Cross-Validation for Binary Classification by Real-Valued Functions: Theoretical Analysis

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, U.K.

Sean B. Holden
Department of Computer Science
University College London
Gower Street
London WC1E 6BT, U.K.

August 5, 1999

Abstract

This paper concerns the use of real-valued functions for binary classification problems. Previous work in this area has concentrated on using as an error estimate the 'resubstitution' error (that is, the empirical error of a classifier on the training sample) or its derivatives. However, in practice, cross-validation and related techniques are more popular. Here, we analyse theoretically the accuracy of the holdout and cross-validation estimators for the case where real-valued functions are used as classifiers. We then introduce two new error estimation techniques, which we call the *adaptive holdout estimate* and the *adaptive cross-validation estimate*, and we perform a similar analysis for these. Finally, we show how our results can be applied to certain types of neural network.

Keywords: Supervised Learning, Cross-Validation Estimate, Holdout Estimate, Sample Complexity, Error Estimation, PAC Learning

1 Introduction

Recently, within probabilistic models of machine learning, attention has focused on the use of real-valued functions for binary classification (as in [4, 7, 8, 24], for instance). It has been shown that in many cases, one can obtain fairly accurate estimates of a classifier's error when the classifier is a real-valued function which achieves the correct classification of a training sample, with a large 'margin' (a notion to be made precise in what follows). Not only have the results obtained for such classification models often improved upon the analogous results for classification by binary-valued classes, but in some cases, error estimates can be given in the new model when none can possibly be given in the binary classification model. In this paper we take a further step in the analysis of classification by real-valued classes. Previous analyses have concentrated on using as an error estimate the 'resubstitution' error (that is, the empirical error of a classifier on the training sample) or its derivatives. However, in practice, cross-validation and related techniques are more popular and more effective. In this paper, we perform a probabilistic performance analysis of the cross-validation and holdout estimates in the case where real-valued functions are used as classifiers. We then devise and analyse two new error estimates, which we call the *adaptive holdout estimate* and the *adaptive cross-validation estimate*. The results we obtain are what Kearns and Ron [20] (in the binary-valued context) have described as 'sanity-check' bounds. This name reflects the fact that the bounds are not better than those one has for the corresponding resubstitution error estimate. Nonetheless, we believe that such sanity-check bounds are worthwhile, first, as they show that our new error estimation techniques work, and second, as they tell us about error estimates that are widely used in practice, rather than the resubstitution estimate, which is not widely used due to its well-known poor performance. The error bounds might be looser than one would like for real, practical algorithms, but that is a consequence of the very general approach we take. In fact, we demonstrate in section 5 that in the present model of learning it is not possible to obtain bounds for the cross-validation estimate that are better than those for the resubstitution estimate. Consequently, an approach differing significantly from the usual approaches used in the literature is likely to be required if bounds better than sanity-check bounds are to be obtained. The possibility is left open of our new techniques out-performing the resubstitution or other techniques in practice, and preliminary experimental results suggest that this can be the case.

We start, in Section 2 with a general discussion of the error estimation problem. In Section 3 we present previously obtained results to set the present work in context. In this section, we provide the basic definitions of the standard holdout and cross-validation estimators, and we describe the PAC probabilistic learning model. The next section 4 then discusses two of the new ideas used in the new work: the use of real-valued functions for binary classification, and the importance of a 'margin' in the classification. We tentatively suggest a modified way of measuring holdout and cross-validation errors, taking into account the classification margin. Section 5 presents the probability results which are at the centre of our analysis (the proofs of which are deferred until Section 9). Section 6, building on the results of Section 5 and the suggestions made in Section 4, formalises the definitions of two new error estimators applicable when using real-valued functions for classification purposes and discusses their accuracy. In Section 7 we obtain general error bounds in terms of the 'fat-shattering dimension', a parameter describing the complexity of a function class. Applications to certain types of neural network are then given in Section 8. Section 9 contains the proofs of the probability results of Section 5. The concluding Section 10 suggests some open problems.

2 Error estimation

In this section we describe the standard classification and cross-validation frameworks and present some previous results on the error estimation abilities of these techniques.

We start by describing a standard form of the 'PAC' model of learning [26, 10, 14]. This concerns two-class pattern classification problems, in which we wish to classify objects $x \in X$ as belonging to class -1 or class 1. (The set X will be called the set of *examples*.) The appearance of examples and their class membership labels is governed by an arbitrary distribution P on $Z = X \times \{-1, 1\}$. During 'learning', we are presented with a *training sample* $\mathbf{z}$ of n *labelled examples* drawn independently at random according to P,

$$\mathbf{z} = ((x_1, y_1), \ldots, (x_n, y_n)).$$

A learning algorithm L takes such samples $\mathbf{z}$ and uses them to select a hypothesis $h : X \to \{-1, 1\}$ from a hypothesis space $\mathcal{H}$, which may be used to classify further objects drawn independently at random according to P. The (true) *error* $\mathrm{er}_P(h)$ of $h = L(\mathbf{z})$ is defined to be

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

and the *empirical error* $\mathrm{er}_{\mathbf{z}}(h)$ is defined to be

$$\mathrm{er}_{\mathbf{z}}(h) = \frac{1}{n} \left| \{i : h(x_i) \neq y_i\} \right|.$$

The empirical error is often referred to as the 'error on the training sample' or, in statistics, the 'resubstitution estimate', and can be regarded as an estimate of the value of the actual error $\mathrm{er}_P(h)$. It is natural to ask how close it is to the actual error, and a significant body of research exists which bounds the probability that the resubstitution estimate is a 'bad' estimate of the actual error in an appropriately defined sense. A typical result is the following, due to Vapnik [27] (see also [10]). (Like all such results, this requires some fairly reasonable measurability conditions on the hypothesis space $\mathcal{H}$. These are not very stringent, and we shall assume without further comment that they hold.)

Theorem 1 *For parameters $\epsilon, \gamma \in (0, 1]$, and for any distribution P, we have*

$$P^n \left\{\mathbf{z} \in Z^n : \exists h \in \mathcal{H} \text{ for which } \mathrm{er}_{\mathbf{z}}(h) \leq (1 - \gamma)\epsilon \text{ and } \mathrm{er}_P(h) > \epsilon\right\} < 8\Pi_{\mathcal{H}}(2n) \exp(-\gamma^2 \epsilon n / 4)$$

where $\Pi_{\mathcal{H}}()$ denotes the growth function *which we define below.*

Theorem 1 tells us that the probability of obtaining a training sequence such that L can choose a hypothesis from $\mathcal{H}$ that simultaneously has empirical error no greater than $(1 - \gamma)\epsilon$ and true error greater than ϵ is bounded by the stated expression. If $\mathcal{H}$ has finite *Vapnik-Chervonenkis (VC) dimension* (see for example [10, 3]) then the upper bound can be made arbitrarily small by choosing a large enough n (which does not depend on P); this will be explained below.

A great deal of research in recent years has concentrated on extending results of this kind to deal with multiple class problems (see for example Anthony and Shawe-Taylor [6]), and real-valued functions. (See for example Haussler [14], Pollard [22], Anthony [2], among others.) However there is another potential direction for extending such results which to date has received much less emphasis, although it is no less interesting. Cheng and Titterington [11] have raised the interesting question of what happens to the result stated in Theorem 1 if a cross-validation estimate is used in place of a resubstitution estimate. This problem has recently been considered by Holden [15, 16, 17] and Kearns and Ron [20], and a related problem has been studied by Blum *et al.* [9]. There are different types of cross-validation estimate. These operate by splitting the sample $\mathbf{z}$ into m sections. A section is removed and the remaining examples used in conjunction with L to select a hypothesis h from $\mathcal{H}$. The removed section is then used to estimate the error of h, using the obvious technique of finding the fraction of the examples in the removed section that h classifies incorrectly (that is, the empirical error on the removed section). In the standard cross-validation techniques, this is repeated for each section and the average of the m individual estimates is then used as a final estimate of the true error that will be obtained if L is applied on the complete training sequence.

It is now well-known (see for example Toussaint [25]) that the resubstitution estimate is not a good estimator of true error in practice due to its considerable optimistic bias. (This simply means that it tends to underestimate the true error.) Clearly therefore the study of error estimates other than the resubstitution estimate is of significant importance if we wish to obtain results applicable to practical learning applications.

3 Previous related work

We now describe some previous work of Holden [16] which derives results like Theorem 1 for cross-validation error estimates in the 'realisable' case. The initial new results of this paper provide similar results in a model of cross-validation where real-valued functions are used for classification.

3.1 PAC learning in the realisable case

The results in [16] apply to the *realisable* learning case, in which (with probability one) only one label occurs with each example. Explicitly, the correct classification for each example is assigned by a *target concept* $c : X \to \{-1, 1\}$, which we assume belongs to our *hypothesis class* $\mathcal{H}$. In this case, the training sample $\mathbf{z}(\mathbf{x}, c) \in (X \times \{-1, 1\})^n$ of length n is obtained by drawing n inputs independently at random according to an arbitrary distribution μ on X in order to obtain a sequence $\mathbf{x} = (x_1, \ldots, x_n) \in X^n$, and then forming the training sample

$$\mathbf{z} = \mathbf{z}(\mathbf{x}, c) = ((x_1, c(x_1)), \ldots, (x_n, c(x_n))).$$

We denote by $\mathbf{z}_{\mathcal{H}}$ the set of all possible training sequences,

$$\mathbf{z}_{\mathcal{H}} = \{\mathbf{z}(\mathbf{x}, c) : c \in \mathcal{H}, \mathbf{x} \in X^n, n \geq 1\}.$$

In general, we will abbreviate $\mathbf{z}(\mathbf{x}, c)$ to $\mathbf{z}$ when the underlying target concept and input sequence are clear from the context. The length of the training sequence is denoted by n throughout this paper. It should be noted that the realisable framework just described is a special case of the framework in which the labelled examples are chosen according to a joint probability distribution P on $Z = X \times \{-1, 1\}$; see [10, 6]. In learning from examples we attempt to use $\mathbf{z}$ to select a hypothesis $h : X \to \{-1, 1\}$ from $\mathcal{H}$. A *learning algorithm* or *learner* L is a technique that accomplishes this, such that $h = L(\mathbf{z})$. Throughout the paper we denote by $L(\mathbf{z})$ the hypothesis produced when L is applied to $\mathbf{z}$. Although the notation suggests that L is a function, we also specifically allow randomized algorithms and all results given hold for such algorithms. When L is a randomized algorithm we may obtain different hypotheses by applying L to some fixed sequence $\mathbf{z}$ more than once. This is a well-known property of many actual learning algorithms. A learning algorithm is *consistent* if for any $\mathbf{z} \in \mathbf{z}_{\mathcal{H}}$, L produces a hypothesis h for which $h(x_i) = c(x_i)$ for $1 \leq i \leq n$. Given any training sequence $\mathbf{z} \in \mathbf{z}_{\mathcal{H}}$ of length n we denote by $\mathcal{H}[\mathbf{z}]$ the set of all hypotheses in $\mathcal{H}$ that are consistent with $\mathbf{z}$,

$$\mathcal{H}[\mathbf{z}] = \{h \in \mathcal{H} : h(x_i) = c(x_i) \text{ for } 1 \leq i \leq n\}.$$

Finally, the *true error* (usually abbreviated to the *error*) of any hypothesis $h \in \mathcal{H}$ is denoted $\mathrm{er}_\mu(h, c)$, and is defined as the probability that h disagrees with c on an input chosen at random according to μ:

$$\mathrm{er}_\mu(h, c) = \mu\{x \in X : h(x) \neq c(x)\}.$$

We abbreviate $\mathrm{er}_\mu(h, c)$ to $\mathrm{er}_\mu(h)$ when the target concept is clear from the context.

In order to describe the results of Holden [16] and to move towards the results of this paper, we now define the *growth function* and *Vapnik-Chervonenkis (VC) dimension*. Let $S = \{x_1, \ldots, x_i\}$ be a subset of X with $|S| = i$ and let F be a class of functions $f : X \to \{-1, 1\}$. With each $f \in F$ we associate the set $A_f = \{x \in X : f(x) = 1\}$. We define the set $\Pi_F(S)$ as

$$\Pi_F(S) = \{S \cap A_f : f \in F\}$$

and we define the growth function $\Pi_F(j)$ by

$$\Pi_F(j) = \max\{|\Pi_F(S)| : S \subseteq X \text{ and } |S| = j\}.$$

The VC dimension of F is denoted $\mathrm{VCdim}(F)$ and defined as

$$\mathrm{VCdim}(F) = \max\{j : \Pi_F(j) = 2^j\}.$$

By 'Sauer's lemma' (see [3, 10, 23]) we know that if $d = \mathrm{VCdim}(F)$ is finite and $j \geq d \geq 1$, then

$$\Pi_F(j) < \left(\frac{ej}{d}\right)^d.$$

The following theorem is a standard result in the concept learning framework.

Theorem 2 ([10]; see also [3], proposition 8.2.3) *For any distribution μ, any $c \in \mathcal{H}$, any ϵ, and any $n \geq 8/\epsilon$,*

$$\mu^n\{\mathbf{x} \in X^n : \exists h \in \mathcal{H}[\mathbf{z}] \text{ such that } \mathrm{er}_\mu(h) \geq \epsilon\} < 2\Pi_{\mathcal{H}}(2n)2^{-\epsilon n/2}.$$

Equivalently, for any n and $\delta \in (0, 1)$, and any distribution μ, with probability at least $1 - \delta$, a sample $\mathbf{x}$ of length n satisfies: if $h \in H$ and $\mathrm{er}_{\mathbf{x}}(h) = 0$, then

$$\mathrm{er}_\mu(h) < \epsilon(n, \delta) = \frac{8}{n}\left(\log\left(\frac{2}{\delta}\right) + \log\left(\Pi_H(2n)\right)\right),$$

where $\log$ *denotes binary logarithm.*

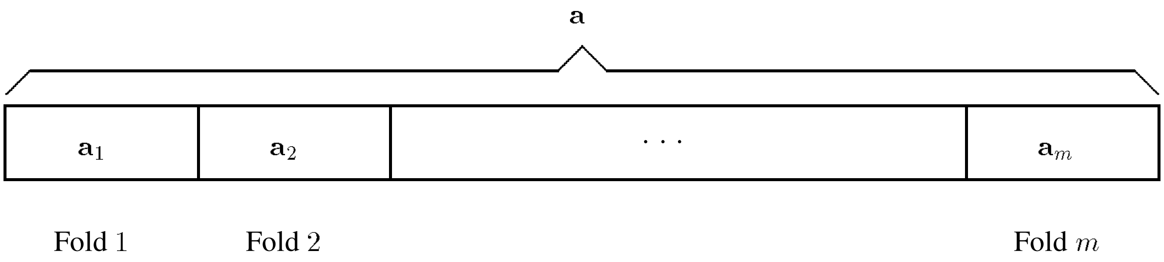

Figure 1: Any sequence $\mathbf{a} \in A^n$ can be divided into m contiguous sections of equal size called *folds*. It is assumed that $m \geq 2$ and m divides n. Each fold contains $n' = (n/m)$ elements of $\mathbf{a}$.

A consequence of this theorem and Sauer's Lemma is that if $\mathcal{H}$ has finite VC dimension, then the estimation error $\epsilon(n, \delta)$ for a consistent algorithm can be made arbitrarily small by making n large enough.

We now take a more general perspective. Assume that we have an *error estimator* which takes the training sequence $\mathbf{z}$ and learner L and uses them to make an estimate of $\text{er}_\mu(L(\mathbf{z}))$. We will denote the value of the estimate by $\text{est}(L, \mathbf{z})$. Holden [16] obtains upper bounds on the quantity

$$P^n\{\mathbf{x} \in X^n : |\text{est}(L, \mathbf{z}) - \text{er}_\mu(L(\mathbf{z}))| > \epsilon\}.$$

when the error estimators arise from cross-validation. Such bounds can, as we shall see, easily be converted into high-confidence bounds on the estimation error. We now describe these estimates in more detail.

3.2 Empirical errors

A slight abuse of set-theoretic notation will be used in describing and dealing with the error estimation techniques of interest. The following definitions are most easily explained diagrammatically (figure 1). Given any sequence $\mathbf{a} = (a_1, \ldots, a_n) \in A^n$ for some set A we will speak of it as being divided into m *folds*, where m is an integer, $m \geq 2$, and m divides n. The folds form contiguous sections of equal size, each fold containing $n' = (n/m)$ elements of $\mathbf{a}$. We denote by $\mathbf{a}_i$ the ith fold,

$$\mathbf{a}_i = (a_{(i-1)n'+1}, \ldots, a_{in'})$$

and we denote by $\mathbf{a}\backslash\mathbf{a}_i$ the sequence obtained by removing the ith fold from $\mathbf{a}$,

$$\mathbf{a}\backslash\mathbf{a}_i = (a_1, \ldots, a_{(i-1)n'}, a_{in'+1}, \ldots, a_n).$$

Given any two sequences $\mathbf{a} \in A^{n_1}$ and $\mathbf{b} \in A^{n_2}$ we denote by $\mathbf{ab}$ the sequence

$$\mathbf{ab} = (a_1, \ldots, a_{n_1}, b_1, \ldots, b_{n_2}) \in A^{n_1+n_2}.$$

Clearly when $\mathbf{a}$ is divided into m folds we have $\mathbf{a} = \mathbf{a}_1\mathbf{a}_2 \cdots \mathbf{a}_m$. If $\mathbf{x} \in X^n$ is any sequence obtained by drawing n inputs independently at random according to μ, then for a hypothesis $h \in$

$\mathcal{H}$ and a target concept $c \in \mathcal{H}$ we define the *empirical error* $\text{er}_{\mathbf{x}}(h, c)$ of h as

$$\text{er}_{\mathbf{x}}(h, c) = \frac{1}{n} \sum_{i=1}^{n} I_{\{h(x_i) \neq c(x_i)\}}$$

where I denotes the indicator function. If $\mathbf{x}$ is divided into m folds, then for m hypotheses $h_1, \ldots, h_m$ each of which is a member of $\mathcal{H}$, and a target concept $c \in \mathcal{H}$, we define the *m-fold empirical error* $\text{er}_{\mathbf{x}}(h_1, \ldots, h_m, c)$ to be

$$\text{er}_{\mathbf{x}}(h_1, \ldots, h_m, c) = \frac{1}{m} \sum_{i=1}^{m} \text{er}_{\mathbf{x}_i}(h_i, c) = \frac{1}{n} \sum_{i=1}^{m} \sum_{j=1}^{n'} I_{\{h_i(x_{(i-1)n'+j}) \neq c(x_{(i-1)n'+j})\}}$$

In general, we abbreviate $\text{er}_{\mathbf{x}}(h, c)$ and $\text{er}_{\mathbf{x}}(h_1, \ldots, h_m, c)$ to $\text{er}_{\mathbf{x}}(h)$ and $\text{er}_{\mathbf{x}}(h_1, \ldots, h_m)$ respectively when the target concept of interest is clear from the context.

3.3 The holdout estimate

Of particular importance in the theory developed in [16], and in this paper, is the *holdout estimate*. Assume we have a training sequence $\mathbf{z}$ of length n obtained from some target concept $c \in \mathcal{H}$ by drawing inputs independently at random according to μ. We wish to make an estimate of the value of $\text{er}_{\mu}(L(\mathbf{z}))$ for our learner L, without making use of any further examples. Informally, the holdout estimate works as follows. We split $\mathbf{z}$ into two sections such that $\mathbf{z} = \mathbf{z}_1\mathbf{z}_2$, use L in conjunction with $\mathbf{z}_1$ to obtain a hypothesis $L(\mathbf{z}_1)$, calculate the empirical error of $L(\mathbf{z}_1)$ using $\mathbf{z}_2$, and use the value obtained as the required estimate. Formally, let $\mathbf{x} \in X^n$ be the sequence used to obtain $\mathbf{z}$. Choose an appropriate value m and divide $\mathbf{x}$ and $\mathbf{z}$ into m folds. The value $\text{est}_H^m(L, \mathbf{z})$ of the holdout estimate is

$$\text{est}_H^m(L, \mathbf{z}) = \text{er}_{\mathbf{x}_m}(L(\mathbf{z} \backslash \mathbf{z}_m)).$$

Note that it is not usually the case that in calculating the holdout estimate in practice we impose the restriction that m divides n. This restriction is imposed here to aid the theoretical analysis.

3.4 The cross-validation estimate

In order to calculate the holdout estimate we split the training sample into folds and use the final fold in conjunction with a hypothesis obtained using the first $(m-1)$ folds to estimate $\text{er}_{\mu}(L(\mathbf{z}))$. The m-fold cross-validation estimate takes this idea a step further in that, rather than using L once to obtain a single hypothesis, it uses L once for each fold to obtain m—possibly different—hypotheses, each of which is used to obtain an estimate. These individual estimates are then averaged to obtain a final estimate. Formally, the value $\text{est}_{CV}^m(L, \mathbf{z})$ of the m-fold cross-validation estimate is given by

$$\text{est}_{CV}^m(L, \mathbf{z}) = \text{er}_{\mathbf{x}}(L(\mathbf{z} \backslash \mathbf{z}_1), \ldots, L(\mathbf{z} \backslash \mathbf{z}_m)) = \frac{1}{m} \sum_{i=1}^{m} \text{er}_{\mathbf{x}_i}(L(\mathbf{z} \backslash \mathbf{z}_i)).$$

If $n = m$ then this estimate is called the leave-one-out cross-validation estimate. We do not consider the latter estimate in this paper, and always assume that $n > m$.

3.5 Previous bounds

Holden [16] obtains results in probability theory. He then applies these to obtain, for both the holdout estimate and the cross-validation estimate, upper bounds on the probability of obtaining a training sequence $\mathbf{z}$ for which the estimate differs from the true error $\text{er}_\mu(L(\mathbf{z}))$ by more than a specified constant $0 < \epsilon \leq 1$, where L is *any* consistent learning algorithm. Specifically, he obtains functions $f_1(n, m, \epsilon, \text{VCdim}(\mathcal{H}))$ and $f_2(n, m, \epsilon, \text{VCdim}(\mathcal{H}))$ such that under quite general conditions,

$$\mu^n \left\{\mathbf{x} \in X^n : |\text{est}_H^m(L, \mathbf{z}) - \text{er}_\mu(L(\mathbf{z}))| > \epsilon\right\} \leq f_1(n, m, \epsilon, \text{VCdim}(\mathcal{H}))$$

and

$$\mu^n \left\{\mathbf{x} \in X^n : |\text{est}_{CV}^m(L, \mathbf{z}) - \text{er}_\mu(L(\mathbf{z}))| > \epsilon\right\} \leq f_2(n, m, \epsilon, \text{VCdim}(\mathcal{H}))$$

where $\mathbf{z}$ is obtained from $\mathbf{x}$ using the target concept c. In each case, if $\text{VCdim}(\mathcal{H})$ is finite, the upper bound can be made arbitrarily small by choosing a large enough n. Also, the upper bounds given are independent of the target concept c and the distribution μ.

The main results in [16] may be stated as follows.

Theorem 3 (Holden [16]) *For any $n > m \geq 3$ where m divides n and $n \geq 2m/\epsilon^2$, any $c \in \mathcal{H}$, any distribution μ, and any ϵ such that $0 < \epsilon \leq 1$, let*

$$\mathcal{P}_{\text{H}} = \mu^n \left\{\mathbf{x} \in X^n : \exists h_m, h \in \mathcal{H} \text{ with } \text{er}_{\mathbf{z}\backslash\mathbf{z}_m}(h_m) = \text{er}_{\mathbf{z}}(h) = 0, \text{er}_\mu(h) > \text{er}_{\mathbf{x}_m}(h_m) + \epsilon\right\}$$

and

$$\mathcal{P}_{\text{CV}} = \mu^n \left\{\mathbf{x} \in X^n : \exists h_1, \ldots, h_m, h \in \mathcal{H} \text{ with } \text{er}_{\mathbf{z}\backslash\mathbf{z}_i}(h_i) = 0 \ (1 \leq i \leq m), \text{er}_{\mathbf{z}}(h) = 0 \text{ and} \right.$$
$$\left. \text{er}_\mu(h) > \frac{1}{m}\sum_{i=1}^{m} \text{er}_{\mathbf{z}_i}(L(\mathbf{z}\backslash\mathbf{z}_i)) + \epsilon\right\}.$$

Then

$$\mathcal{P}_{\text{H}} < 2\Pi_{\mathcal{H}}^2\left(n\left(1 + \frac{1}{m}\right)\right) 2^{-n\epsilon/2m},$$

and $\mathcal{P}_{\text{CV}} \leq m\mathcal{P}_{\text{H}}$.

Theorem 4 (Holden [16]) *For any $n > m \geq 3$ where m divides n and $n \geq 2m/\epsilon^2$, any $c \in \mathcal{H}$, any distribution μ, and any ϵ such that $0 < \epsilon \leq 1$, we have*

$$\mu^n \left\{\mathbf{x} \in X^n : |\text{est}_H^m(L, \mathbf{z}) - \text{er}_\mu(L(\mathbf{z}))| > \epsilon\right\} < 2\Pi_{\mathcal{H}}^2\left(n\left(1 + \frac{1}{m}\right)\right) 2^{-n\epsilon/2m}$$

where $\Pi_{\mathcal{H}}^2()$ denotes $[\Pi_{\mathcal{H}}()]^2$, and

$$\mu^n \left\{\mathbf{x} \in X^n : |\text{est}_{CV}^m(L, \mathbf{z}) - \text{er}_\mu(L(\mathbf{z}))| > \epsilon\right\} < 2\, m\, \Pi_{\mathcal{H}}^2\left(n\left(1 + \frac{1}{m}\right)\right) 2^{-n\epsilon/2m}.$$

Using Sauer's Lemma, these bounds can be translated into bounds of the form f_1, f_2 described above; see Holden [16] for details.

From this we have the following estimation error bounds.

Theorem 5 *With the notation as above, if $\delta \in (0, 1)$ is a constant, and $d = \mathrm{VCdim}(\mathcal{H})$ is finite and $d \geq 1$, then for any consistent learning algorithm L, any $c \in \mathcal{H}$ and any μ, we have:*

1. *with probability at least $1 - \delta$,*

$$\mathrm{er}_\mu(L(\mathbf{z})) < \mathrm{est}_H^m(L, \mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\left(\log\left(\frac{2}{\delta}\right) + 2\log\left(\Pi_{\mathcal{H}}(n + n/m)\right)\right)\right\},$$

2. *with probability at least $1 - \delta$,*

$$\mathrm{er}_\mu(L(\mathbf{z})) < \mathrm{est}_{CV}^m(L, \mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\left(\log\left(\frac{2m}{\delta}\right) + 2\log\left(\Pi_{\mathcal{H}}(n + n/m)\right)\right)\right\}.$$

Sauer's Lemma shows that if the VC-dimension of $\mathcal{H}$ is finite, then the estimation error can be made arbitrarily small by choosing n large.

4 Using real-valued function classes

4.1 Measuring error with a margin

We now consider how one might use a class of real-valued functions for binary classification, and how we might apply a type of cross-validation to estimate the classification error. Our aim is to obtain theorems like those described above. The techniques used are similar to those used by Holden [16] and Anthony and Bartlett [4, 5].

As before, X will denote the set of all possible *examples* to be classified. (For instance, X would be $\mathbf{R}^k$ in the case of a neural network having k real inputs.) Z will denote $X \times \{-1, 1\}$, the set of all possible *labelled examples*. We work in the general framework in which there is some unknown probability distribution P on Z, representing the concept to be learned. (As noted earlier, such a set-up includes as a special case the realisable framework in which there is a $\{-1, 1\}$-valued target *concept* $c \in \mathcal{H}$ on X and a probability distribution μ on X.) We consider the case where a fixed class of functions $\mathcal{H}$ mapping into the real interval $[-1, 1]$ is used for binary classification of the examples: given $f \in \mathcal{H}$ and $x \in X$, the resulting binary classification of x by f is simply $\mathrm{sgn}(f(x))$, where $\mathrm{sgn}(a) = 1$ if $a \geq 0$ and $\mathrm{sgn}(a) = -1$ if $a < 0$, and the *error* of f (with respect to P) is

$$\mathrm{er}_P(f) = P\left(\{(x, y) \in X \times \{-1, 1\} : \mathrm{sgn}(f(x)) \neq y\}\right).$$

The use of real-valued functions for binary classification has been considered in [4, 7, 8, 24] within versions of the PAC model of learning and versions of 'agnostic' PAC learning (see Kearns et al. [19] and Haussler [14]), and it has been shown that there are advantages in considering the values of the real function during training, rather than merely its sign. In particular, as suggested in [13, 27], classification with a large 'margin' (the distance between the classification boundary and classified examples) appears to lead in many cases to better estimates of error. Motivated by this, we study here a model of cross-validation applicable to binary classification by real-valued functions, in which one takes into account the size of the margin.

For $\gamma \in (0, 1)$, the *error of f at margin γ* (with respect to P) is

$$\mathrm{er}_P^{\gamma}(f) = P\left(\{(x, y) \in Z : yf(x) < \gamma\}\right),$$

the probability that if $y = -1$ then $f(x) > -\gamma$, and if $y = 1$ then $f(x) < \gamma$. Notice that if $\mathrm{sgn}(f(x)) \neq y$ then (remembering that $y \in \{-1, 1\}$), we certainly have $yf(x) < 0 < \gamma$, so for any $\gamma \in (0, 1)$, and any $f \in H$, $\mathrm{er}_P^{\gamma}(f) \geq \mathrm{er}_P(f)$. It will also be useful to observe that if $\gamma < \gamma'$ then $\mathrm{er}_P^{\gamma}(f) \leq \mathrm{er}_P^{\gamma'}(f)$. The concept of a margin γ will be used in this way throughout this paper. Intuitively, the margin can be thought of as specifying how stringent we wish to be in considering f to be making an error. If the margin is small then $f(x)$ must be a significant distance from y for us to decide that an error has been made, but if the margin is large we will tolerate only a small distance between $f(x)$ and y. We can also consider the case of a large margin γ to be a method for forcing f to be "confident" in order to have a small error at margin γ, in the sense that $|f(x) - y|$ must be small.

The *empirical error* $\mathrm{er}_{\mathbf{z}}(f)$ of f on $\mathbf{z} \in Z^n$ is defined to be the proportion of examples in $\mathbf{z}$ sign-misclassified by f; that is, if $z_i = (x_i, y_i) \in X \times \{-1, 1\}$, then

$$\mathrm{er}_{\mathbf{z}}(f) = \frac{1}{n} \left|\{i : \mathrm{sgn}(f(x_i)) \neq y_i\}\right|.$$

4.2 Measuring cross-validation error

A straightforward way in which to extend the preceding theory would be to use this measure of empirical error in place of the previous definition of empirical error. It is clear that results analogous to Theorems 4 and 5 would hold, with the VC-dimension of $\mathcal{H}$ replaced by the VC-dimension of $\{\mathrm{sgn}(h) : h \in \mathcal{H}\}$, where $\mathrm{sgn}(h)(x) = \mathrm{sgn}(h(x))$. However, we take a different approach in which we take into consideration not simply whether h gives the correct sign on a particular example, but the actual real value it gives.

Suppose that $\gamma \in (0, 1)$ is a constant. We start by consider cross-validation for algorithms which correctly classify with a margin of at least γ. Thus, such algorithms 'interpolate' to accuracy $1-\gamma$. (Later, we consider more general algorithms.) Explicitly, suppose that, given a sample $\mathbf{z} \in Z^n = (X \times \{-1, 1\})^n$, our learning algorithm L produces a function $f \in H$ such that for each (x_i, y_i) in $\mathbf{z}$, $y_i f(x_i) \geq \gamma$. (Thus, not only does $f(x_i)$ have the correct sign on each of the examples, but the correct classification is achieved with a margin of at least γ. Note that this is, in a sense, a generalisation of consistency as defined for the standard cross-validation model, but is not as demanding as requiring that $f(x_i) = y_i$, which would be inappropriately strong.) We call such an algorithm a *γ-margin algorithm*. Consider how we might attempt to generalise the definition of the holdout estimate to apply to such algorithms. For a labelled sample $\mathbf{z}$, let $\mathbf{z}_m$ and

$\mathbf{z}\backslash\mathbf{z}_m$ be defined in the obvious way (in analogy with the definitions for the standard holdout estimate). One approach to cross-validation would be to use the empirical error $\mathrm{er}_{\mathbf{z}_m}(L(\mathbf{z}\backslash\mathbf{z}_m))$ as an error estimate. However, we shall find some advantage in using a 'margin-based' estimate of the error on $\mathbf{z}_m$. For a constant $\gamma \in (0,1)$, $f \in \mathcal{H}$ and $\mathbf{z} \in Z^n$, the *empirical error at margin* γ is

$$\mathrm{er}_{\mathbf{z}}^{\gamma}(f) = \frac{1}{n}\left|\{i : y_i f(x_i) < \gamma\}\right|.$$

We shall see that in many cases, with high probability, the empirical error of $L(\mathbf{z}\backslash\mathbf{z}_m)$ at margin γ, on $\mathbf{z}_m$ is a good estimate of the true error $\mathrm{er}_P(L(\mathbf{z}))$ of L on the whole sample $\mathbf{z}$. That is, we shall obtain results concerning the estimator

$$\mathrm{est}_{H,\gamma}^{m}(L,\mathbf{z}) = \mathrm{er}_{\mathbf{z}_m}^{\gamma}(L(\mathbf{z}\backslash\mathbf{z}_m)).$$

As in [16], we shall relate the cross-validation estimate to the holdout estimate and thereby obtain a similar result for the corresponding cross-validation estimate

$$\mathrm{est}_{CV,\gamma}^{m}(L,\mathbf{z}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{er}_{\mathbf{z}_i}^{\gamma}(L(\mathbf{z}\backslash\mathbf{z}_i)).$$

5 Results for real-valued cross-validation models

5.1 Covering numbers

To state our main result, we require the important notion of *covering number*. Given $W \subseteq \mathbf{R}^k$ and a positive real number ϵ, we say that $C \subseteq \mathbf{R}^k$ is a d_∞ ϵ*-cover for* W if $C \subseteq W$ and for every $w \in W$ there is a $v \in C$ such that

$$\max\{|w_i - v_i| : i = 1,\ldots,k\} < \epsilon.$$

We say that W is *totally bounded* if for each $\epsilon > 0$ there is a *finite* ϵ-cover for W. In this case, we define the d_∞ ϵ*-covering number of* W, $\mathcal{N}(W,\epsilon,d_\infty)$, to be the minimum cardinality of a d_∞ ϵ-cover for W.

Suppose that F is a class of functions from X to $\mathbf{R}$. Given $\mathbf{x} = (x_1, x_2, \ldots, x_k) \in X^k$, we let $F_{|\mathbf{x}}$ be the subset of $\mathbf{R}^k$ given by

$$F_{|\mathbf{x}} = \{(f(x_1), f(x_2), \ldots, f(x_k)) : f \in F\}.$$

For a positive number ϵ, we define the *(uniform) covering number* $\mathcal{N}_\infty(F,\epsilon,k)$ to be the maximum, over all $\mathbf{x} \in X^k$, of the covering number $\mathcal{N}(F_{|\mathbf{x}},\epsilon,d_\infty)$ (and we take it to be infinite if there is no bound on these covering numbers); that is,

$$\mathcal{N}_\infty(F,\epsilon,k) = \max\left\{\mathcal{N}(F_{|\mathbf{x}},\epsilon,d_\infty) : \mathbf{x} \in X^k\right\}.$$

The covering number is a generalisation of the growth function. To see this, suppose that functions in H map into $\{-1,1\}$. Then, for all $\mathbf{x} \in X^k$, $H_{|\mathbf{x}}$ is finite and, for all $\epsilon < 1$, $\mathcal{N}(H_{|\mathbf{x}},\epsilon,d_\infty) = \left|H_{|\mathbf{x}}\right|$, so $\mathcal{N}_\infty(H,\epsilon,k) = \Pi_H(k)$.

5.2 Main probability theorems

Our main theorem for the form of holdout estimate defined above is as follows. Note that this result is concerned with showing that $\mathrm{est}^m_{H,\gamma}(L, \mathbf{z})$ does not *underestimate* by too much the true error $\mathrm{er}_P(L(\mathbf{z}))$ of $L(\mathbf{z})$. It does not concern the probability that $\mathrm{est}^m_{H,\gamma}(L, \mathbf{z})$ overestimates the true error. That is, we bound the probability that $\mathrm{er}_P(L(\mathbf{z})) > \mathrm{est}^m_{H,\gamma}(L, \mathbf{z}) + \epsilon$, rather than the probability that $|\mathrm{er}_P(L(\mathbf{z})) - \mathrm{est}^m_{H,\gamma}(L, \mathbf{z})| > \epsilon$. (If, in fact, $\mathrm{er}_P(L(\mathbf{z})) < \mathrm{est}^m_{H,\gamma}(L, \mathbf{z}) - \epsilon$ then using the error estimate does not, in any case, mislead us into thinking that a bad hypothesis is in fact a good one.)

Theorem 6 *Let $\gamma \in (0, 1)$. For $n > m \geq 3$, where m divides n and $n \geq 2m/\epsilon^2$, for any distribution P on $Z = X \times \{-1, 1\}$, and any $\epsilon \in (0, 1]$, let*

$$\mathcal{P}_{\mathrm{H}} = P^n \left(\left\{ \mathbf{z} \in Z^n : \exists h_m, h \in \mathcal{H} \text{ with } \mathrm{er}^\gamma_{\mathbf{z} \setminus \mathbf{z}_m}(h_m) = 0, \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0 \text{ and } \mathrm{er}_P(h) \geq \mathrm{er}^\gamma_{\mathbf{z}_m}(h_m) + \epsilon \right\} \right).$$

Then

$$\mathcal{P}_{\mathrm{H}} < 2 \left(\mathcal{N}_\infty \left(\mathcal{H}, \gamma/2, n + n/m \right) \right)^2 2^{-n\epsilon/2m}.$$

Equivalently, for $\delta, \gamma \in (0, 1)$ and for n, m as above, with probability at least $1 - \delta$, if $h_m, h \in \mathcal{H}$ and $\mathrm{er}^\gamma_{\mathbf{z} \setminus \mathbf{z}_m}(h_m) = \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0$ then

$$\mathrm{er}_P(h) < \mathrm{er}^\gamma_{\mathbf{z}_m}(h_m) + \epsilon(n, m, \delta, \gamma),$$

where

$$\epsilon(n, m, \delta, \gamma) = \max \left\{ \sqrt{\frac{2m}{n}}, \frac{2m}{n} \log\left(\frac{2}{\delta}\right) + \frac{4m}{n} \log\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n/m\right)\right) \right\}.$$

In particular, therefore, if L is a γ-margin algorithm, we have that, with probability at least $1 - \delta$,

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{est}^m_{H,\gamma}(L, \mathbf{z}) + \epsilon(n, m, \delta, \gamma).$$

For the modified cross-validation estimate, we have the following result.

Theorem 7 *For $n > m \geq 3$, where m divides n and $n \geq 2m/\epsilon^2$, and with the notation as above, if*

$$\mathcal{P}_{\mathrm{CV}} = P^n \left(\left\{ \mathbf{z} \in Z^n : \exists h_1, \ldots, h_m, h \in \mathcal{H} \text{ with } \mathrm{er}^\gamma_{\mathbf{z} \setminus \mathbf{z}_i}(h_i) = 0 \ (1 \leq i \leq m), \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0 \text{ and } \right.\right.$$

$$\left.\left. \mathrm{er}_P(h) > \frac{1}{m} \sum_{i=1}^m \mathrm{er}^\gamma_{\mathbf{z}_i}(h_i)) + \epsilon \right\} \right),$$

then

$$\mathcal{P}_{\mathrm{CV}} < 2m \left(\mathcal{N}_\infty \left(\mathcal{H}, \gamma/2, n + n/m \right) \right)^2 2^{-n\epsilon/2m}.$$

Equivalently, for $\delta, \gamma \in (0,1)$ *and for* n, m *as above, with probability at least* $1-\delta$*, if* $h_1, \ldots, h_m, h \in \mathcal{H}$ *and* $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_i}(h_i) = 0$ *for* $1 \le i \le m$*, and* $\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$ *then*

$$\mathrm{er}_P(h) < \frac{1}{m}\sum_{i=1}^{m} \mathrm{er}^{\gamma}_{\mathbf{z}_i}(h_i) + \epsilon(n, m, \delta, \gamma),$$

where

$$\epsilon(n, m, \delta, \gamma) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{2m}{\delta}\right) + \frac{4m}{n}\log\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n/m\right)\right)\right\}.$$

Hence, for any γ*-margin algorithm* L*, we have that, with probability at least* $1-\delta$*,*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{est}^{m}_{CV,\gamma}(L, \mathbf{z}) + \epsilon(n, m, \delta, \gamma).$$

It is perhaps unnatural to specify in advance the parameter γ. Indeed, it would seem more appropriate to compute γ on the basis of the training sample, and this will be the basis of the algorithm to be discussed later, in Section 6 (in which we define 'sample-based' modified holdout and cross-validation estimators). The following results will be useful.

Theorem 8 *For* $\delta \in (0,1)$ *and for* $n > m \ge 3$*, where* m *divides* n*, with probability at least* $1-\delta$*,* for all $\gamma \in (0,1)$*, if* $h_m, h \in \mathcal{H}$ *and* $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_m}(h_m) = \mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$ *then*

$$\mathrm{er}_P(h) < \mathrm{er}^{\gamma}_{\mathbf{z}_m}(h_m) + \varepsilon(n, m, \delta, \gamma),$$

where

$$\varepsilon(n, m, \delta, \gamma) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4}{\delta\gamma}\right) + \frac{4m}{n}\log\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/4, n + n/m\right)\right)\right\}.$$

Theorem 9 *For* $\delta \in (0,1)$ *and for* $n > m \ge 3$*, where* m *divides* n*, with probability at least* $1-\delta$*,* for all $\gamma \in (0,1)$*, if* $h_1, \ldots, h_m, h \in \mathcal{H}$ *and* $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_i}(h_i) = 0$ *for* $1 \le i \le m$*, and* $\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$*, then*

$$\mathrm{er}_P(h) < \frac{1}{m}\sum_{i=1}^{m} \mathrm{er}^{\gamma}_{\mathbf{z}_i}(h_i) + \varepsilon(n, m, \delta, \gamma),$$

where

$$\varepsilon(n, m, \delta, \gamma) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4m}{\delta\gamma}\right) + \frac{4m}{n}\log\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/4, n + n/m\right)\right)\right\}.$$

5.3 Discussion

Comparison with the binary cross-validation bounds

If we ignore the margin of classification and simply use the sign $\mathrm{sgn}(h)$ of $h : X \to [-1, 1]$, then Theorem 4 will provide bounds on the hold-out and cross-validation estimates, with $\mathcal{H}$ replaced by $\mathrm{sgn}(\mathcal{H}) = \{\mathrm{sgn}(h) : h \in \mathcal{H}\}$. These bounds will involve the growth function of the class $\mathrm{sgn}(\mathcal{H})$. If $\mathrm{sgn}(\mathcal{H})$ has finite VC-dimension then, by Sauer's Lemma, this growth function will be subexponential and the estimation error can be made arbitrarily small by choosing n large enough. On the other hand, taking account of the margin and using the results just given, then the relevant bounds involve the covering numbers. In many cases, these are subexponential even when $\mathrm{sgn}(\mathcal{H})$ has infinite dimension. (This will happen if $\mathrm{sgn}(\mathcal{H})$ has infinite VC-dimension but $\mathcal{H}$ has finite fat-shattering dimension–see later.) Thus the new results provide useful bounds on the error in situations where the previous bounds tell us nothing at all.

Sanity-checking and existential anxiety

Comparing our new bounds to ones which apply to an appropriate resubstitution error, however, makes it clear that our bounds are 'sanity-check' bounds, to use the terminology of Kearns and Ron [20]. Explicitly, it has been shown in [7, 8, 24] that the following holds for $\gamma, \delta \in (0, 1)$: with probability at least $1 - \delta$, if $\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$ then

$$\mathrm{er}_P(h) < \frac{4}{n} \ln\left(\frac{4}{\delta}\right) + \frac{4}{n} \ln\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, 2n\right)\right).$$

Thus, the γ-resubstitution estimate $\mathrm{er}^{\gamma}_{\mathbf{z}}(h)$ has a smaller estimation error than we have been able to obtain for the cross-validation estimator. The same is true of the binary-valued case (see Holden [16]). The looseness in our bound is a result of the way we derive it from a probability bound. Explicitly, we bound the probability $\mathcal{P}_{\mathrm{CV}}$ of the set of $\mathbf{z} \in Z^n$ such that $\exists h_1, \ldots, h_m, h \in \mathcal{H}$ with $\mathrm{er}^{\gamma}_{\mathbf{z}\backslash\mathbf{z}_i}(h_i) = 0$ for $1 \leq i \leq m$, $\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$, and

$$\mathrm{er}_P(h) > \frac{1}{m} \sum_{i=1}^{m} \mathrm{er}^{\gamma}_{\mathbf{z}_i}(h_i) + \epsilon.$$

To obtain a bound for cross-validation, the existential quantifiers are not needed: that is, instead of asking whether there *exists* h_i and h with the required property, we more specifically question whether the particular functions $L(\mathbf{z}\backslash\mathbf{z}_i)$ and $L(\mathbf{z})$ have the property. That is to say, the functions $h_1, h_2, \ldots, h_m, h$ are not independently chosen but arise from L. Now, if $g \in \mathcal{H}$ is such that $\mathrm{er}^{\gamma}_{\mathbf{z}}(g) = 0$ and $\mathrm{er}_P(g) > \epsilon$, then, taking each h_i, and h, to equal g, we see that $\mathbf{z}$ belongs to the set just described. Thus,

$$\mathcal{P}_{\mathrm{CV}} \geq P^n \left\{\mathbf{z} : \exists g \in \mathcal{H}, \mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0, \mathrm{er}_P(h) > \epsilon\right\}.$$

The estimation error bounds for the γ-resubstitution error follow by bounding the right-hand probability from above, so it follows that, using the general method of this paper, it is impossible to obtain bounds which show that cross-validation is better than resubstitution. It might therefore be argued that the approach we take is over-general and that a different approach would yield better than mere 'sanity-check' bounds. This is an open problem. We suggest, however, that our probability bounds are of interest in their own right. In any case, our results are of use in that they provide reasonable bounds for cross-validation in situations where no previous bounds were known.

6 New estimators

6.1 Modified estimators

Theorem 8 and Theorem 9 allow one to compute a suitable γ on the basis of the performance of the learning algorithm in the training sample. Consider for the moment the holdout estimate. Suppose L is a learning algorithm and that a sample $\mathbf{z}$ is given. Let $\gamma(\mathbf{z})$ be the largest γ for which

$$\mathrm{er}_{\mathbf{z}\backslash\mathbf{z}_m}(L(\mathbf{z}\backslash\mathbf{z}_m)) = \mathrm{er}_{\mathbf{z}}(L(\mathbf{z})) = 0.$$

Then, using this value of γ, and applying Theorem 8, we have a high-probability bound on the true error of $L(\mathbf{z})$: explicitly, $\mathrm{er}_P(L(\mathbf{z}))$ is no more than $\varepsilon(n, m, \delta, \gamma)$ plus $\mathrm{er}^{\gamma}_{\mathbf{z}_m}(L(\mathbf{z}\backslash\mathbf{z}_m))$, the error of $L(\mathbf{z}\backslash\mathbf{z}_m)$ at margin γ on $\mathbf{z}_m$.

For $\mathbf{w} = ((x_1, y_1), \ldots, (x_k, y_k)) \in Z^k$ and $f \in H$, let

$$\mathrm{mar}(f, \mathbf{w}) = \min\{f(x_i)y_i : 1 \le i \le k\}.$$

We formalise the above as follows.

Definition 10 (Adaptive holdout estimate) *For $n > m \ge 1$, for a learning algorithm L, and for $\mathbf{z} \in Z^n$, let*

$$\mathrm{Hmar}(L, \mathbf{z}) = \min\left\{\mathrm{mar}\left(L(\mathbf{z}\backslash\mathbf{z}_m), \mathbf{z}\backslash\mathbf{z}_m\right), \mathrm{mar}\left(L(\mathbf{z}), \mathbf{z}\right)\right\}.$$

Then the adaptive holdout estimate $\mathrm{Mest}^m_{\mathrm{H}}(L, \mathbf{z})$ *is defined to be* 1 *if* $\mathrm{Hmar}(L, \mathbf{z}) \le 0$*, and otherwise is given by*

$$\mathrm{Mest}^m_{\mathrm{H}}(L, \mathbf{z}) = \mathrm{est}^m_{H,\mathrm{Hmar}(L,\mathbf{z})}(L, \mathbf{z}) = \mathrm{er}^{\mathrm{Hmar}(L,\mathbf{z})}_{\mathbf{z}_m}(L(\mathbf{z}\backslash\mathbf{z}_m)).$$

In a similar way, we define a modified cross-validation estimator.

Definition 11 (Adaptive cross-validation estimate) *For $n > m \ge 1$, for a learning algorithm L, and for $\mathbf{z} \in Z^n$, let*

$$\mathrm{CVmar}(L, \mathbf{z}) = \min\left\{\min_{1\le i\le m} \mathrm{mar}\left(L(\mathbf{z}\backslash\mathbf{z}_i), \mathbf{z}\backslash\mathbf{z}_i\right), \mathrm{mar}\left(L(\mathbf{z}), \mathbf{z}\right)\right\}.$$

Then the adaptive cross-validation estimate $\mathrm{Mest}^m_{\mathrm{CV}}(L, \mathbf{z})$ *is defined to be* 1 *if* $\mathrm{CVmar}(L, \mathbf{z}) \le 0$*, and otherwise is given by*

$$\mathrm{Mest}^m_{\mathrm{CV}}(L, \mathbf{z}) = \mathrm{est}^m_{CV,\mathrm{CVmar}(L,\mathbf{z})}(L, \mathbf{z}) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{er}^{\mathrm{CVmar}(L,\mathbf{z})}_{\mathbf{z}_i}(L(\mathbf{z}\backslash\mathbf{z}_i)).$$

Although the formal definitions for these two estimates are a little intricate, both estimates are simple to apply in practice; we merely obtain the maximum suitable value for γ by examining the behaviour of L for the training examples (step 1), and then use this value when calculating the estimates (step 2). There is a simple intuitive justification for estimating error in this way. Assume that the value of γ we find in the first step is in fact large. This means that in all cases L provides us with a hypothesis h which has learned the training examples, in such a way that $h(x_i)$ is always very close to y_i. Intuitively, we might reasonably expect this to indicate the presence of over-fitting. As over-fitting tends to result in poor generalization, a sensible course of action is then to make our estimate of error more demanding, in the sense that it will tend to produce a higher value than if no potential over-fitting had been detected. This is precisely the effect of using the large γ obtained in step 1 when we calculate the estimate in step 2.

6.2 Performance bounds

Using Theorem 8 and Theorem 9, we obtain the following bounds on the error of the modified estimators.

Theorem 12 *For $\delta \in (0,1)$ and for $n > m \geq 3$, where m divides n, with probability at least $1-\delta$,*

1. *if* $\mathrm{Hmar}(L,\mathbf{z}) > 0$*, then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}^m_{\mathrm{H}}(L,\mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4}{\delta\gamma}\right) + \frac{4m}{n}\log\left(\mathcal{N}_\infty\left(\mathcal{H},\gamma/4,n+n/m\right)\right)\right\},$$

   where $\gamma = \mathrm{Hmar}(L,\mathbf{z})$.

2. *if* $\mathrm{CVmar}(L,\mathbf{z}) > 0$*, then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}^m_{\mathrm{CV}}(L,\mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4m}{\delta\gamma}\right) + \frac{4m}{n}\log\left(\mathcal{N}_\infty\left(\mathcal{H},\gamma/4,n+n/m\right)\right)\right\},$$

   where $\gamma = \mathrm{CVmar}(L,\mathbf{z})$.

7 General 'dimension-based' bounds

In this section, we return to the general theory. The general bounds obtained earlier concern the covering numbers of the function class $\mathcal{H}$. It is often useful to have bounds in terms of certain dimensions associated with the function class, analogous to the VC-dimension for binary-valued classes.

A number of ways of measuring the 'expressive power' of a class $\mathcal{H}$ of functions have been found to be useful. This power can be quantified by associating a 'dimension' to the class. Sometimes this is simply one number depending on $\mathcal{H}$ and sometimes—in what is known as a *scale-sensitive dimension*—it is a function depending on $\mathcal{H}$.

An important example of the first type of dimension is the *pseudo-dimension* [14, 22]. We say that a finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *shattered* if there is $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \mathbf{R}^d$ such that for every $b = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_b \in \mathcal{H}$ with $h_b(x_i) > r_i$ if $b_i = 1$ and $h_b(x_i) < r_i$ if $b_i = 0$. The *pseudo-dimension* of $\mathcal{H}$, denoted $\mathrm{Pdim}(\mathcal{H})$, is the largest cardinality of a shattered set, or infinity if there is no bound on the cardinalities of the shattered sets. The pseudo-dimension is a well-understood and useful measure of expressive power. One attractive feature of this dimension is that if the set of functions is a vector space then its pseudo-dimension coincides with its linear dimension (see for example [14]).

The most important scale-sensitive dimension which has been used to date in the development of the theory of learning real-valued functions is the *fat-shattering dimension*. This is a scale-sensitive version of the pseudo-dimension and was introduced by Kearns and Schapire [21]. Suppose that $\mathcal{H}$ is a set of functions from X to $[-1, 1]$ and that $\gamma \in (0, 1)$. We say that a finite subset $S = \{x_1, x_2, \ldots, x_d\}$ of X is *γ-shattered* if there is $\mathbf{r} = (r_1, r_2, \ldots, r_d) \in \mathbf{R}^d$ such that for every $b = (b_1, b_2, \ldots, b_d) \in \{0,1\}^d$, there is a function $h_b \in \mathcal{H}$ with $h_b(x_i) \geq r_i + \gamma$ if $b_i = 1$ and $h_b(x_i) \leq r_i - \gamma$ if $b_i = 0$. Thus, S is γ-shattered if it is shattered with a 'width of shattering' of at least γ. We define the *fat-shattering dimension*, $\mathrm{fat}_{\mathcal{H}} : \mathbf{R}^+ \to \mathbf{N}_0 \cup \{\infty\}$, as

$$\mathrm{fat}_{\mathcal{H}}(\gamma) = \max \left\{ |S| : S \subseteq X \text{ is } \gamma\text{-shattered by } \mathcal{H} \right\},$$

or $\mathrm{fat}_{\mathcal{H}}(\gamma) = \infty$ if the maximum does not exist. (Here, $\mathbf{N}_0$ denotes the set of nonnegative integers.) It is easy to see that $\mathrm{Pdim}(\mathcal{H}) = \lim_{\gamma \to 0} \mathrm{fat}_{\mathcal{H}}(\gamma)$. It should be noted, however, that it is possible for the pseudo-dimension to be infinite, even when $\mathrm{fat}_{\mathcal{H}}(\gamma)$ is finite for all γ. We shall say that $\mathcal{H}$ has *finite fat-shattering dimension* whenever it is the case that for all $\gamma \in (0, 1)$, $\mathrm{fat}_{\mathcal{H}}(\gamma)$ is finite.

The following result is due to Alon *et al.* [1]. It bounds the d_∞-covering number of $\mathcal{H}$ in terms of the fat-shattering function of $\mathcal{H}$.

Theorem 13 *Suppose that $\mathcal{H}$ is a set of real functions from a domain X to the bounded interval $[-1, 1]$ and that $\epsilon > 0$. Then*

$$\mathcal{N}_\infty(\mathcal{H}, \epsilon, m) < 2\left(mb^2\right)^{\lceil \log_2 y \rceil} < 2\left(\frac{16m}{\epsilon^2}\right)^{d \log_2(4em/(d\epsilon))},$$

where

$$b = \left\lfloor \frac{4}{\epsilon} \right\rfloor,$$

and, with $d = \mathrm{fat}_{\mathcal{H}}(\epsilon/4)$,

$$y = \sum_{i=1}^{d} \binom{m}{i} b^i.$$

Using Theorem 8 and Theorem 9 we obtain the following bounds in terms of the fat-shattering dimension. (The proofs are omitted.)

Theorem 14 *For $\delta \in (0, 1)$ and for $n > m \geq 3$, where m divides n, with probability at least $1 - \delta$,*

1. *for all* $\gamma \in (0,1)$, *if* $h_m, h \in \mathcal{H}$ *and* $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_m}(h_m) = \mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$ *then*

$$\mathrm{er}_P(h) < \mathrm{er}^{\gamma}_{\mathbf{z}_m}(h_m) + \varepsilon(n, m, \delta, \gamma),$$

   where

$$\varepsilon(n, m, \delta, \gamma) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4}{\delta\gamma}\right) + \frac{16m}{n}\mathrm{fat}_H(\gamma/16)\log^2\left(\frac{23n}{\gamma}\right)\right\}.$$

2. *for all* $\gamma \in (0,1)$, *if* $h_1, \ldots, h_m, h \in \mathcal{H}$ *and* $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_i}(h_i) = 0$ *for* $1 \le i \le m$, *and* $\mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$, *then*

$$\mathrm{er}_P(h) < \frac{1}{m}\sum_{i=1}^{m}\mathrm{er}^{\gamma}_{\mathbf{z}_i}(h_i) + \varepsilon(n, m, \delta, \gamma),$$

   where

$$\varepsilon(n, m, \delta, \gamma) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4m}{\delta\gamma}\right) + \frac{16m}{n}d\log^2\left(\frac{23n}{\gamma}\right)\right\}.$$

In particular, for the modified holdout and cross-validation estimators, we have the following results.

Theorem 15 *For* $n > m \ge 3$, *where* m *divides* n *and* $n \ge 2m/\epsilon^2$, *and with the notation as above, we have that, with probability at least* $1 - \delta$,

1. *if* $\mathrm{Hmar}(L, \mathbf{z}) > 0$, *then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}^{m}_{\mathrm{H}}(L, \mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4}{\delta\gamma}\right) + \frac{16m}{n}d\log^2\left(\frac{23n}{\gamma}\right)\right\},$$

   where $\gamma = \mathrm{Hmar}(L, \mathbf{z})$ *and where* $d = \mathrm{fat}_{\mathcal{H}}(\mathrm{Hmar}(L, \mathbf{z})/16)$.

2. *if* $\mathrm{CVmar}(L, \mathbf{z}) > 0$, *then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}^{m}_{\mathrm{CV}}(L, \mathbf{z}) + \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n}\log\left(\frac{4m}{\delta\gamma}\right) + \frac{16m}{n}d\log^2\left(\frac{23n}{\gamma}\right)\right\},$$

   where $\gamma = \mathrm{CVmar}(L, \mathbf{z})$ *and* $d = \mathrm{fat}_{\mathcal{H}}(\gamma/16)$.

As indicated earlier, these results apply in many cases where the binary results applied to $\mathrm{sgn}(\mathcal{H})$ tell us nothing. Explicitly, if $\mathcal{H}$ has finite fat-shattering dimension but $\mathrm{VCdim}(\mathrm{sgn}(\mathcal{H}))$ is infinite, then the binary results are useless. (There are many such classes. For example, the set of functions $f : [-1, 1] \to [-1, 1]$ of total variation at most 1 is such a class.)

8 Application to neural networks

We now use bounds of Bartlett [4, 7, 8] on the fat-shattering dimension of function classes computable by certain feedforward sigmoid neural networks. We deal here with networks having as activation function $\sigma(y) = \tanh(y/2) = 1 - 1/(1 + e^{-y})$ and, for convenience, we assume that all thresholds are set at 0. (There are similar results for other types of sigmoid function, and for nets with thresholds, but to illustrate the results as concisely as possible, we shall make these assumptions.) We define

$$F_0 = \{x \mapsto x_i : x = (x_1, \ldots, x_n) \in [-1,1]^r,\ i \in \{1, \ldots, r\}\},$$

and, for $i \geq 1$, we let

$$F_i = \left\{ \sigma\left(\sum_{j=1}^{N} w_j f_j\right) : N \in \mathbf{N},\ f_j \in \bigcup_{k=0}^{i-1} F_k,\ \sum_{j=1}^{N} |w_j| \leq V \right\}.$$

Thus, $\mathcal{H} = F_\ell$ is the class of functions that can be computed by an ℓ-layer feed-forward network in which each unit has the sum of the magnitudes of its weights bounded by V.

Bartlett has shown ([4], following [7, 8]) that

$$\mathrm{fat}_{\mathcal{H}}(\gamma) \leq \frac{1}{6}\left(\frac{48}{\gamma}\right)^{2\ell} (2V)^{\ell(\ell+1)} \log(2r+2),$$

provided $V \geq 1/2$, and $\gamma \leq V$.

We therefore have the following result, which follows from the results of the previous section.

Theorem 16 *Let $\delta \in (0,1)$ and suppose $n > m \geq 3$, where m divides n. Let $\mathcal{H}$ be the class of functions computable by the type of neural network just described (with weight-bound V and r inputs), and assume that $V \geq 1/2$. Then, for any learning algorithm L, with probability at least $1 - \delta$:*

1. *if* $\mathrm{Hmar}(L, \mathbf{z}) > 0$, *then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}_{\mathrm{H}}^{m}(L, \mathbf{z}) + \epsilon(n, m, \delta, \gamma),$$

where $\epsilon(n, m, \delta, \gamma)$ *is*

$$\max\left\{\sqrt{\frac{2m}{n}},\ \frac{2m}{n}\log\left(\frac{4}{\delta\gamma}\right) + \frac{8m}{3n}\left(\frac{768}{\gamma}\right)^{2\ell}(2V)^{\ell(\ell+1)}\log(2r+2)\log^2\left(\frac{23n}{\gamma}\right)\right\},$$

2. *if* $\mathrm{CVmar}(L, \mathbf{z}) > 0$, *then*

$$\mathrm{er}_P(L(\mathbf{z})) < \mathrm{Mest}_{\mathrm{CV}}^{m}(L, \mathbf{z}) + \epsilon(n, m, \delta, \gamma),$$

where $\epsilon(n, m, \delta, \gamma)$ *is*

$$\max\left\{\sqrt{\frac{2m}{n}},\ \frac{2m}{n}\log\left(\frac{4m}{\delta\gamma}\right) + \frac{8m}{3n}\left(\frac{768}{\gamma}\right)^{2\ell}(2V)^{\ell(\ell+1)}\log(2r+2)\log^2\left(\frac{23n}{\gamma}\right)\right\},$$

where $\gamma = \mathrm{CVmar}(L, \mathbf{z})$.

9 Proofs

9.1 Relating cross-validation estimate to holdout estimate

We start by defining, for any $\mathbf{z}$, the following two sets of hypothesis sequences:

$$\mathcal{H}^m_{H,\gamma}(\mathbf{z}) = \left\{ \mathbf{h} = (h_m, h) \in \mathcal{H}^2 : \mathrm{er}^\gamma_{\mathbf{z}\backslash\mathbf{z}_m}(h_m) = 0, \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0 \right\}$$

and

$$\mathcal{H}^m_{CV,\gamma}(\mathbf{z}) = \left\{ \mathbf{h} = (h_1, \ldots, h_m, h) \in \mathcal{H}^{m+1} : \mathrm{er}^\gamma_{\mathbf{z}\backslash\mathbf{z}_i}(h_i) = 0 \text{ for } 1 \leq i \leq m \text{ and } \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0. \right\}$$

We then define the sets

$$Q^{\epsilon,\gamma}_{n,m} = \left\{ \mathbf{z} \in Z^n : \exists \mathbf{h} \in \mathcal{H}^m_{H,\gamma}(\mathbf{z}) \text{ such that } \mathrm{er}_P(h) > \mathrm{er}_{\mathbf{z}_m}(h_m) + \epsilon \right\}$$

and

$$R^{\epsilon,\gamma}_{n,m} = \left\{ \mathbf{z} \in Z^n : \exists \mathbf{h} \in \mathcal{H}^m_{CV,\gamma}(\mathbf{z}) \text{ such that } \mathrm{er}_P(h) > \mathrm{er}_{\mathbf{z}}(h_1, \ldots, h_m) + \epsilon \right\}.$$

As in the corresponding analysis of [16], an upper bound for $P^n(Q^{\epsilon,\gamma}_{n,m})$ immediately implies one for $P^n(R^{\epsilon,\gamma}_{n,m})$, as we now show.

Lemma 17 *For any P, any $n \geq 2$, any $\gamma \in (0,1)$, any m such that $m \geq 2$ and m divides n, and any $\epsilon \in (0,1]$,*

$$P^n(R^{\epsilon,\gamma}_{n,m}) \leq m P^n(Q^{\epsilon,\gamma}_{n,m}).$$

Proof Recalling the definition of the set $R^{\epsilon,\gamma}_{n,m}$,

$$\begin{aligned}
R^{\epsilon}_{n,m} &= \left\{ \mathbf{z} \in Z^n : \exists \mathbf{h} \in \mathcal{H}^m_{CV,\gamma}(\mathbf{z}) \text{ such that } \mathrm{er}_P(h) > \mathrm{er}_{\mathbf{z}}(h_1, \ldots, h_m) + \epsilon \right\} \\
&= \left\{ \mathbf{z} \in Z^n : \exists \mathbf{h} \in \mathcal{H}^m_{CV,\gamma}(\mathbf{z}) \text{ such that } \mathrm{er}_P(h) > \frac{1}{m} \sum_{i=1}^m \mathrm{er}^\gamma_{\mathbf{z}_i}(L(\mathbf{z}\backslash\mathbf{z}_i)) + \epsilon \right\} \\
&\subseteq \left\{ \mathbf{z} \in Z^n : \exists \mathbf{h} \in \mathcal{H}^m_{CV,\gamma}(\mathbf{z}) \text{ s.t. } \mathrm{er}_P(h) > \mathrm{er}^\gamma_{\mathbf{z}_i}(L(\mathbf{z}\backslash\mathbf{z}_i)) + \epsilon \text{ for some } i, 1 \leq i \leq m \right\} \\
&\subseteq \bigcup_{i=1}^m A^{\epsilon,\gamma}_{n,m}(i)
\end{aligned}$$

where the sets $A^{\epsilon,\gamma}_{n,m}(i)$ are defined as

$$A^{\epsilon,\gamma}_{n,m}(i) = \left\{ \mathbf{z} \in Z^n : \exists h, h_i \in \mathcal{H} \text{ with } \mathrm{er}^\gamma_{\mathbf{z}\backslash\mathbf{z}_i}(h_i) = 0, \mathrm{er}^\gamma_{\mathbf{z}}(h) = 0, \;\; \mathrm{er}_P(h) > \mathrm{er}_{\mathbf{z}_i}(h_i) + \epsilon \right\}$$

for $1 \leq i \leq m$. Consequently,

$$P^n(R^{\epsilon,\gamma}_{n,m}) \leq P^n\left(\bigcup_{i=1}^m A^{\epsilon,\gamma}_{n,m}(i) \right) \leq \sum_{i=1}^m P^n(A^{\epsilon,\gamma}_{n,m}(i)).$$

By symmetry, we have $P^n(A^{\epsilon,\gamma}_{n,m}(i)) = P^n(Q^{\epsilon,\gamma}_{n,m})$ for $1 \leq i \leq m$, and the required result follows.
□

9.2 Symmetrization

We will present the proof of Theorem 6 using a series of lemmas. The techniques are similar to those used in [16] and [5], which in turn extend the standard techniques of symmetrization and combinatorial bounding used in [3, 10, 28]. For notational convenience we will divide any $\mathbf{z} \in Z^n$ into two parts and write $\mathbf{z} = \mathbf{z}'\mathbf{z}_m$ where $\mathbf{z}' = \mathbf{z}\backslash\mathbf{z}_m = (z_1, \ldots, z_{n-n'})$ and as usual $\mathbf{z}_m = (z_{n-n'+1}, \ldots, z_n)$. We will also introduce a further sequence $\mathbf{w} = (w_1, \ldots, w_k) \in Z^k$, where the value of k will be fixed below. Define the set

$$S_{n,m}^{\epsilon,\gamma}(k) = \left\{\mathbf{z}'\mathbf{z}_m\mathbf{w} \in Z^{n+k} : \exists \mathbf{h} \in \mathcal{H}_{H,\gamma}^m(\mathbf{z}) \text{ such that } \mathrm{er}_{\mathbf{w}}(h) > \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon/2\right\}.$$

Note that the error measure $\mathrm{er}_{\mathbf{w}}(h)$ in this definition is the (standard) empirical error; that is, the proportion of examples (x_i, y_i) in $\mathbf{w}$ such that $\mathrm{sgn}(h(x_i)) \neq y_i$.

Lemma 18 *For any $n \geq 2$, any m such that $m \geq 2$ and m divides n, any ϵ such that $0 < \epsilon \leq 1$, any $\gamma \in (0,1)$, any distribution P, and any $k \geq 2/\epsilon^2$,*

$$P^n(Q_{n,m}^{\epsilon,\gamma}) \leq 2P^{n+k}(S_{n,m}^{\epsilon,\gamma}(k)).$$

Proof If $h \in \mathcal{H}$ satisfies $\mathrm{er}_P(h) > \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon$ and $\mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2$, then

$$\mathrm{er}_{\mathbf{w}}(h) > \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon/2.$$

so

$$\begin{aligned} P^{n+k}(S_{n,m}^{\epsilon,\gamma}) &\geq P^{n+k}\left(\left\{\mathbf{z}\mathbf{w} : \exists \mathbf{h} \in \mathcal{H}_{H,\gamma}^m(\mathbf{z}),\ \mathrm{er}_P(h) \geq \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon \text{ and } \mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2\right\}\right) \\ &= \int_{Z^n} I_Q(\mathbf{z})\, P^k\left(\left\{\mathbf{w} : \exists h \in \mathcal{H},\ \mathrm{er}_P(h) \geq \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon \text{ and}\right.\right. \\ &\qquad\qquad \left.\left. \mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2\right\}\right)\, dP^n(\mathbf{z}), \end{aligned} \tag{1}$$

where I_Q denotes the indicator (or characteristic) function of the set $Q_{n,m}^{\epsilon,\gamma}$.

Now, for $\mathbf{z} \in Q_{n,m}^{\epsilon,\gamma}$ fix an $h \in \mathcal{H}$ with $\mathrm{er}_P(h) \geq \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon$. For this h, we shall show that for $k \geq 2/\epsilon^2$,

$$P^k\left(\{\mathbf{w} : \mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2\}\right) \geq 1/2. \tag{2}$$

It will then follow that, for any $\mathbf{z} \in Q = Q_{n,m}^{\epsilon,\gamma}$, we have

$$P^k\left(\left\{\mathbf{w} : \exists h \in \mathcal{H},\ \mathrm{er}_P(h) \geq \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon \text{ and } \mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2\right\}\right) \geq 1/2,$$

and combining this result with (1) shows that $P^{2m}(R) \geq P^m(Q)/2$.

To complete the proof, we show that (2) holds for any $h \in \mathcal{H}$. For a fixed h, notice that $k\,\mathrm{er}_{\mathbf{w}}(h)$ is a binomial random variable, with expectation $k\,\mathrm{er}_P(h)$ and variance $\mathrm{er}_P(h)(1-\mathrm{er}_P(h))k$. Chebyshev's inequality bounds the probability that $\mathrm{er}_{\mathbf{w}}(h) \geq \mathrm{er}_P(h) - \epsilon/2$ by

$$\frac{\mathrm{er}_P(h)(1-\mathrm{er}_P(h))k}{(\epsilon k/2)^2},$$

which is less than $1/(\epsilon^2 k)$ (using the fact that $x(1-x) \leq 1/4$ for x between 0 and 1). This is at most $1/2$ for $k \geq 2/\epsilon^2$, which implies (2). □

9.3 Using a group action

Our next task is to seek an upper bound on $P^{n+k}(S^{\epsilon,\gamma}_{n,m}(k))$. We will do this using the next lemma, which is a minor variation on a standard result. Let G_i be the group of all permutations of $\{1,\ldots,i\}$ and let σ be a permutation in G_i. For any sequence $\mathbf{a} = (a_1,\ldots,a_i)$ we denote by $\sigma\mathbf{a}$ the sequence $(a_{\sigma(1)},\ldots,a_{\sigma(i)})$.

Lemma 19 (see [10], and [3], Lemma 8.3.3) *Let S be a subset of S^{n+k} and let P be any distribution on Z. Let $\mathbf{zw}$ be a sequence in Z^{n+k} and let G be any subgroup of G_{n+k}. Then,*

$$P^{n+k}(S) \leq \frac{1}{|G|} \max_{\mathbf{zw} \in Z^{n+k}} N(S, \mathbf{zw})$$

where $N(S,\mathbf{zw})$ is defined as

$$N(S,\mathbf{zw}) = |\{\sigma \in G : \sigma\mathbf{zw} \in S\}|.$$

In the light of lemma 19, our next step is therefore to obtain a bound on $N(S^{\epsilon,\gamma}_{n,m}(k), \mathbf{zw})$ that is independent of $\mathbf{zw}$, using a suitable subgroup G. For any $(h_i, h_j) \in \mathcal{H}^2$, define the set,

$$S(h_i,h_j) = \left\{\mathbf{zw} \in Z^{n+k} : \mathrm{er}^{\gamma}_{\mathbf{z}}(h_i) = 0, \mathrm{er}^{\gamma}_{\mathbf{z}'}(h_j) = 0 \text{ and } \mathrm{er}_{\mathbf{w}}(h_i) \geq \mathrm{er}^{\gamma}_{\mathbf{z}_m}(h_j) + \frac{\epsilon}{2}\right\}$$

where as usual $\mathbf{z} = \mathbf{z}'\mathbf{z}_m$. We now require that $m \geq 3$, and we choose $k = n' = (n/m)$. We define G as follows. Consider the section $\mathbf{z}_{m-2}\mathbf{z}_{m-1}\mathbf{z}_m\mathbf{w}$ of the sequence $\mathbf{s} = \mathbf{zw}$ where $\mathbf{s} = (z_1,\ldots,z_n,z_{n+1},\ldots,z_{n+n'})$. For any i in the range $(n-3n'+1) \leq i \leq (n-n')$ let σ_i be the permutation of $\{1,\ldots,n+n'\}$ having the following properties:

1. $\sigma_i(j) = j$ for $j \neq i$ and $j \neq i + 2n'$.
2. $\sigma_i(i) = i + 2n'$ and $\sigma_i(i+2n') = i$.

Consequently, if $\mathbf{s} = \mathbf{zw}$ then each σ_i, when used to obtain $\sigma_i\mathbf{s}$, swaps an element in the $\mathbf{z}_{m-2}\mathbf{z}_{m-1}$ section with a corresponding element in the $\mathbf{z}_m\mathbf{w}$ section. This is illustrated in Figure 2. We define G to be the subgroup of $G_{n+n'}$ generated by the set of permutations $\{\sigma_i\}$ for i in the range defined above. Clearly $|G| = 2^{2n'}$.

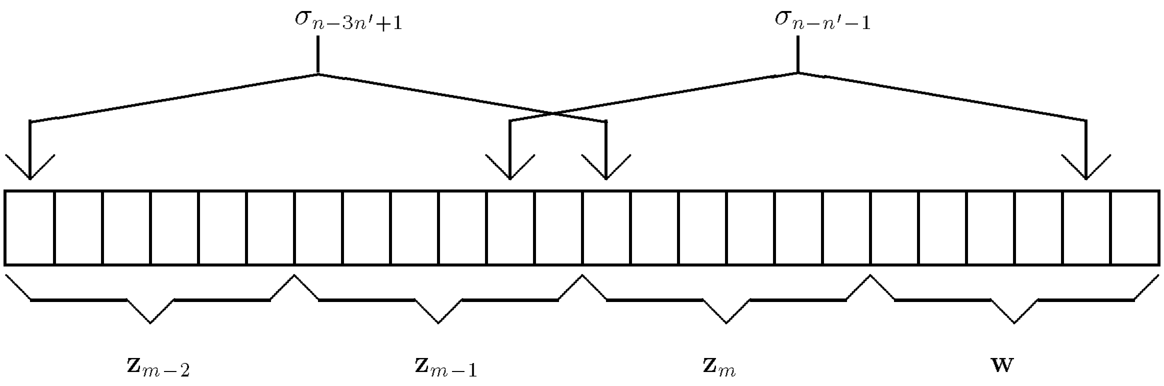

Figure 2: Illustration of the effect of the permutations σ_i when applied to $\mathrm{s} = \mathbf{zw}$.

9.4 Using covers

We now show how the problem of bounding $N(S, \mathbf{zw})$ can be reduced to a more manageable problem through the use of covers. Fix $\mathrm{s} = \mathbf{zw} \in Z^{n+n'}$ and suppose that $\gamma \in (0,1)$. Suppose that $\mathrm{s} = (z_1, \ldots, z_n, z_{n+1}, \ldots, z_{n+n'})$ where $z_i = (x_i, y_i) \in Z$. Let $\mathcal{C}$ be a $\gamma/2$-cover (in the d_∞ sense) of $\mathcal{H}_{|\mathbf{x}}$, where $\mathrm{x} = (x_1, x_2, \ldots, x_{n+n'})$, and suppose that $\mathcal{C}$ has minimum possible cardinality, $\mathcal{N}_\infty(\mathcal{H}, \gamma/2, n+n')$. Then, for each $h \in \mathcal{H}$ there is $\hat{h} \in \mathcal{C}$ such that for $1 \le i \le n+n'$, $|h(x_i) - \hat{h}(x_i)| < \gamma/2$. Suppose that s belongs to $S = S_{n,m}^{\epsilon,\gamma}(n')$. Then, by definition, there are $h, h_m \in \mathcal{H}$ such that

$$\mathrm{er}_{\mathbf{z}'}^{\gamma}(h_m) = 0, \quad \mathrm{er}_{\mathbf{z}}^{\gamma}(h) = 0, \quad \mathrm{er}_{\mathbf{w}}(h) > \mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) + \epsilon/2,$$

where, as usual, $\mathbf{z} = \mathbf{z}'\mathbf{z}_m$. If $g \in \mathcal{H}$ satisfies $y_i g(x_i) \ge \gamma$ then, for $\hat{g}$ a member of $\mathcal{C}$ which is $\gamma/2$-close to g, we have

$$\hat{g}(x_i)y_i \ge g(x_i)y_i - |\hat{g}(x_i) - g(x_i)| > \gamma - \gamma/2 = \gamma/2.$$

Thus the conditions $\mathrm{er}_{\mathbf{z}'}^{\gamma}(h_m) = 0$ and $\mathrm{er}_{\mathbf{z}}^{\gamma}(h) = 0$ imply that

$$\mathrm{er}_{\mathbf{z}'}^{\gamma/2}(\hat{h}_m) = \mathrm{er}_{\mathbf{z}}^{\gamma/2}(\hat{h}) = 0,$$

where $\hat{h}_m \in \mathcal{C}$ is $\gamma/2$-close to h_m, and $\hat{h} \in \mathcal{C}$ is $\gamma/2$-close to h. Now, if $\mathrm{sgn}(g(x_i)) \ne y_i$ then $g(x_i)y_i < 0$ and $\hat{g}(x_i)y_i < 0 + \gamma/2$, Thus,

$$\mathrm{er}_{\mathbf{w}}^{\gamma/2}(\hat{h}) \ge \mathrm{er}_{\mathbf{w}}(h).$$

Furthermore, if $\hat{g}(x_i)y_i < \gamma/2$ then $g(x_i)y_i < \gamma/2 + \gamma/2 = \gamma$. Therefore,

$$\mathrm{er}_{\mathbf{z}_m}^{\gamma}(h_m) \ge \mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(\hat{h}_m).$$

These observations show that if $\mathrm{s} = \mathbf{z}'\mathbf{z}_m\mathrm{w} \in S$ then there exist $\hat{h}, \hat{h}_m \in \mathcal{C}$ such that

$$\mathrm{er}_{\mathbf{z}'}^{\gamma/2}(\hat{h}_m) = 0, \quad \mathrm{er}_{\mathbf{z}}^{\gamma/2}(\hat{h}) = 0, \quad \mathrm{er}_{\mathbf{w}}^{\gamma/2}(\hat{h}) > \mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(\hat{h}_m) + \epsilon/2.$$

It follows that

$$N(S, \mathbf{zw}) \leq N(\hat{S}, \mathbf{zw}), \tag{3}$$

where

$$\hat{S} = \left\{\mathbf{z}'\mathbf{z}_m\mathbf{w} : \exists f, g \in \mathcal{C} \text{ s.t. } \mathrm{er}_{\mathbf{z}'}^{\gamma/2}(f) = 0, \ \ \mathrm{er}_{\mathbf{z}}^{\gamma/2}(g) = 0, \ \ \mathrm{er}_{\mathbf{w}}^{\gamma/2}(g) > \mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(f) + \epsilon/2\right\}.$$

Now,

$$\begin{aligned}
N(\hat{S}, \mathrm{s}) &= \left|\{\sigma \in G : \sigma\mathrm{s} \in \hat{S}\}\right| \\
&\leq \left|\bigcup_{(f,g)\in\mathcal{C}^2} \left\{\sigma \in G : \sigma\mathrm{s} \in \hat{S}(f,g)\right\}\right| \\
&\leq \sum_{(f,g)\in\mathcal{C}^2} \left|\left\{\sigma \in G : \sigma\mathrm{s} \in \hat{S}(f,g)\right\}\right|, \qquad (4)
\end{aligned}$$

where

$$\hat{S}(f,g) = \left\{\mathbf{z}'\mathbf{z}_m\mathbf{w} : \mathrm{er}_{\mathbf{z}'}^{\gamma/2}(f) = 0, \ \ \mathrm{er}_{\mathbf{z}}^{\gamma/2}(g) = 0, \ \ \mathrm{er}_{\mathbf{w}}^{\gamma/2}(g) > \mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(f) + \epsilon/2\right\}.$$

We now bound the quantity in the summation of (4) independently of s and (f, g).

Lemma 20 *With* s, f, g *as above,*

$$|\{\sigma \in G : \sigma\mathrm{s} \in \hat{S}(f,g)\}| \leq 2^{2n' - n'\epsilon/2}. \tag{5}$$

Proof Denote by $N(f, g, \mathrm{s})$ the quantity on the left of (5), and suppose it is nonzero (for otherwise, the bound certainly holds). Then there is at least one $\sigma \in G$ such that $\sigma\mathrm{s} = \sigma\mathbf{zw} \in \hat{S}(f, g)$. Since G is a group, $N(f, g, \mathrm{s})$ is then equal to the number of permutations $\sigma' \in G$ for which $(\sigma'\sigma)\mathrm{s} \in \hat{S}(f, g)$ and consequently we can assume that $\mathrm{s} \in \hat{S}(f, g)$.

Consider any permutation $\sigma \in G$. When used to obtain $\sigma\mathbf{zw}$ this permutation has the effect of swapping pairs of elements, using combinations of the individual permutations σ_i described above. Whether $\sigma\mathbf{zw} \in \hat{S}(f, g)$ depends on which pairs are swapped. If σ swaps any pair including an (x_i, y_i) in $\mathbf{z}_m$ for which $f(x_i)y_i < \gamma/2$, then $\sigma\mathbf{zw} \notin \hat{S}(f, g)$. Similarly, if σ swaps any pair including an (x_i, y_i) in w such that $g(x_i)y_i < \gamma/2$ then $\sigma\mathbf{zw} \notin \hat{S}(f, g)$. Consequently,

$$N(f, g, \mathrm{s}) \leq 2^{n_1 + n_2}$$

where $n_1 = n' - n'\mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(f)$ is the number of entries (x_i, y_i) of $\mathbf{z}_m$ for which $f(x_i)y_i \geq \gamma/2$, and $n_2 = n' - n'\mathrm{er}_{\mathbf{w}}^{\gamma/2}(g)$ is the number of (x_i, y_i) in w such that $g(x_i)y_i \geq \gamma/2$. Now, because $\mathrm{s} \in \hat{S}(f, g)$, we have

$$\mathrm{er}_{\mathbf{w}}^{\gamma/2}(g) > \mathrm{er}_{\mathbf{z}_m}^{\gamma/2}(f) + \epsilon/2,$$

so

$$n' - n_2 > n - n_1 + n'\epsilon/2;$$

hence $n_1 - n_2 > n'\epsilon/2$ and

$$n_1 + n_2 = 2n_1 - (n_1 - n_2) < 2n_1 - n'\epsilon/2 \leq 2n' - n'\epsilon/2.$$

It follows, as required, that

$$N(f, g, \mathbf{s}) \leq 2^{2n' - n'\epsilon/2}.$$

□

9.5 Proof of the main theorems

Theorem 6 follows on piecing together the results we have obtained. We have, by (4) and (5),

$$\begin{aligned} N(\hat{S}, \mathbf{s}) &\leq \sum_{(f,g) \in \mathcal{C}^2} 2^{2n' - n'\epsilon/2} \\ &\leq |\mathcal{C}|^2 \, 2^{2n' - n'\epsilon/2} \\ &= \left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n'\right)\right)^2 2^{2n' - n'\epsilon/2}. \end{aligned}$$

By Lemma 18, Lemma 19, and (3), we then have, for $k = n' = n/m \geq 2/\epsilon^2$,

$$\begin{aligned} P^n(Q^{\epsilon,\gamma}_{n,m}) &\leq 2P^{n+k}(S^{\epsilon,\gamma}_{n,m}(k)) \\ &\leq \frac{2}{|G|} \max_{\mathbf{s} \in Z^{n+k}} N(S, \mathbf{s}) \\ &\leq \frac{2}{|G|} \max_{\mathbf{s} \in Z^{n+k}} N(\hat{S}, \mathbf{s}) \\ &\leq \frac{2}{2^{2n'}} \left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n'\right)\right)^2 2^{2n' - n'\epsilon/2} \\ &= 2\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n'\right)\right)^2 2^{-n'\epsilon/2} \\ &= 2\left(\mathcal{N}_\infty\left(\mathcal{H}, \gamma/2, n + n/m\right)\right)^2 \, 2^{-n\epsilon/2m}, \end{aligned}$$

the required bound.

Theorem 7 follows immediately on using Lemma 17.

To obtain Theorem 8 and Theorem 9 we use a technique from [7]. We give the proof of Theorem 8, that of Theorem 9 being similar.

Proof of Theorem 8: For $0 < \alpha_1, \alpha_2, \delta < 1$, let $E(\alpha_1, \alpha_2, \delta)$ be the set

$$\{\mathbf{z} \in Z^n : \exists h_m, h \in \mathcal{H} \text{ s.t. } \mathrm{er}^{\alpha_2}_{\mathbf{z} \setminus \mathbf{z}_m}(h_m) = \mathrm{er}^{\alpha_2}_{\mathbf{z}}(h) = 0, \text{ and } \mathrm{er}_P(h) \geq \mathrm{er}^{\alpha_2}_{\mathbf{z}_m}(h_m) + \epsilon(n, m, \delta, \alpha_1)\},$$

where

$$\epsilon(n, m, \delta, \alpha_1) = \max\left\{\sqrt{\frac{2m}{n}}, \frac{2m}{n} \log\left(\frac{2}{\delta}\right) + \frac{4m}{n} \log\left(\mathcal{N}_\infty\left(\mathcal{H}, \alpha_1/2, n + n/m\right)\right)\right\}.$$

Theorem 6 asserts that $P^n(E(\alpha,\alpha,\delta)) < \delta$ for all α. Furthermore, as is easily checked, if $0 < \alpha_1 \le \alpha \le \alpha_2 < 1$ and $\delta_1 \le \delta$, then $E(\alpha_1,\alpha_2,\delta_1) \subseteq E(\alpha,\alpha,\delta)$. We may argue, as in [7], that

$$\begin{aligned}
P^n\left(\bigcup_{\gamma\in(0,1)} E(\gamma/2,\gamma,\delta\gamma/2)\right) &\le P^n\left(\bigcup_{i=0}^{\infty}\bigcup_{\gamma\in(2^{-(i+1)},2^{-i}]} E(\gamma/2,\gamma,\delta\gamma/2)\right)\\
&\le P^n\left(\bigcup_{i=0}^{\infty} E(2^{-(i+1)},2^{-(i+1)},\delta 2^{-(i+1)})\right)\\
&\le \sum_{i=0}^{\infty} \delta 2^{-(i+1)} = \delta,
\end{aligned}$$

and the result follows on observing that the conclusion of Theorem 8 is precisely that

$$P^n\left(\bigcup_{\gamma\in(0,1)} E(\gamma/2,\gamma,\delta\gamma/2)\right) \le \delta.$$

10 Conclusions and further work

In this paper we have presented new error estimation techniques, applicable when real-valued functions are being used for binary classification. We have derived results in probability theory in order to provide bounds on the estimation error of the new techniques. As discussed earlier in the paper, the probability results are in a sense stronger than one needs to analyse the cross-validation and other techniques for particular 'sensible' learning algorithms, and it is an open question as to whether better estimation error bounds can be found using different methods. One potential approach is to use the idea of *error-stability* explored in [20] and Holden [17].

One specific aspect of our bounds which it might be possible to improve concerns the $\sqrt{2m/n}$ term appearing in the estimation error bounds. This arises from the requirement in the proof of Lemma 18 that $k \ge 2/\epsilon^2$. We have been unable to replace this by a condition of the form $k \ge O(1/\epsilon)$, but if this were done then the square-root term could be dropped from the bounds. Perhaps one approach (used for the γ-resubstitution error in [7]) is to develop some result on *relative* deviation of error along the lines of that obtained by Vapnik [27] (see also [6]).

The new holdout and cross-validation techniques we have developed require that each training example be classified correctly, and with a margin, by the relevant function(s). It might be worth allowing a small number of misclassifications. To analyse this in the same vein, we would have to derive probability results similar to those considered here but in which (for the holdout case, for instance), the condition $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_m}(h_m) = \mathrm{er}^{\gamma}_{\mathbf{z}}(h) = 0$ is replaced by $\mathrm{er}^{\gamma}_{\mathbf{z}\setminus\mathbf{z}_m}(h_m), \mathrm{er}^{\gamma}_{\mathbf{z}}(h) \le \eta$ for some small η.

Quite apart from theoretical analysis of the performance of the new technique, further experimental work remains to be done to test its performance on real data, and this is something we plan to do.

Acknowledgements

Part of this work was carried out while the authors were guests of the Isaac Newton Institute for Mathematical Sciences, Cambridge University. We thank the organisers of the Neural Networks and Machine Learning programme. We are grateful to David Hand for helpful discussion.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the Symposium on Foundations of Computer Science*. IEEE Press, 1993.

[2] M. Anthony. Probabilistic 'generalization' of functions and dimension-based uniform convergence results. *Statistics and Computing*, 8(1): 5–14, 1998.

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

[4] M. Anthony and P. Bartlett. *Theory of Learning in Neural Networks* (working title). To be published by Cambridge University Press.

[5] M. Anthony and P. Bartlett. Function learning from interpolation. Extended abstract in *Proceedings EuroCOLT'95*, Springer-Verlag, pages 211–221, 1995.

[6] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1994.

[7] P. Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important that the size of the network. *IEEE Transactions on Information Theory*, 44(2): 525–536, 1998.

[8] P. Bartlett. For valid generalisation, the size of the weights is more important than the size of the network. In *Advances in Neural Information Processing Systems, 9*. Morgan Kaufmann, 1996.

[9] Avrim Blum, Adam Kalai and John Langford. Beating the hold-out: bounds for k-fold and progressive cross-validation. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory*, pages 203–208, 1999.

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

[11] B. Cheng and D. M. Titterington. Neural networks: a review from a statistical perspective. *Statistical Science*, 9(1):2–54, 1994.

[12] D. Cohn and G. Tesauro. How tight are the Vapnik-Chervonenkis bounds? *Neural Computation*, 4(2):249–269, 1992.

[13] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis*. John Wiley, 1973.

[14] D. Haussler, Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, Sept. 1992.

[15] S. B. Holden. PAC-like Upper Bounds for the Sample Complexity of Leave-One-Out Cross-Validation. In *Proceedings of the Ninth Annual Conference on Computational Learning Theory*, pages 41–50, ACM Press, 1996.

[16] S. B. Holden. Cross-validation and the PAC learning model. Research Note RN/96/64, Department of Computer Science, University College London, December 1996.

[17] S. B. Holden. On Algorithmic Stability and the Analysis of the Cross-Validation and Hold-out Estimates. Research Note RN/97/73, Department of Computer Science, University College London.

[18] S. B. Holden and M. Niranjan. On the practical applicability of VC dimension bounds. *Neural Computation*, 7(6):1265–1288, 1995.

[19] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 341–352. ACM Press, New York, NY, 1992.

[20] M. J. Kearns and D. Ron. Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. In *Proceedings of the 10th Annual Workshop on Computational Learning Theory*, pages 152–162. ACM Press, New York, NY, 1997.

[21] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings of the 31st Symposium on the Foundations of Computer Science*, pages 382–391. IEEE Computer Society Press, Los Alamitos, CA, 1990.

[22] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[23] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.

[24] J. Shawe-Taylor, P. Bartlett. R.C. Williamson, M. Anthony. Structural risk minimisation over data-dependent hierarchies. To appear, *IEEE Transactions on Information Theory*.

[25] G. T. Toussaint. Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory*, 20(4):472–479, 1974.

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

[27] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[28] 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.