxes. To simplify, we will bound the covering numbers of $2\Delta\mathcal{G}_B$, noting that an α-covering of $2\Delta\mathcal{G}_B$ is an $\alpha/2$-covering of $\Delta\mathcal{G}_B$. So, consider the class $\mathcal{F} = \mathcal{F}_B = \{f_{\mathcal{S}} : h_{\mathcal{S}} \in H_B\}$, where $f_{\mathcal{S}} = 2\Delta g^{\mathcal{S}}$; that is, $f_{\mathcal{S}} : [0,1]^n \to \mathbb{R}^C$ is given by

$$(f_{\mathcal{S}})_k(x) = \min_{l \neq k} d(x, S_l) - d(x, S_k).$$

We use some of the same notations and ideas as developed in the proof of Theorem 2.2. We will define $\hat{\mathcal{F}}$ to be the subset of $\mathcal{F} = \mathcal{F}_B$ in which each S_k is of the form

$$S_k = \bigcup_{j=1}^{B_k} \mathbf{I}(\hat{u}(k,j), \hat{v}(k,j))$$

where $\hat{u}(k,j), \hat{v}(k,j) \in A^n_\gamma$. We will show that $\hat{\mathcal{F}}$ is a $\gamma/2$-cover for $\mathcal{F}_B$, and hence a $\gamma/4$ cover for $\Delta\mathcal{G}_B$.

Suppose $f_{\mathcal{S}} \in \mathcal{F}_B$, where $\mathcal{S} = (S_1, S_2, \ldots, S_C)$. Suppose $\hat{u}(k,j), \hat{v}(j,k) \in A^n_\gamma$ are chosen as in the proof of Theorem 2.2, and that $\hat{S}_k = \bigcup_{j=1}^{B_k} \mathbf{I}(\hat{u}(k,j), \hat{v}(k,j))$ and $\hat{\mathcal{S}} = (\hat{S}_1, \ldots, \hat{S}_C)$. Let $\hat{f}_{\mathcal{S}}$ be the corresponding function in $\hat{\mathcal{F}}_B$: $\hat{f}_{\mathcal{S}} = f_{\hat{\mathcal{S}}}$. As argued in the proof of Theorem 2.2, for all k and all x, $|d(x, S_k) - d(x, \hat{S}_k)| \leq \gamma/4$. For any k, and for any x,

$$\begin{aligned} \left| f_{\mathcal{S}}(x) - \hat{f}_{\mathcal{S}}(x) \right| &= \left| \min_{l \neq k} d(x, S_l) - d(x, S_k) - \left(\min_{l \neq k} d(x, \hat{S}_l) - d(x, \hat{S}_k) \right) \right| \\ &\leq \left| \min_{l \neq k} d(x, S_l) - \min_{l \neq k} d(x, \hat{S}_l) \right| + \left| d(x, \hat{S}_k) - d(x, S_k) \right|. \end{aligned}$$

The second term in this last line is bounded by $\gamma/4$, by the observation just made. Consider the first term. Suppose $\min_{l \neq k} d(x, S_l) = d(x, S_r)$. Then, since $|d(x, S_r) - d(x, \hat{S}_r)| \leq \gamma/4$, it follows that

$$\min_{l \neq k} d(x, \hat{S}_k) \leq d(x, \hat{S}_r) \leq d(x, S_r) + \gamma/4 = \min_{l \neq k} d(x, S_l) + \gamma/4.$$

Similarly, $\min_{l \neq k} d(x, S_k) \leq \min_{l \neq k} d(x, \hat{S}_k) + \gamma/4$. So,

$$\left| \min_{l \neq k} d(x, S_l) - \min_{l \neq k} d(x, \hat{S}_l) \right| \leq \frac{\gamma}{4}.$$

Therefore, $|f_{\mathcal{S}}(x) - \hat{f}_{\mathcal{S}}(x)| \leq \gamma/4 + \gamma/4 = \gamma/2$. This establishes that $\hat{\mathcal{F}}_B$ is a $\gamma/2$-cover of $\mathcal{F}_B$ with respect to the supremum metric on functions $X \to \mathbb{R}^C$, meaning that for every $f_{\mathcal{S}} \in \mathcal{F}_B$, there is $\hat{f}_{\mathcal{S}} \in \hat{\mathcal{F}}_B$ with

$$\sup_{x \in X} \|f_{\mathcal{S}}(x) - \hat{f}_{\mathcal{S}}(x)\|_\infty \leq \gamma/2.$$

In particular, therefore, for any m and for any $\mathbf{x} \in X^m$, $\hat{\mathcal{F}}_B$ is a $\gamma/2$ cover for $\mathcal{F}_B$ with respect to the $d_{\mathbf{x}}$ metric.

We now see that the covering number $\mathcal{N}(\gamma/4, \Delta\mathcal{G}_B, 2m)$ is, for all m, bounded above by $|\hat{\mathcal{F}}_B|$. Bounding this cardinality as in the proof of Theorem 2.2 gives

$$\mathcal{N}(\gamma/4, \Delta\mathcal{G}_B, 2m) \leq \left(\frac{6n}{\gamma}\right)^{2nB} C^B.$$

Taking $\delta/2^B$ in place of δ, using the covering number bound now obtained, and noting that $R(g^{\mathcal{S}}) = \text{er}_P(h_{\mathcal{S}})$ and $R_{\gamma,\mathbf{z}}(g^{\mathcal{S}}) = E^\gamma_{\mathbf{z}}(h_{\mathcal{S}})$, Theorem 2.6 shows that, with probability at least $1 - \delta/2^B$, for all $\gamma \in (0,1)$, if $h_{\mathcal{S}} \in H_B$, then

$$\text{er}_P(h_{\mathcal{S}}) \leq E^\gamma_{\mathbf{z}}(h_{\mathcal{S}}) + \sqrt{\frac{2}{m}\left(2nB\ln\left(\frac{6n}{\gamma}\right) + 2B\ln C + \ln\left(\frac{2}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

(We have used $B\ln 2 \leq B\ln C$.) So, with probabilty at least $1 - \delta$, *for all* B, this bound holds, completing the proof.

3. Conclusions

This paper has analyzed the generalization performance of a type of multi-category classifier introduced in [13], which has a very natural interpretation. Based on boxes, each of which contains points of one particular classification, remaining points (outside the boxes) are categorised as belonging to the same class as the nearest box, where distance is measured by the d_1, or Manhattan, metric. The generalization error bounds we derive involve 'margin error', and we have two types of bound: one applying to the case in which the margin error is zero, and the other to the general situation.

It is certainly possible to use metrics other than the Manhattan metric (as we explored in [5], for the two-class classification case, using the d_∞ metric). Additionally, and perhaps simultaneously, one might consider regions other than box regions, one example being to use the Euclidean metric and to consider spheres (as in [17], where a different approach to obtaining error bounds, using sample compression bounds, was taken).

Acknowledgements

This work was supported in part by the IST Programme of the European Community, under the PASCAL2 Network of Excellence, IST-2007-216886. Part of the work was carried out while Martin Anthony was a visitor at Rutgers University Center for Operations Research. We are grateful to the anonymous referees for their helpful comments.

References

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

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

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

[4] Martin Anthony and Normal L. Biggs. *Computational Learning Theory: an Introduction.*Cambridge Tracts in Theoretical Computer Science (30), Cambridge University Press, 1992.

[5] M. Anthony and J. Ratsaby. The performance of a new hybrid classifier based on boxes and nearest neighbors. In *International Symposium on Artificial Intelligence and Mathematics*, 2012. (Also RUTCOR Research Report RRR 17-2011, Rutgers University, 2011).

[6] M. Anthony and J. Ratsaby. Robust cutpoints in the logical analysis of numerical data. *Discrete Applied Mathematics*, 160:355–364, 2012.

[7] 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):525–536, 1998.

[8] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

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

[10] 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):292–306, 2000.

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

[12] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929, 1978.

[13] Giovanni Felici, Bruno Simeone, and Vincenzo Spinelli. Classification techniques and error control in logic mining. In *Data Mining 2010*, volume 8 of *Annals of Information Systems*, pages 99–119, 2010.

[14] Yann Guermeur. VC theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8:2551–2594, 2007.

[15] P.L. Hammer, Y. Liu, S. Szedmák, and B. Simeone. Saturated systems of homogeneous boxes and the logical analysis of numerical data. *Discrete Applied Mathematics*, 144(1-2):103–109, 2004.

[16] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, September 1992.

[17] M. Marchand and J. Shawe-Taylor. The set covering machine. *Journal of Machine Learning Research*, 3, 2002.

[18] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

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

[20] Vladimir N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag, New York, 1995.

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