RUTCOR
RESEARCH
REPORT

Sample width for multi-category classifiers

Martin Anthony[a] Joel Ratsaby[b]

RRR 29-2012, November 2012

RUTCOR
Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
http://rutcor.rutgers.edu/~rrr

[a]Department of Mathematics, The London School of Economics and Political Science, Houghton Street, London WC2A 2AE, UK. m.anthony@lse.ac.uk

[b]Electrical and Electronics Engineering Department, Ariel University Center of Samaria, Ariel 40700, Israel. ratsaby@ariel.ac.il

Rutcor Research Report
RRR 29-2012, November 2012

Sample width for multi-category classifiers

Martin Anthony Joel Ratsaby

Abstract. In a recent paper, the authors introduced the notion of *sample width* for binary classifiers defined on the set of real numbers. It was shown that the performance of such classifiers could be quantified in terms of this sample width. This paper considers how to adapt the idea of sample width so that it can be applied in cases where the classifiers are multi-category and are defined on some arbitrary metric space.

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

1 Introduction

By a (multi-category) classifier on a set X, we mean a function mapping from X to $[C] = \{1, 2, \ldots, C\}$ where $C \geq 2$ is the number of possible categories. Such a classifier indicates to which of the C different classes objects from X belong and, in supervised machine learning, it is arrived at on the basis of a *sample*, a set of objects from X together with their classifications in $[C]$. In [4], the notion of *sample width* for binary classifiers ($C = 2$) mapping from the real line $X = \mathbb{R}$ was introduced and in [5], this was generalized to finite metric spaces. In this paper, we consider how a similar approach might be taken to the situation in which C could be larger than 2, and in which the classifiers map not simply from the real line, but from some metric space (which would not generally have the linear structure of the real line). The definition of sample width is given below, but it is possible to indicate the basic idea at this stage: we define sample width to be at least γ if the classifier achieves the correct classifications on the sample and, furthermore, for each sample point, the minimum distance to a point of the domain having a different classification is at least γ.

A key issue that arises in machine learning is that of *generalization error*: given that a classifier has been produced by some learning algorithm on the basis of a (random) sample of a certain size, how can we quantify the accuracy of that classifier, where by its accuracy we mean its likely performance in classifying objects from X correctly? In this paper, we seek answers to this question that involve not just the sample size, but the sample width.

2 Probabilistic modelling of learning

We work in a version of the popular 'PAC' framework of computational learning theory (see [14, 7]). This model assumes that the sample $\mathbf{s}$ consists of an ordered set (x_i, y_i) of labeled examples, where $x_i \in X$ and $y_i \in Y = [C]$, and that each (x_i, y_i) in the training sample $\mathbf{s}$ has been generated randomly according to some fixed (but unknown) probability distribution P on $Z = X \times Y$. (This includes, as a special case, the situation in which each x_i is drawn according to a fixed distribution on X and is then labeled deterministically by $y_i = t(x_i)$ where t is some fixed function.) Thus, a sample $\mathbf{s}$ of length m can be thought of as being drawn randomly according to the product probability distribution P^m. An appropriate measure of how well $h : X \to Y$ would perform on further randomly drawn points is its *error*, $\mathrm{er}_P(h)$, the probability that $h(X) \neq Y$ for random (X, Y).

Given any function $h \in H$, we can measure how well h matches the training sample through its *sample error*

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

(the proportion of points in the sample incorrectly classified by h). Much classical work in learning theory (see [7, 14], for instance) related the error of a classifier h to its sample error. A typical result would state that, for all $\delta \in (0,1)$, with probability at least $1-\delta$, for all h belonging to some specified set if functions, we have $\mathrm{er}_P(h) < \mathrm{er}_{\mathbf{s}}(h) + \epsilon(m,\delta)$, where $\epsilon(m,\delta)$ (known as a *generalization error bound*) is decreasing in m and δ. Such results can be derived using uniform convergence theorems from probability theory [15, 11, 8], in which case $\epsilon(m,\delta)$ would typically involve a quantity known as the growth function of the set of classifiers [15, 7, 14, 2]. More recently, emphasis has been placed on 'learning with a large margin'. (See, for instance [13, 2, 1, 12].) The rationale behind margin-based generalization error bounds in the two-category classification case is that if a binary classifier can be thought of as a geometrical separator between points, and if it 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 generalization error bound can be obtained. Margin-based results apply when the binary classifiers are derived from real-valued function by 'thresholding' (taking their sign). Margin analysis has been extended to multi-category classifiers in [9].

3 The width of a classifier

We now discuss the case where the underlying set of objects X forms a metric space. Let X be a set on which is defined a metric $d : X \times X \to \mathbb{R}$. For a subset S of X, define the distance $d(x,S)$ from $x \in X$ to S as follows:

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

We define the *diameter* of X to be

$$\mathrm{diam}(X) := \sup_{x,y \in X} d(x,y).$$

We will denote by $\mathcal{H}$ the set of all possible functions h from X to $[C]$.

The paper [4] introduced the notion of the width of a binary classifier at a point in the domain, in the case where the domain was the real line $\mathbb{R}$. Consider a set of points $\{x_1, x_2, \ldots, x_m\}$ from $\mathbb{R}$, which, together with their true classifications $y_i \in \{-1,1\}$, yield a *training sample*

$$\mathbf{s} = ((x_j, y_j))_{j=1}^m = ((x_1,y_1),(x_2,y_2),\ldots,(x_m,y_m)).$$

We say that $h : \mathbb{R} \to \{-1,1\}$ achieves sample margin at least γ on $\mathbf{s}$ if $h(x_i) = y_i$ for each i (so that h correctly classifies the sample) and, furthermore, h is constant on each of

the intervals $(x_i - \gamma, x_i + \gamma)$. It was then possible to obtain generalization error bounds in terms of the sample width. In this paper we use an analogous notion of width to analyse multi-category classifiers defined on a metric space.

For each k between 1 and C, let us denote by S_k^h (or simply by S_k) the sets corresponding to the function $h : X \to [C]$, defined as follows:

$$S_k := h^{-1}(k) = \{x \in X : h(x) = k\}. \tag{1}$$

We define the *width* $w_h(x)$ of h at a point $x \in X$ as follows:

$$w_h(x) := \min_{l \neq h(x)} d(x, S_l).$$

In other words, it is the distance from x to the set of points that are labeled differently from $h(x)$. The term 'width' is appropriate since the functional value is just the geometric distance between x and the complement of $S_{h(x)}$.

Given $h : X \to [C]$, for each k between 1 and C, we define $f_k^h : X \to \mathbb{R}$ by

$$f_k^h(x) = \min_{l \neq k} d(x, S_l) - d(x, S_k),$$

and we define $f^h : X \to \mathbb{R}^C$ by setting the kth component function of f^h to be f_k^h: that is, $(f^h)_k = f_k^h$.

Note that if $h(x) = k$, then $f_k^h(x) \geq 0$ and $f_j^h(x) \leq 0$ for $j \neq k$. The function f contains geometrical information encoding how 'definitive' the classification of a point is: if $f_k^h(x)$ is a large positive number, then the point x belongs to category k and is a large distance from differently classified points. We will regard h as being in error on $x \in X$ if $f_k^h(x)$ is not positive, where $k = h(x)$ (that is, if the classification is not unambiguously revealed by f^h). The error $\mathrm{er}_P(h)$ of h can then be expressed in terms of the function f^h:

$$\mathrm{er}_P(h) = P\left(f_Y^h(X) < 0\right). \tag{2}$$

We define the class $\mathcal{F}$ of functions as

$$\mathcal{F} := \left\{f^h(x) : h \in \mathcal{H}\right\}. \tag{3}$$

Note that f^h is a mapping from X to the bounded set $[-\mathrm{diam}(X), \mathrm{diam}(X)]^C \subseteq \mathbb{R}^C$. Henceforth, we will use $\gamma > 0$ to denote a *width parameter* whose value is in the range $(0, \mathrm{diam}(X)]$.

For a positive width parameter $\gamma > 0$ and a training sample $\mathbf{s}$, the *empirical* (sample) γ-width error is defined as

$$E_{\mathbf{s}}^{\gamma}(h) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{I}\left(f_{y_j}^h(x_j) \leq \gamma\right).$$

(Here, $\mathbb{I}(A)$ is the indicator function of the set, or event, A.) We shall also denote $E^\gamma_{\mathbf{s}}(h)$ by $E^\gamma_{\mathbf{s}}(f)$ where $f = f^h$. Note that

$$\begin{aligned} f^h_y(x) \leq \gamma &\iff \min_{l \neq y} d(x, S_l) - d(x, S_y) \leq \gamma \\ &\iff \exists l \neq y \text{ such that } d(x, S_l) \leq d(x, S_y) + \gamma. \end{aligned}$$

So the empirical γ-width error on the sample is the proportion of points in the sample which are either misclassified by h or which are classified correctly, but lie within distance γ of the set of points classified differently. (We recall that $h(x) = y$ implies $d(x, S_y) = 0$.) Our aim is to show that (with high probability) the generalization error $\mathrm{er}_P(h)$ is not much greater than $E^\gamma_{\mathbf{s}}(h)$. (In particular, as a special case, we want to bound the generalization error given that $E^\gamma_{\mathbf{s}}(h) = 0$.) This will imply that if the learner finds a hypothesis which, for a large value of γ, has a small γ-width error, then that hypothesis is likely to have small error. What this indicates, then, is that if a hypothesis has a large width on most points of a sample, then it will be likely to have small error.

4 Covering numbers

4.1 Covering numbers

We will make use of some results and ideas from large-margin theory. One central idea in large-margin analysis is that of *covering numbers*. We will discuss different types of covering numbers, so we introduce the idea in some generality to start with.

Suppose (A, d) is a (pseudo-)metric space and that $\alpha > 0$. Then an α-cover of A (with respect to d) is a finite subset C of A such that, for every $a \in A$, there is some $c \in C$ such that $d(a, c) \leq \alpha$. If such a cover exists, then the mimimum cardinality of such a cover is the *covering number* $\mathcal{N}(A, \alpha, d)$.

We are working with the set $\mathcal{F}$ of vector-valued functions from X to $\mathbb{R}^C$, as defined earlier. We define the sup-metric d_∞ on F as follows: for $f, g : X \to \mathbb{R}^C$,

$$d_\infty(f, g) = \sup_{x \in X} \max_{1 \leq k \leq C} |f_k(x) - g_k(x)|,$$

where f_k denotes the kth component function of f. (Note that each component function is bounded, so the metric is well-defined.)

We can bound the covering numbers $\mathcal{N}(\mathcal{F}, \alpha, d_\infty)$ of $\mathcal{F}$ (with respect to the sup-metric) in terms of the covering numbers of X with respect to its metric d. The result is as follows.

Theorem 4.1 *For $\alpha \in (0, \mathrm{diam}(X)]$,*

$$\mathcal{N}(\mathcal{F}, \alpha, d_\infty) \leq \left(\frac{9\,\mathrm{diam}(X)}{\alpha}\right)^{CN_\alpha},$$

where $N_\alpha = \mathcal{N}(X, \alpha/3, d)$.

4.2 Smoothness of the function class

As a first step towards establishing this result, we prove that the functions in $\mathcal{F}$ satisfy a certain Lipschitz (or smoothness) property.

Proposition 4.2 *For every $f \in \mathcal{F}$, and for all $x, x' \in X$,*

$$\max_{1\leq k\leq C} |f_k(x) - f_k(x')| \leq 2\, d(x, x'). \tag{4}$$

Proof: Let $x, x' \in X$ and fix k between 1 and C. We show that

$$|f_k(x) - f_k(x')| \leq 2\, d(x, x').$$

Recall that, since $f \in \mathcal{F}$, there is some $h : X \to [C]$ such that, for all x,

$$f_k(x) = \min_{l\neq k} d(x, S_l) - d(x, S_k)$$

where, for each i, $S_i = h^{-1}(i)$. We have

$$\begin{aligned} |f_k(x) - f_k(x')| &= \left|\min_{l\neq k} d(x, S_l) - d(x, S_k) - \min_{l\neq k} d(x', S_l) + d(x', S_k)\right| \\ &\leq \left|\min_{l\neq k} d(x, S_l) - \min_{l\neq k} d(x', S_l)\right| + |d(x, S_k) - d(x', S_k)| \end{aligned}$$

We consider in turn each of the two terms in this final expression. We start with the second. Let $\epsilon > 0$ and suppose $r \in S_k$ is such that $d(x, r) < d(x, S_k) + \epsilon$. Such an r exists since $d(x, S_k) = \inf_{y\in S_k} d(x, y)$. Let $s \in S_k$ be such that $d(x', s) < d(x', S_k) + \epsilon$. Then

$$d(x, S_k) \leq d(x, r) \leq d(x, s) + \epsilon,$$

since $d(x, r) < d(x, S_k) + \epsilon$ and $d(x, S_k) \leq d(x, s)$. So,

$$d(x, S_k) \leq d(x', s) + d(x, x') + \epsilon < d(x', S_k) + \epsilon + d(x, x') + \epsilon = d(x', S_k) + 2\epsilon.$$

A similar argument establishes

$$d(x', S_k) \leq d(x', s) \leq d(x', r) + \epsilon \leq d(x, r) + d(x, x') + \epsilon < d(x, S_k) + d(x, x') + 2\epsilon.$$

Combining these two inequalities shows $|d(x, S_k) - d(x', S_k)| \leq d(x, x') + 2\epsilon$. Since this holds for all $\epsilon > 0$, it follows that $|d(x, S_k) - d(x', S_k)| \leq d(x, x')$. Next we show

$$\left| \min_{l \neq k} d(x, S_l) - \min_{l \neq k} d(x', S_l) \right| \leq d(x, x').$$

Suppose that $\min_{k \neq l} d(x, S_l) = d(x, S_p)$ and that $\min_{k \neq l} d(x, S_l) = d(x, S_q)$. Let $\epsilon > 0$. Suppose that $y \in S_p$ is such that $d(x, y) < d(x, S_p) + \epsilon$ and that $y' \in S_q$ is such that $d(x', y') < d(x', S_q) + \epsilon$. Then

$$d(x', S_q) \leq d(x', S_p) \leq d(x', y) \leq d(x, x') + d(x, y) < d(x, x') + d(x, S_p) + \epsilon$$

and

$$d(x, S_p) \leq d(x, S_q) \leq d(x, y') \leq d(x, x') + d(x', y') = d(x, x') + d(x', S_q) + \epsilon.$$

From these inequalities, it follows that $|d(x, S_p) - d(x', S_q)| \leq d(x, x') + \epsilon$, for all ϵ, and it follows that $|d(x, S_p) - d(x', S_q)| \leq d(x, x')$. □

Next, we exploit this 'smoothness' to construct a cover for $\mathcal{F}$.

4.3 Covering $\mathcal{F}$

Let the subset $C_\gamma \subseteq X$ be a *minimal* size $\alpha/3$-cover for X with respect to the metric d. So, for every $x \in X$ there is some $\hat{x} \in C_\gamma$ such that $d(x, \hat{x}) \leq \alpha/3$. Denote by N_α the cardinality of C_α.

Let

$$\Lambda_\alpha = \left\{ \lambda_i = i\alpha : i = -\left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil, \ldots, -1, 0, 1, 2, \ldots, \left\lceil \frac{3\,\mathrm{diam}(X)}{\alpha} \right\rceil \right\} \tag{5}$$

and define the class $\hat{F}$ to be all functions $\hat{f} : C_\alpha \to (\Lambda_\alpha)^C$. Then $\hat{F}$ is of a finite size equal to $|\Lambda_\alpha|^{C\,N_\alpha}$. For any $\hat{f} \in \hat{F}$ define the extension $\hat{f}_{ext} : X \to \mathbb{R}^C$ of $\hat{f}$ to the whole domain X as follows: given $\hat{f}$ (which is well-defined on the points $\hat{x}_i$ of the cover) then for every point x in the ball $B_{\alpha/3}(\hat{x}_i) = \{x \in X : d(x, \hat{x}_i) \leq \alpha/3\}$, we let $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$, for all $\hat{x}_i \in C_\alpha$ (where, if, for a point x there is more than one point $\hat{x}_i$ such that $x \in B_{\alpha/3}(\hat{x}_i)$, we arbitrarily pick one of the points $\hat{x}_i$ in order to assign the value of $\hat{f}_{ext}(x)$). There is a one-to-one correspondence

between the functions $\hat{f}$ and the functions $\hat{f}_{ext}$. Hence the set $\hat{F}_{ext} = \left\{ \hat{f}_{ext} : \hat{f} \in \hat{F} \right\}$ is of cardinality equal to $|\Lambda_\alpha|^{C N_\alpha}$.

We claim that for any $f \in \mathcal{F}$ there exists an $\hat{f}_{ext}$ such that $d_\infty(f, \hat{f}_{ext}) \leq \alpha$. To see this, first for every point $\hat{x}_i \in C_\alpha$, consider $f(\hat{x}_i)$ and find a corresponding element in Λ_α^C, (call it $\hat{f}(\hat{x}_i)$) such that

$$\max_{1 \leq k \leq C} |(f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \leq \alpha/3. \tag{6}$$

(That there exists such a value follows by design of Λ_α.) By the above definition of extension, it follows that for all points $x \in B_{\alpha/3}(\hat{x}_i)$ we have $\hat{f}_{ext}(x) = \hat{f}(\hat{x}_i)$. Now, from (4) we have for all $f \in \mathcal{F}$,

$$\max_{1 \leq i \leq k} \sup_{x \in B_{\alpha/3}(\hat{x}_i)} |(f(x))_k - (f(\hat{x}_i))_k| \;\; \leq \;\; 2d(x, \hat{x}_i) \leq 2\alpha/3. \tag{7}$$

Hence for any $f \in \mathcal{F}$ there exists a function $\hat{f} \in \hat{F}$ with a corresponding $\hat{f}_{ext} \in \hat{F}_{ext}$ such that, given an $x \in X$, there exists $\hat{x}_i \in C_\alpha$ such that, for each k between 1 and C, $|(f(x))_k - (\hat{f}_{ext}(x))_k| = |(f(x))_k - (\hat{f}_{ext}(\hat{x}_i))_k|$. The right hand side can be expressed as

$$\begin{aligned}
|(f(x))_k - (\hat{f}_{ext}(\hat{x}_i))_k| &= |(f(x))_k - (\hat{f}(\hat{x}_i))_k| \\
&= |(f(x))_k - (f(\hat{x}_i))_k + (f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \\
&\leq |(f(x))_k - (f(\hat{x}_i))_k| + |(f(\hat{x}_i))_k - (\hat{f}(\hat{x}_i))_k| \\
&\leq 2\alpha/3 + \alpha/3 \qquad (8) \\
&= \alpha.
\end{aligned}$$

where (8) follows from (6) and (7).

Hence the set $\hat{F}_{ext}$ forms an α-covering of the class $\mathcal{F}$ in the sup-norm. Thus we have the following covering number bound.

$$\mathcal{N}(\mathcal{F}, \alpha, d_\infty) \;\; \leq \;\; |\Lambda_\alpha|^{C N_\alpha} = \left(2 \left\lceil \frac{3 \operatorname{diam}(X)}{\alpha} \right\rceil + 1 \right)^{C N_\alpha}. \tag{9}$$

Theorem 4.1 now follows because (for $0 < \alpha \leq \operatorname{diam}(X)$)

$$2 \left\lceil \frac{3 \operatorname{diam}(X)}{\alpha} \right\rceil + 1 \leq 2 \left(\frac{3 \operatorname{diam}(X)}{\alpha} + 1 \right) + 1 = \frac{6 \operatorname{diam}(X)}{\alpha} + 3 \leq \frac{9 \operatorname{diam}(X)}{\alpha}.$$

5 Generalization error bounds

We present two results. The first bounds the generalization error in terms of a width parameter γ for which the γ-width error on the sample is zero; the second (more general but looser

in that special case) bounds the error in terms of γ and the γ-width error on the sample (which could be non-zero).

Theorem 5.1 *Suppose that X is a metric space of diameter* $\operatorname{diam}(X)$. *Suppose P is any probability measure on $Z = X \times [C]$. Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the following holds for $\mathbf{s} \in Z^m$: for any function $h : X \to [C]$, and for any $\gamma \in (0, \operatorname{diam}(X)]$, if $E^\gamma_{\mathbf{s}}(h) = 0$, then*

$$\operatorname{er}_P(h) \leq \frac{2}{m}\left(C\mathcal{N}(X,\gamma/12,d)\log_2\left(\frac{36\operatorname{diam}(X)}{\gamma}\right) + \log_2\left(\frac{4\operatorname{diam}(X)}{\delta\gamma}\right)\right).$$

Theorem 5.2 *Suppose that X is a metric space of diameter* $\operatorname{diam}(X)$. *Suppose P is any probability measure on $Z = X \times [C]$. Let $\delta \in (0,1)$. Then, with probability at least $1-\delta$, the following holds for $\mathbf{s} \in Z^m$: for any function $h : X \to [C]$, and for any $\gamma \in (0, \operatorname{diam}(X)]$,*

$$\operatorname{er}_P(h) \leq E^\gamma_{\mathbf{s}}(h) + \sqrt{\frac{2}{m}\left(C\mathcal{N}(X,\gamma/6,d)\ln\left(\frac{18\operatorname{diam}(X)}{\gamma}\right) + \ln\left(\frac{2\operatorname{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

What we have in Theorem 5.2 is a high probability bound that takes the following form: for all h and for all $\gamma \in (0, \operatorname{diam}(X)]$,

$$\operatorname{er}_P(h_{\mathcal{S}}) \leq E^\gamma_{\mathbf{s}}(h) + \epsilon(m,\gamma,\delta),$$

where ϵ tends to 0 as $m \to \infty$ and ϵ decreases as γ increases. The rationale for seeking such a bound is that there is likely to be a trade-off between width error on the sample and the value of ϵ: taking γ small so that the error term $E^\gamma_{\mathbf{s}}(h)$ is zero might entail a large value of ϵ; and, conversely, choosing γ large will make ϵ relatively small, but lead to a large sample 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.

Proof of Theorem 5.1

The proof uses techniques similar to those first used in [15, 14, 8, 11] and in subsequent work extending those techniques to learning with real-valued functions, such as [10, 3, 1, 6]. The first observation is that if

$$Q = \{\mathbf{s} \in Z^m : \exists h \in \mathcal{H} \text{ with } E^\gamma_{\mathbf{s}}(h) = 0,\ \operatorname{er}_P(h) \geq \epsilon\}$$

and

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

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

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

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

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, for instance, [11, 2].)

Let $\mathcal{F} = \{f^h : h \in \mathcal{H}\}$ be the set of vector-valued functions derived from $\mathcal{H}$ as before, and let $\hat{\mathcal{F}}$ be a minimal $\gamma/2$-cover of $\mathcal{F}$ in the d_∞-metric. Theorem 4.1 tells us that the size of $\hat{\mathcal{F}}$ is no more than

$$\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right)^{CN},$$

where $N = \mathcal{N}(X, \gamma/6, d)$.

Fix $\mathbf{z} \in Z^{2m}$. Suppose $\tau\mathbf{z} = (\mathbf{s}, \mathbf{s}') \in T$ and that h is such that $E^\gamma_{\mathbf{s}}(h) = 0$ and $E^0_{\mathbf{s}'}(h) \geq \epsilon/2$. Let $\hat{f} \in \hat{\mathcal{F}}$ be such that $d_\infty(\hat{f}, f^h) \leq \gamma/2$. Then the fact that $E^\gamma_{\mathbf{s}}(h) = 0$ implies that $E^{\gamma/2}_{\mathbf{s}}(\hat{f}) = 0$. And, the fact that $E^0_{\mathbf{s}'}(h) \geq \epsilon/2$ implies that $E^{\gamma/2}_{\mathbf{s}'}(\hat{f}) \geq \epsilon/2$. To see the first claim, note that we have $f^h_y(x) > \gamma$ for all (x, y) in the sample $\mathbf{s}$ and that, since, for all k, and all x, $|\hat{f}_k(x) - f^h_k(x)| \leq \gamma/2$, we have also that for all such (x, y), $\hat{f}_y(x) > \gamma/2$. For the second claim, we observe that if $f^h_y(x) \leq 0$ then $\hat{f}_y(x) \leq \gamma/2$.

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

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

By symmetry, $\mathbb{P}\left(\sigma\mathbf{z} \in R(\hat{f})\right) = \mathbb{P}\left(\sigma(\tau\mathbf{z}) \in R(\hat{f})\right)$. Suppose that $E^{\gamma/2}_{\mathbf{s}'}(\hat{f}) = r/m$, where $r \geq \epsilon m/2$ is the number of (x_i, y_i) in $\mathbf{s}'$ on which $\hat{f}$ is such that $\hat{f}_{y_i}(x_i) \leq \gamma/2$. Then those

permutations σ such that $\sigma(\tau\mathbf{z}) \in R(\hat{f})$ 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 $\hat{f} \in \hat{\mathcal{F}}$,

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

We therefore have

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

So,

$$P^m(Q) \leq 2\, P^{2m}(T) \leq 2\, |\hat{\mathcal{F}}|\, 2^{-\epsilon m/2} \leq 2\, \left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right)^{CN},$$

where $N = \mathcal{N}(X, \gamma/6, d)$. This is at most δ when

$$\epsilon = \frac{2}{m}\left(CN \log_2\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Next, we use this to obtain a result in which γ is not prescribed in advance. 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 $h \in \mathcal{H}$ with $E^{\alpha_2}_{\mathbf{z}}(h) = 0$ and $\mathrm{er}_P(h) \geq \epsilon_1(m, \alpha_1, \delta)$, where

$$\epsilon_1(m, \alpha_1, \delta) = \frac{2}{m}\left(C\mathcal{N}(X, \alpha_1/6, d) \log_2\left(\frac{18\,\mathrm{diam}(X)}{\alpha_1}\right) + \log_2\left(\frac{2}{\delta}\right)\right).$$

Then the result just obtained 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 result from [6], that

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

In other words, with probability at least $1-\delta$, *for all* $\gamma \in (0, \mathrm{diam}(X)]]$, we have that if $h \in \mathcal{H}$ satisfies $E^{\gamma}_{\mathbf{s}}(h) = 0$, then $\mathrm{er}_P(h) < \epsilon_2(m, \gamma, \delta)$, where

$$\epsilon_2(m, \gamma, \delta) = \frac{2}{m}\left(C\mathcal{N}(X, \gamma/12, d) \log_2\left(\frac{36\,\mathrm{diam}(X)}{\gamma}\right) + \log_2\left(\frac{4\,\mathrm{diam}(X)}{\delta\gamma}\right)\right).$$

Note that γ now need not be prescribed in advance.

Proof of Theorem 5.2

Guermeur [9] has developed a framework in which to analyse multi-category classification, and we can apply one of his results to obtain the bound of Theorem 5.2, which is generalization error bound applicable to the case in which the γ-width sample error is not zero. In that framework, there is a set $\mathcal{G}$ of functions from X into $\mathbb{R}^C$, and a typical $g \in \mathcal{G}$ is represented by its component functions g_k for $k = 1$ to C. Each $g \in \mathcal{G}$ satisfies the constraint

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

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 $M(v, k) = \frac{1}{2}\left(v_k - \max_{l \neq k} v_l\right)$ 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 $\mathbf{s} \in (X \times [C])^m$, let

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

A result following from [9] is (in the above notation) as follows:

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

$$R(g) \leq R_{\gamma, \mathbf{s}}(g) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\Delta\mathcal{G}, \gamma/4, d_\infty) + \ln\left(\frac{2d}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

(In fact, the result from [9] involves emprical covering numbers rather than d_∞-covering numbers. The latter are at least as large as the empirical covering numbers, but we use

these because we have bounded them earlier in this paper.) For each function $h : X \to [C]$, let $g = g^h$ be the function $X \to \mathbb{R}^C$ defined by

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

where, as before, $S_j = h^{-1}(j)$. Let $\mathcal{G} = \{g^h : h \in \mathcal{H}\}$ be the set of all such g. 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,

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

From the definition of g, the event that $g_y(x) \leq \max_{k\neq y} g_k(x)$ is equivalent to the event that

$$\frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_y) \leq \max_{k\neq y}\left(\frac{1}{C}\sum_{i=1}^{C} d(x, S_i) - d(x, S_k)\right),$$

which is equivalent to $\min_{k\neq y} d(x, S_k) \leq d(x, S_y)$. It can therefore be seen that $R(g) = \mathrm{er}_P(h)$. Similarly, $R_{\gamma,\mathbf{s}}(g) = E_{\mathbf{s}}^{\gamma}(h)$.

Noting that $\Delta\mathcal{G} = (1/2)\mathcal{F}$, so that an $\alpha/2$ cover of $\mathcal{F}$ will provide an $\alpha/4$ cover of $\Delta\mathcal{G}$, we can therefore apply Guermeur's result to see that with probability at least $1-\delta$, for all h and for all $\gamma \in (0, \mathrm{diam}(X)]$,

$$\mathrm{er}_P(h) \leq E_{\mathbf{s}}^{\gamma}(h) + \sqrt{\frac{2}{m}\left(\ln \mathcal{N}(\mathcal{F}, \gamma/2, d_\infty) + \ln\left(\frac{2\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m}$$

$$\leq E_{\mathbf{s}}^{\gamma}(h) + \sqrt{\frac{2}{m}\left(C\mathcal{N}(X, \gamma/6, d)\ln\left(\frac{18\,\mathrm{diam}(X)}{\gamma}\right) + \ln\left(\frac{2\,\mathrm{diam}(X)}{\gamma\delta}\right)\right)} + \frac{1}{m}.$$

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

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

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

[3] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615631, 1997.

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

[5] M. Anthony and J. Ratsaby. Learning on finite metric spaces. RUTCOR Research Report RRR-19-2012, June 2012.

[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):525–536, 1998.

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

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

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

[10] D. Haussler. Decision theoretic generalizations of the pac model for neural net and other learning applications, *Information and Computation*, 100: 78–150, 1992.

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

[12] 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.

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

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

[15] 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): 264–280, 1971.