An Improved Predictive Accuracy Bound for Averaging Classifiers

John Langford
School of Computer Science
Carnegie Mellon
jcl@cs.cmu.edu
`http://www.cs.cmu.edu/~jcl`

Matthias Seeger
Institute for Adaptive and Neural Computation
University of Edinburgh
seeger@dai.ed.ac.uk
`http://www.dai.ed.ac.uk/homes/seeger/`

Nimrod Megiddo
IBM Almaden Research Center
megiddo@almaden.ibm.com
`http://theory.stanford.edu/~megiddo/`

Abstract

We present an improved bound on the difference between training and test errors for voting classifiers. This improved averaging bound provides a theoretical justification for popular averaging techniques such as *Bayesian classification*, *Maximum Entropy discrimination*, *Winnow* and *Bayes point machines* and has implications for learning algorithm design.

1. Introduction

Averaging is a standard technique in applied machine learning for combining multiple classifiers to achieve greater accuracy. Examples include *Bayesian classification* [4], *boosting* [7], *bagging* [2], *Winnow* [13], *Maximum Entropy discrimination* [11], and *Bayes point machines* [9]. Despite the prevalence of this technique there is only weak theoretical justification so far for the practice. This paper provides a new stronger theoretical justification for the practice of averaging. In particular, we state and prove a bound on the gap between the training set error rate and the predictive error rate which improves as more hypotheses are averaged over.

Until 1998, theoretical bounds such as the Occam's razor bound [3] suggested that averaging was *wrong* because it increased the description length of the resulting hypothesis.[1] The Occam's razor bound *only suggests* that averaging *may* be bad since there is no corresponding lower bound. Schapire, Freund, Bartlett and Lee [16] showed a great improvement on the naive bound for an average-of-classifiers hypothesis. Loosely speaking, their margin bound states that if the average has a small empirical error rate (i.e., it is accurate on most training examples) and has a large "margin" (defined in 2.1), then its true error rate is also small. The proof itself works in a very intuitive manner by showing that the accuracy of a large margin classifier is close to the accuracy of a simple classifier, for which standard bounds are tight.

The problem with this result is that the value of the bound depends only on the empirical margin which does not necessarily improve with an average over a larger number of hypotheses. This bound suggests using the simple criteria: choose the average to maximize the margin. However, empirical results [8] indicate that this procedure is not optimal.

In this paper, we prove a new bound on the true error rate, which suggests a new optimization criterion, namely, optimize for a large margin *and* for a uniform average over as many hypotheses as possible.

The layout of this paper is as follows:

1. Discussion of the relationship with prior relevant results.
2. Development of the improved theoretical bound with an emphasis on simplicity.
3. A proof of the bound.
4. An example of the benefit of the new bound on a toy problem.
5. Discussion of implications of the new bound on prior work.

[1] We note, however, that it is the *minimum* description length that should be used in the bound.

2. The setting and important earlier results

2.1 The setting

We first explain the setting, which is the same as the one used in [16].

An input space $\mathcal{X}$ is given, where the members of $\mathcal{X}$ are also referred to as *examples*. The set $\mathcal{X} \times \{-1, 1\}$ is the space of *labeled* examples. A *base hypothesis* h is a mapping from the input space $\mathcal{X}$ into $\{-1, 1\}$. A (possibly infinite) space $\mathcal{H}$ (the hypothesis space) is given and the goal is to construct an *averaging classifier* $c : \mathcal{X} \to \{-1, 1\}$ as a weighted average of base hypotheses:

$$c(x) = \operatorname{sign} \sum_{j=1}^{k} q_j h_j(x) \quad (x \in \mathcal{X}) \ ,$$

where $q_j \geq 0$, $j = 1, \ldots, k$, and $\sum_{j=1}^{k} q_i = 1$. The fundamental assumption here is that labeled training examples are drawn independently, with replacement, from some probability distribution D over $\mathcal{X} \times \{-1, 1\}$. In all the theorems we discuss, D is assumed to be unknown to the procedure which constructs the classifier, and the results hold for all D. Probabilities and expectations over D will be denoted by the subscript D; for example, the true error of an averaging classifier is denoted by:

$$e_D(c) = E_D[I(c(x) \neq y)] = E_{(x,y) \sim D}[I(c(x) \neq y)]. \tag{1}$$

Here, $I(\cdot)$ is the indicator function, which is 1 if its argument evaluates to *true* and 0 otherwise. Probabilities with respect to D are written as $\Pr_D$.

With S, we denote a sample $\{(x_i, y_i) \,|\, i = 1, \ldots, m\}$ drawn independently and identically distributed (i.i.d.) from $D(x, y)$. The i.i.d assumption is the one fundamental assumption we make in this work. The subscript S denotes empirical expectation or probability over S, for example the empirical error of an averaging classifier is given by:

$$e_S(c) = E_S[I(c(x) \neq y)] = \frac{1}{m} \sum_{i=1}^{m} I(c(x_i) \neq y) \tag{2}$$

Probabilities with respect to S are written as $\Pr_S$.

2.2 Quantities used in the bound

The basic learning model needs to be augmented with a few definitions for the analysis.

Given a subset $\{h_1, \ldots, h_k\} \subseteq \mathcal{H}$, the set $\{q_1, \ldots, q_k\}$ can be interpreted as a probability distribution over the set $\mathcal{H}$.[2] This distribution will be denoted by Q. We will also often use the unsigned version of the classifier:

$$f(x) = E_{h \sim Q}[h(x)] = \sum_{j=1}^{k} q_j h_j(x) \ .$$

It is important to note that we make no assumption about how the weights $q_1, \ldots, q_k$ are obtained, so our results are applicable to many algorithms.

The derived bounds depend on the powerful concept of the *margin*, $t(x, y)$, of a labeled example with respect to a classifier, namely,

$$t(x, y) = y \sum_{j=1}^{k} q_j h_j(x) = y f(x) \ .$$

The margin is a quantitative measure of how decided the average is. Obviously, $-1 \leq t(x, y) \leq 1$. If $t(x, y) = 1$ (resp. $t(x, y) = -1$), then all the base hypotheses classify correctly (resp. incorrectly). When $t(x, y)$ is close to zero, the classifier is, in some sense, undecided. Note that $c(x) = y$ iff $t(x, y) > 0$.

2.3 Earlier results

The new averaging bound arises from improving one critical step in the proof of the original margin bound, which we state here for reference.

Theorem 1 (Margin Bound [16]) *Let $\delta \in (0, 1)$. With probability at least $1 - \delta$ over random samples S from D we have that for all distributions $Q = (q_1, \ldots, q_k)$ over the finite hypothesis space $\mathcal{H}$ and all $\theta \in (0, 1]$:*

$$\Pr_D[yf(x) \leq 0] \leq \Pr_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{\theta^{-2} \ln |\mathcal{H}| \log m + \ln \delta^{-1}}{m}}\right), \tag{3}$$

where $f(x) = E_{h \sim Q}[h(x)] = \sum_j q_j h_j(x)$.

Here, the notation $b(m) = O(a(m))$ means there exists a constant C such that $b(m) \leq C \cdot a(m)$ for all m. This margin bound implies that if most training examples have a large margin θ (i.e. $t(x, y) > \theta$ for most $(x, y) \in S$) and the hypothesis space is not too large, then the generalization error cannot be large.

[2] In the case of finite or countably infinite $\mathcal{H}$, this is achieved by assigning all hypotheses outside the subset the weight zero. If $\mathcal{H}$ is finite, we will usually work with $\mathcal{H} = \{h_1, \ldots, h_k\}$ for simplicity. For uncountable spaces, we define Q as $\sum_j q_j \delta(h, h_j)$, where $\delta(h, h_j)$ is the delta distribution centered on h_j.

To improve on this bound, we employ a PAC-Bayes bound from McAllester [14]. In the PAC-Bayes setting, a classifier is also defined by a distribution Q over the hypothesis space. However, each classification[3] is carried out according to a hypothesis sampled from Q rather than by the averaging classifier c defined by Q. We are interested in the gap between the *expected* generalization error and the *expected* empirical error, where both expectations are taken with respect to Q. We need to introduce the *relative entropy* (or *Kullback-Leibler (KL) divergence*; e.g., [5]):

$$\mathrm{D}(Q\|P) = E_{h\sim Q}\left[\ln \frac{q(h)}{p(h)}\right] , \tag{4}$$

where q, p denote the probability densities of the distributions Q, P. If $\mathcal{H}$ is finite, we have[4]

$$\mathrm{D}(Q\|P) = \sum_{j=1}^{k} q_j \ln \frac{q_j}{p_j}, \tag{5}$$

where $Q = (q_1, \ldots, q_k),\ P = (p_1, \ldots, p_k)$. The relative entropy is an asymmetric distance measure between probability distributions, with $\mathrm{D}(Q\|P) = 0$ if and only if $Q = P$ almost everywhere.

Theorem 2 (PAC-Bayes [14]) *Let P be any prior distribution over $\mathcal{H}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D we have that for all distributions Q over the hypothesis space $\mathcal{H}$:*

$$\mathrm{Pr}_{D,Q}[h(x) \neq y] \leq \mathrm{Pr}_{S,Q}[h(x) \neq y] + \sqrt{\frac{D(Q\|P) + \ln \delta^{-1} + \ln m + 2}{2m-1}}$$

Here, $\mathrm{Pr}_{D,Q}[\cdot]$ is short for $E_{h\sim Q}[\mathrm{Pr}_D[\cdot]]$, and $\mathrm{Pr}_{S,Q}[\cdot]$ stands for $E_{h\sim Q}[\mathrm{Pr}_S[\cdot]]$. This theorem holds for finite and infinite hypothesis spaces. The PAC-Bayes theorem guarantees a tighter bound (except at low order) than earlier results such as the following Occam's razor theorem.

Theorem 3 (Occam's Razor [3]) *Let P be a distribution over a hypothesis space $\mathcal{H}$ and $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D, for all hypotheses $h \in \mathcal{H}$:*

$$\mathrm{Pr}_D[h(x) \neq y] \leq \mathrm{Pr}_S[h(x) \neq y] + \sqrt{\frac{\ln(1/p(h)) + \ln \delta^{-1}}{2m}} . \tag{6}$$

[3] Such classifiers are called *Gibbs classifiers* (e.g. [10]).

[4] Here and elsewhere, we agree on the definition $0 \log 0 = \lim_{t\to 0+} t \log t = 0$.

Note that, for finite $\mathcal{H}$ and up to low order terms, theorem 3 is a special case of theorem 2, where we choose delta distributions $Q = (0, \ldots, 0, 1, 0, \ldots, 0)$ in theorem 2. The essence of our improvement of the standard margin bound comes from the application of the PAC-Bayes bound instead of the Occam's razor bound within the standard proof of the margin bound.

3. An improved averaging bound

In this section, we state and prove our main result, a PAC-Bayes generalization error bound for averaging classifiers. Before we do this, we provide a discussion of a special case in order to put across an intuition of how our bound can improve upon theorem 1. In this discussion, we limit ourselves to a finite $\mathcal{H} = \{h_1, \ldots, h_k\}$, while the main result will be stated for arbitrary $\mathcal{H}$. Our bound relies on a "posterior" distribution $Q = (q_1, \ldots, q_k)$ over $\mathcal{H}$, from which the average $f(x)$ is defined as $f(x) = \sum_k q_k h_k(x)$. The "posterior" Q may depend on the training sample S in an arbitrary way.[5]

The *entropy* $H(Q)$ (e.g. [5]) of Q is defined as $H(Q) = -\sum_i q_i \ln q_i$. It measures the "uncertainty" in Q, in that delta distributions $Q = (0, \ldots, 0, 1, 0, \ldots, 0)$ have minimum entropy 0 and the uniform distribution has maximum entropy $\ln k$. We can state a special case of our main result as follows.

Theorem 4 (Special Case) *Let $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S from D, for all distributions Q over the hypothesis space $\mathcal{H}$ and for all $\theta \in (0,1]$:*

$$\mathrm{Pr}_D[yf(x) \leq 0] \leq \mathrm{Pr}_S[yf(x) \leq \theta] + O\left(\sqrt{\frac{\theta^{-2}(\ln|\mathcal{H}| - H(Q))\ln m + \ln m + \ln \delta^{-1}}{m}}\right). \tag{7}$$

This theorem is just a simplification of theorem 5 to finite hypothesis spaces with a uniform prior P.

How much can the improvement help us? About the best case we could hope for is a uniform average over half the hypothesis space[6]. In that case, the complexity term $(\ln|H| - H(Q))\ln m$ is quite small: $\ln m \ln 2$. In the worst case, when the average is over only a number of hypotheses k similar to the number of examples

[5] An example is, in Bayesian classification, the *posterior distribution* over $\mathcal{H}$. The corresponding averaging classifier $f(x)$ is called *Bayes classifier* or *Bayes-optimal classifier* (e.g. [10]).

[6] An average over a subset of the hypothesis space $\mathcal{H}$ includes only those hypotheses h_j with coefficients q_j significantly different from zero.

m, there is no significant improvement over the original margin bound.

It is easy to generalize the improved averaging bound to continuous spaces with arbitrary priors by carefully applying the PAC-Bayes bound.

Theorem 5 (Main theorem) *Let P be any continuous probability distribution over $\mathcal{H}$ and let $\delta \in (0,1)$. With probability at least $1-\delta$ over random samples S of D, for all $\theta > 0$ and for every distribution Q over $\mathcal{H}$:*

$$\begin{aligned}\Pr_D\left[yf(x) \le 0\right] &\le \Pr_S\left[yf(x) \le \theta\right] \\ &+ O\left(\sqrt{\frac{\theta^{-2} D(Q\|P)\ln m + \ln m + \ln \delta^{-1}}{m}}\right)\end{aligned} \tag{8}$$

where $f(x) = E_{h\sim Q}[h(x)]$.

The proof is given in 3.1.

Theorem 5 holds also for the case of bounded real-valued hypotheses, without any loss in the tightness of the bound. The theorem can also be tightened in several quantitatively important ways. Details can be found in [12].

There exists an alternative approach for deriving a bound similar to Theorem 5 which needs to be mentioned. Essentially, starting with the covering number based approach of [1] we can use the covering number results from theorem 3.6 of [18] to arrive at a similar bound. The principle advantage of our approach over this one is simplicity of argumentation combined with quantitatively tighter results.

The continuous form of the improved averaging bound applies to arbitrary averages over continuous hypothesis spaces, the *finite* averages defined in subsection 2.2 are special cases. Note that in this setting, the average needs to be an integral over an uncountably infinite set of hypotheses, otherwise the KL-divergence does not converge. In practice, this is not a significant problem because machine learning algorithms over large hypothesis spaces typically have some parameter stability. In other words, a small shift in the parameters of the learned model produces a small change in the prediction of the hypothesis. With hypothesis stability, we can convert any average over a finite set of hypotheses into an average over an infinite set of hypotheses without significantly altering the predictions of the average.

3.1 Proof of main theorem

The proof has the same structure as the original margin bound proof 1 with one step replaced by the application of the PAC-Bayes theorem 2.

Our averaging classifier is specified by

$$c(x) = \text{sign } E_{h\sim Q}[h(x)] \ .$$

Let N be any natural number; later, the choice of N will be optimized. For every distribution Q, we construct a random function $g = g_Q$ as follows. Draw N hypotheses i.i.d. from Q and define

$$g(x) = \frac{1}{N}\sum_{j=1}^{N} h_j(x). \tag{9}$$

The set of all possible g's is denoted

$$\mathcal{H}_N = \left\{ \frac{1}{N}\sum_{j=1}^{N} h_j(x) \;\middle|\; h_j \in \mathcal{H} \right\}, \tag{10}$$

and we denote the distribution of g (i.e., over the set $\mathcal{H}_N$) by Q^N. Note that for a fixed pair (x,y), the quantities $h_j(x)$ in the expression for $g(x)$ (see (9)) are i.i.d. Bernoulli variables (over $\{-1,1\}$) with mean

$$yE_{h_j\sim Q}[h_j(x)] = yf(x) \ . \tag{11}$$

Therefore, $yE_{g\sim Q^N}[g(x)] = yf(x)$. Since $g(x)$ is the average over N i.i.d. Bernoulli variables, Hoeffding's bound (see [6], p.122) applies. Thus, for every $x \in \mathcal{X}$, $y \in \{-1,+1\}$, the probabilities with respect to the sampling of $h_1,\ldots,h_N$ satisfy

$$\Pr_{g\sim Q^N}\left[y(g(x) - f(x)) > \epsilon\right] \le e^{-\frac{1}{2}N\epsilon^2} \tag{12}$$

For every $\theta > 0$ and for every (fixed) $g \in \mathcal{H}_N$, the following simple inequality holds:

$$\begin{aligned}&\Pr_D[yf(x) \le 0] \\ &= \Pr_D[yg(x) \le \tfrac{1}{2}\theta,\, yf(x) \le 0] \\ &+ \Pr_D[yg(x) > \tfrac{1}{2}\theta,\, yf(x) \le 0] \\ &\le \Pr_D[yg(x) \le \tfrac{1}{2}\theta] + \Pr_D[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \le 0].\end{aligned} \tag{13}$$

Note that the left-hand side does not depend on g. By taking the expectation over $g \sim Q^N$ (and exchanging the order of expectations in the second term on the right-hand side), we arrive at

$$\begin{aligned}\Pr_D[yf(x) \le 0] &\le E_{g\sim Q^N}\left[\Pr_D[yg(x) \le \tfrac{1}{2}\theta]\right] \\ &+ E_D\left[P_{g\sim Q^N}[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \le 0]\right].\end{aligned} \tag{14}$$

As discussed above, we are now ready to apply Hoeffding's inequality (12) with $\epsilon = \theta/2$. For any fixed (x,y) we have

$$Pr_{g\sim Q^N}[yg(x) > \tfrac{1}{2}\theta \,|\, yf(x) \le 0] \le e^{-\frac{1}{8}N\theta^2}, \tag{15}$$

so

$$\Pr_D[yf(x) \le 0] \le E_{g\sim Q^N}\left[\Pr_D[yg(x) \le \tfrac{1}{2}\theta]\right] + e^{-\frac{1}{8}N\theta^2}. \quad (16)$$

We would like to apply the PAC-Bayes theorem 2 to the right-hand side. For simplicity we stated theorem 2 for the common *zero-one loss* $I(h(x) \neq y)$, but it holds more generally for arbitrary binary loss functions (see [14]). Here we use the loss function $I(yg(x) \le \theta/2)$. Recall that theorem 2 applies for any fixed hypothesis space and "prior" distribution. The hypothesis space here will be $\mathcal{H}_N$. We use as the "prior" the distribution P^N over $\mathcal{H}_N$, which is constructed from the prior P over $\mathcal{H}$ exactly as Q^N is constructed from Q (see (9)). It is easy to see that $\mathrm{D}(Q^N\|P^N) = N\mathrm{D}(Q\|P)$.[7]

It follows from Theorem 2 that with probability at least $1-\delta$ over random choices of S, for every Q,

$$\begin{aligned} & E_{g\sim Q^N}\left[\Pr_D[yg(x) \le \tfrac{1}{2}\theta]\right] \\ & \le E_{g\sim Q^N}\left[\Pr_S[yg(x) \le \tfrac{1}{2}\theta]\right] \\ & + \sqrt{\frac{N\mathrm{D}(Q\|P) + \ln m + \ln(1/\delta) + 2}{2m-1}}. \end{aligned} \quad (17)$$

By the same argument as in (13), for every $g \in \mathcal{H}_N$:

$$\begin{aligned} & \Pr_S[yg(x) \le \tfrac{1}{2}\theta] \\ & \le \Pr_S[yg(x) \le \tfrac{1}{2}\theta,\ yf(x) > \theta] + \Pr_S[yf(x) \le \theta] \\ & \le \Pr_S[yg(x) \le \tfrac{1}{2}\theta \,|\, yf(x) > \theta] + \Pr_S[yf(x) \le \theta]. \end{aligned} \quad (18)$$

Again, we take expectations over $g \sim Q^N$ on both sides, interchange the order of the expectations and apply Hoeffding's inequality (12) with $\epsilon = \theta/2$:

$$E_S\left[P_{g\sim Q^N}\left[yg(x) \le \tfrac{1}{2}\theta \,|\, yf(x) > \theta\right]\right] \le e^{-\frac{1}{8}N\theta^2}, \quad (19)$$

to arrive at

$$E_S\left[P_{g\sim Q^N}[yg(x) \le \tfrac{1}{2}\theta]\right] \le e^{-\frac{1}{8}N\theta^2} + \Pr_S\left[yf(x) \le \theta\right]. \quad (20)$$

Combining (16), (17) and (20), we conclude that with probability at least $1-\delta$, for every Q

$$\begin{aligned} \Pr_D\left[yf(x) \le 0\right] - \Pr_S\left[yf(x) \le \theta\right] \le 2e^{-\frac{1}{8}N\theta^2} \\ + \sqrt{\frac{N\mathrm{D}(Q\|P) + \ln m + \ln(1/\delta) + 2}{2m-1}}. \end{aligned} \quad (21)$$

[7]Note that this reveals a tradeoff between N and $\mathrm{D}(Q\|P)$. Namely, for large N, $g \sim Q^N$ will be a close approximation to the averaging classifier f, which keeps (15) small, but if $\mathrm{D}(Q\|P)$ is not very small, Q^N will be rather far from P^N in terms of relative entropy, as a consequence of the strict factorized forms of the two distributions (they are constructed using i.i.d. samples of size N).

This bound holds for any fixed N and θ, which is not yet what we need here, since we want to allow these to depend on the data S. We apply a standard technique to resolve this problem. In essence, the bound we proved so far is a statement about certain events, parameterized by N and θ, namely the probability of each event is smaller than δ. However, we need to prove that the probability of the *union* of all these events is smaller than δ. To this end, we first observe that this union is contained in the union over only a *countable* number of events. Note that if $g \in \mathcal{H}_N$ (see (9)), then $g(x) \in \{(2k-N)/N | k = 0, 1, \ldots, N\}$. Thus, even with all the possible (positive) values of θ, there are no more than $N+1$ events of the form $\{yg(x) \le \theta/2\}$. Denote by $k(\theta, N)$ the largest integer $k \le N$ such that $k/N \le \theta/2$. We observe that for every $\theta > 0$, every $g \in \mathcal{H}_N$ and every distribution over (x, y):

$$Pr\left[yg(x) \le \theta/2\right] = Pr\left[yg(x) \le k(\theta, N)/N\right]. \quad (22)$$

This means that the middle step in the proof above, i.e. the application of theorem 2, depends on (N, θ) only through (N, k). Since the other steps, i.e. the applications of Hoeffding's inequality, are true with probability one, we see that we can restrict ourselves to the union of countably many events, indexed by (N, k). Now, we "allocate" parts of the confidence quantity δ to each of these events, namely (N, k) receives $\delta_{N,k} = \delta/(N(N+1)^2)$, $N = 1, 2, \ldots;\ k = 0, \ldots, N$. It follows easily that the union of all these events has probability at most $\sum_{N,k} \delta_{N,k} = \delta$. Therefore we have proved that with probability at least $1-\delta$ over random choices of S, for *all* N and *all* $\theta > 0$,

$$\begin{aligned} & \Pr_D\left[yf(x) \le 0\right] - \Pr_S\left[yf(x) \le \theta\right] \\ & \le 2e^{-\frac{1}{8}N\theta^2} + \sqrt{\frac{N\mathrm{D}(Q\|P) + \ln m + \ln(1/\delta_{N,k}) + 2}{2m-1}} \\ & \le 2e^{-\frac{1}{8}N\theta^2} \\ & + \sqrt{\frac{N\mathrm{D}(Q\|P) + \ln m + \ln(1/\delta) + 3\ln(N+1) + 2}{2m-1}} \end{aligned} \quad (23)$$

where $k = k(\theta, N)$. The asymptotic bound stated in the theorem can be derived by choosing N (with respect to θ and Q) so as to approximately minimize the bound we have derived above. We can choose

$$N = \left\lceil 4\theta^{-2} \ln \frac{m}{\mathrm{D}(Q\|P) + 1} \right\rceil.$$

4. Implications

We wish to apply the preceding theory to two general learning methods: Maximum Entropy discrimination[11] and Bayes as well as Bayes Point Classifiers [15] [9]. We choose these two learning methods because the average in these cases is over many hypotheses, so that the low order terms in the bound are not very significant. We begin with a simple toy example that illustrates the bound application.

4.1 Example

A quick example will illustrate the advantage of the improved bound. Suppose the input space is $\mathcal{X} = \{-1,1\}^n$, and let $\mathcal{H} = \{h_1,\ldots,h_n\}$, where for every $x = (x_1,\ldots,x_n) \in \mathcal{X}$, $h_i(x) = x_i$, $i = 1,\ldots,n$. Fix a parameter $0 < \theta < 1$.

The setting falls within the naive Bayes probability model. The probability distribution D can be described as follows: First, the value of y is 1 with probability 0.5, and -1 with probability 0.5. Given y, the entries of an instance $(x_1,\ldots,x_n,y) \in \mathcal{X} \times \{-1,1\}$ are (conditionally) independent. For every i, x_i equals y with probability $\frac{1}{2} + \frac{1}{2}\theta$, and $-y$ with probability $\frac{1}{2} - \frac{1}{2}\theta$.

It follows, that for every i, $yh_i(x) = 1$ (i.e., h_i predicts correctly) with probability $\frac{1}{2} + \frac{1}{2}\theta$, so the expected value of $yh_i(x)$ is $\frac{1}{2} + \frac{1}{2}\theta - (\frac{1}{2} - \frac{1}{2}\theta) = \theta$. Thus, the expected value of $t(x,y) = \sum_{i=1}^n q_i(yh_i(x))$ is also θ. For a large number of independent hypotheses, with uniform weights q_i, the value of $t(x,y)$ is probably approximately θ.

What will the old margin bound suggest using? The old margin bound depends purely on the proportion of examples at some margin so it suggests averaging over the few hypotheses which happen to do better than expected on this particular sample set.

What does the improved averaging bound suggest? The improved averaging bound will include many more hypotheses because it becomes tighter with a more uniform average over the hypothesis space.

We implemented two quick learning algorithms to explore the implications of this bound to this problem.

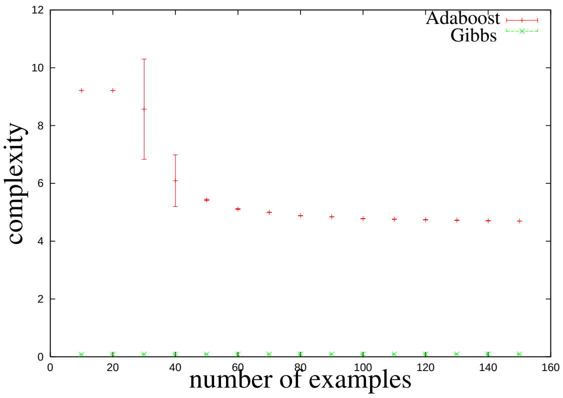
The first algorithm, which motivated the development of the original margin bound, is Adaboost [7]. Our Adaboost implementation uses a weak learning algorithm which simply selects the hypotheses with smallest error under the distribution over examples and we set the number of rounds to 100.

The second algorithm is a Gibbs averaging algorithm. The Gibbs averaging algorithm picks a weight for each hypothesis proportional to

$$e^{es(h)/T},$$

where T is a "temperature" parameter. Motivated by the variance of a binomial distribution, we set $T = 1/\sqrt{n}$. After finding the weights of all hypotheses we create an averaging classifier by taking the sign of the expectation with respect to the Gibbs distribution.

In real world examples, the number of hypotheses is typically much larger than the number of examples so we use examples with 10000 hypotheses/features and 10 to 150 examples. For all experiments we set the true margin in data generation to 0.5.



This is a plot of the complexity $\mathrm{D}(Q\|P)$ of the averaging hypotheses returned by Adaboost and Gibbs-averaging versus the number of examples. All error bars are at one standard deviation.

Adaboost just picks one hypothesis (the one which happens to get every example correct) when the number of examples is small and eventually limits to a near uniform distribution over as many hypotheses as it has iterations. The Gibbs-averaging hypothesis instead always controls complexity well. The error bars are very small everywhere except for 30 and 40 examples where Adaboost suddenly starts using more then one hypothesis.

This example is arranged so the "right" answer is to use every hypothesis with the same weight. In general, the goals of complexity control and error minimization are often opposed and the averaging bound suggests how to trade off between these goals.

4.2 Maximum Entropy Discrimination

Maximum Entropy discrimination (MED) is tailor-made to take advantage of our new bound. This is especially interesting because it was proposed before the averaging bound was developed. Consequently, the application of the the improved averaging bound to the MED framework provides an additional motivation for its use.

Maximum Entropy discrimination (MED) is founded on Minimum Relative Entropy discrimination (MRED) which is equivalent to MED when the prior happens to be uniform. In the MRED paradigm, one starts with some prior distribution P over the hypothesis space and the goal is to find a distribution Q which minimizes the KL-divergence $\mathrm{D}(Q||P)$ to the prior subject to classification constraints (further explained below). The latter are stated in terms of the expectation over Q of the so-called discriminant function (see below).

For each hypothesis h, the discriminant function $L(x) = L(x|h)$ assigns real numbers to examples $x \in \mathcal{X}$. The value of $L(x|h)$ can be interpreted as a "confidence-rated" classification. Thus, if $L(x|h)$ is large, the hypothesis h places great confidence on the classification of x as positive.

For every h, the discriminant function $L(\cdot|h)$ is determined by a parameter triple $\Theta^h = \{\theta_+^h, \theta_-^h, b^h\}$. It is derived from some parameterized family $P(\cdot|\theta)$ of probability distributions over the set $\mathcal{X}$. The discriminant function is:

$$L(x|h) = \ln \frac{P(x|\theta_+^h)}{P(x|\theta_-^h)} + b^h \ .$$

Intuitively, $P(\cdot|\theta_+^h)$ (resp. $P(\cdot|\theta_-^h)$) is the posterior distribution over $\mathcal{X}$, given a "positive" (resp. "negative") classification. Thus, the discriminant is a "biased" (by b^h) log-likelihood ratio with respect to the distributions $P(\cdot|\theta_+^h)$ and $P(\cdot|\theta_-^h)$. Every distribution Q over $\mathcal{H}$ induces an average over the discriminant functions of the individual hypotheses. The constraints imposed on Q guarantee a desired margin θ:

$$\forall x, y \in S: \ y \int_{\mathcal{H}} L(x|h)\, dQ(h) \geq \theta \ .$$

Classification is then done in the MED framework by calculating the expected value of the discriminant function under the distribution Q:

$$c(x) = \mathrm{sign} \int_{\mathcal{H}} L(x|h)\, dQ(h) \ .$$

For details on how to find Q subject to these constraints see [11].

How does our theoretical result apply to MRED? The latter averages over real-valued discriminant functions $L(x|h)$ instead of binary hypotheses h, as in our setting. As already mentioned above, our main theorem 5 holds also for spaces $\mathcal{H}$ of *bounded real-valued* hypotheses, without further loss. If the $L(x|h)$ in an MRED application are bounded, our result therefore applies, using the hypothesis space $\{L(x|h)|h \in \mathcal{H}\}$ instead of $\mathcal{H}$ directly. However, in most MRED applications, the discriminant functions are not bounded, and an extension of our result to this case is subject to future work.

The algorithm directly motivated by the averaging bound would be "Minimum Relative Entropy Classification" (MREC) which is identical to the MRED framework except that instead of averaging over real-valued discrminants the average is done over binary valued classifiers. It is unclear whether the MREC criteria is actually better then the MRED framework in practice for either accuracy or ease of solution.

4.3 Bayes and Bayes Point Classifiers

In the Bayesian approach to classification, given an i.i.d. training sample S, a posterior distribution Q over the hypotheses is derived according to Bayes law:

$$dQ(h) = d\,\mathrm{Pr}(h|S) = \frac{\mathrm{Pr}(S|h)\, d\,\mathrm{Pr}(h)}{\int \mathrm{Pr}(S|h)\, d\,\mathrm{Pr}(h)} \ .$$

Here, $\mathrm{Pr}(S|h)$ is the likelihood of the sample S, given the hypothesis h, and $\mathrm{Pr}(h)$ is a prior distribution over $\mathcal{H}$. The classification for a new example x is done by calculating the expectation over Q, as in the MRED classifier above:

$$c(x) = \mathrm{sign} \int_{\mathcal{H}} h(x)\, dQ(h) \ .$$

The Bayes (optimal) classifier is an average over many classifiers, and so our improved averaging bound applies with a prior P, given by $P(h) = \mathrm{Pr}(h)$, and a "posterior" Q, which in this case is the Bayesian posterior distribution.

One significant drawback of this technique is that it is very often intractable. Bayes Point classifiers attempt to address this intractability by finding a *single* hypothesis $h_{BP}(x) \in \mathcal{H}$, which is close to the Bayes (optimal) classifier $c(x)$. Thus, if $h_{BP}(x)$ is a good approximation to $c(x)$, then the improved averaging bound will approximately apply to the Bayes Point classifier as well.

5. Conclusion and Future Work

We have presented a simple qualitative improvement to the margin bound, which motivates the techniques of several learning algorithms and validates the intuition that "averaging is good". There are many directions for interesting future work including the following:

1. The improved averaging bound has some messy low order constants, which are probably removable with an improved argument.
2. Can we give a stronger theoretical motivation of the Maximum Entropy discrimination framework with unbounded discriminant functions?
3. Empirical application of the bound. When applying this bound in the boosting framework, can we get quantitatively interesting results on real world problems?

Acknowledgments

We thank Ralf Herbrich, Chris Williams, Robert Schapire, and David McAllester for helpful discussions. MS is supported through a research studentship from *Microsoft Research Ltd.*

References

[1] Peter 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, 1998

[2] L. Breiman, "Bagging Predictors" Machine Learning, Vol. 24, No. 2, pp. 123-140.

[3] A. Blumer, A. Ehrenfucht, D. Haussler, and M. K. Warmuth, "Occam's Razor" Information Processing Letters, 24:377-380, April 1987

[4] P. Chesseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman. Autoclass: A Bayesian classification system. In Proc. Fifth Intl. Conf. Machine Learning, pages 54–64, 1988.

[5] Thomas Cover and Joy Thomas, "Elements of Information Theory" Wiley, New York 1991

[6] L. Devroye, L. Györfi and G. Lugosi "Applications of Mathematics: Stochastic Modelling and Applied Probability" Springer 1996

[7] Yoav Freund and Robert E. Schapire, "A Decision Theoretic Generalization of On-line Learning and an Application to Boosting" Eurocolt 1995

[8] Adam J. Grove and Dale Schuurmans "Boosting in the limit: Maximizing the margin of learned ensembles" In Proceedings of the Fifteenth National Conference on Artificial Intelligence 1998

[9] Ralf Herbrich, Thore Grapel, and Colin Campbell, "Bayes Point Machines: Estimating the Bayes Point in Kernel Space", IJCAI 1999 pages 23-29.

[10] David Haussler, Michael Kearns, and Robert Schapire, "Bounds on the Sample Complexity of Bayesian Learning Using Information Theory and the VC Dimension", Machine Learning 1994 14:83–113

[11] T. Jaakkola, M. Meila, T. Jebara, "Maximum Entropy Discrimination" NIPS 1999.

[12] J. Langford, M. Seeger, "Bounds for Averaging Classifiers", Technical report, Carngeie Mellon, 2001, CMU-CS-01-102

[13] Nick Littlestone. Redundant noisy attributes, attribute errors, and linear threshold learning using winnow. In COLT-91, pages 147–156, 1991.

[14] David McAllester, "PAC-Bayesian Model Averaging" COLT 1999

[15] Thomas Minka, "Expectation Propagation for approximate Bayesian inference", thesis.

[16] Robert E. Schapire, Yoav Freund, Peter Bartlett, and Wee Sun Lee, "Boosting the Margin: A new explanation for the effectiveness of voting methods" The Annals of Statistics, 26(5):1651-1686, 1998.

[17] John Shawe-Taylor, Peter Bartlett, Robert Williamson and Martin Anthony, "A framework for Structural Risk Minimization", COLT-96 pages 68-76

[18] Tong Zhang, "Analysis of Certain Regularized Linear Function Classes with Special Emphasis on Classification", IBM Research Report RC-21572