The golden chain

Peter J. Bickel
University of California
at Berkeley

Ya'acov Ritov
The Hebrew University
of Jerusalem

July 8, 2003

And through the palpable obscure find out / His uncouth way.
J. Milton, *Paradise Lost.*

Jiang, Lugosi and Vayatis, and Zhang in part explicitly and in part implicitly, have done a great deal in explaining the nature of boosting from a statistical point of view.

The problem all consider is that of finding classifiers that approximate the Bayes classifier using only a training sample (X_i, Y_i), $i = 1, \dots, n$, $(X_i, Y_i) \sim (X, Y)$, with $Y = \pm 1$ (for simplicity). The Bayes classifier is described as sgn $(F_p(X))$, where $F_p(X) = q \circ \log\bigl(p[Y = 1 \mid X]/P[Y = -1 \mid X]\bigr)$, for any strictly increasing function q with $q(0) = 0$.

The methods of approximation discussed by these and previous authors cited in their papers have the common setting that the approximating values are sgn $(\hat{F}(X))$, where $\hat{F} \in \tilde{\mathcal{F}} \equiv \bigcup_{k=1}^{\infty} \mathcal{F}_k$, $\mathcal{F}_k = \bigl\{\sum_{j=1}^{k} \lambda_j h_j : h_1, \dots, h_k \in \mathcal{H}, \lambda_1, \dots, \lambda_k \in R\bigr\}$ and $\mathcal{H}$ is a set of base classifiers, $h : \mathcal{X} \to \{-1, 1\}$.

All methods are based on the following two observations:

(i) Given W convex, $W = R \to R^+$, then, at least formally, if $\tilde{\mathcal{F}}$ is rich enough and P denotes expectation, then $F_p = \arg\min PW\bigl(YF(X)\bigr)$ as above. The validity of this identity is studied extensively by Zhang who relates it to minimizing the Bregman divergence between F and F_p. The function $W(t) = e^t$ correspond to classical ADAboost, while $W(t) = -2t + t^2$ is "L_2 boosting", See Bühlmann and Yu (2001), Friedman et al. (2000).

(ii) One "optimizes" $P_n W\bigl(YF(X)\bigr)$ over $\tilde{\mathcal{F}}$ where P_n is the empirical distribution of (X_i, Y_i), $i = 1, \dots, n$ in the same way to obtain $\hat{F}$. The classical prescription of Breiman (2000) is to optimize greedily starting at $F_0 \equiv 1$ using the Gauss-Southwell approach moving from $\mathcal{F}_m$ to $\mathcal{F}_{m+1}$ on the mth step.

Unfortunately, as is made fairly explicit in these papers, unless P is discrete, $\inf_{F \in \tilde{\mathcal{F}}} P_n W(YF(X))$ =0, and optimizing to the bitter end leads to overfitting.

Jiang shows for classical ADAboost that that, under some conditions, given convergence of the population algorithm, it is possible to stop the sample algorithm early and achieve consistency, i.e., convergence to the Bayes classifier. Lugosi and Vayatis and Zhang separately show that by regularizing, effectively changing what is being optimized, convergence to the Bayes classifier is possible quite generally and obtain rates for their procedures. Such approaches via sieves have already been considered by Baraud (2002) for "L_2 boosting".

We see four distinct questions:

(i) When are greedy algorithms consistent in the population case?

(ii) When does early stopping in the sample case lead to a consistent procedure?

(iii) How can early stopping be implemented by cross-validation?

(iv) How can one directly modify the greedy algorithm retaining its simple sequential structure and yet achieve optimal rate upon stopping suitably?

In our remark we address points (i) and (ii). Point (ii) is treated separately by Bickel and Ritov (2003), Yu and Zhang (2003) and Bühlmann (2002) and (iv) is in progress.

1 Weak consistency

Here is a very general framework.

Let $\Theta_1 \subset \Theta_2 \subset \dots$ be a sequence of sets contained in a separable metric space with metric ρ, $\Theta = \overline{\cup\Theta_m}$ where $\overline{}$ denotes closure. Let K be a target function, and $\vartheta_\infty = \arg\min_\Theta K(\vartheta)$. Let $\Pi_m : \Theta_{m+1} \to \Theta_m$. Finally, let K_n be a sample based approximation of K. We assume:

A1: For any $m, \vartheta_0 :\ M$, $\Theta_m \cap \{\vartheta : \rho(\vartheta, \vartheta_0) < M\}$ is compact. Let $K : \Theta \to R$ and assume that $\vartheta_\infty = \arg\min_{\vartheta \in \Theta} K(\vartheta)$ is unique.

A2: K is strictly convex and $K(\vartheta) \le K(\vartheta') \;\Rightarrow\; \rho(\vartheta, \vartheta_\infty) \le A\rho(\vartheta', \vartheta_\infty)$ for some $A < \infty$.

A3: If $\rho(\vartheta_m, \vartheta_0) \to \infty$ for some, and hence all ϑ_0, then $K(\vartheta_m) \to \infty$.

Let $\Pi_m : \Theta_m \to 2^{\Theta_{m+1}}$ be a sequence of point to set ρ-continuous mappings, where distance between sets is defined as $\rho(A, B)$ is the Hausdorff distance between the closures of A and B, and define the following algorithm generating a sequence $\bar{\vartheta}_m \in \Theta_m$, $m = 1, 2, \ldots$ given an initial point ϑ_0:

(i) $\bar{\vartheta}_{m+1} \in \Pi_m(\bar{\vartheta}_m)$.

(ii) $K(\bar{\vartheta}_{m+1}) = \inf_{\vartheta \in \Pi_m(\bar{\vartheta}_m)} K(\vartheta)$.

Suppose:

A4: if $\{\vartheta_m\}$ is defined as above with any initial ϑ_0, then $\rho(\vartheta_m, \vartheta_\infty) \to 0$.

In boosting, given P, $\Theta = \{F(X), F \in \tilde{\mathcal{F}}\}$, ρ is a metric of convergence in probability, $\Theta_m = \{\sum_{j=1}^m \lambda_j h_j, h_j \in \mathcal{H}\}$ and $\Pi_m(F) = \{F + \lambda h, \lambda \in R, h \in \mathcal{H}\}$. Moreover, $K(F) = \mathrm{E}\, W(YF(X))$.

Now suppose $K_n(\cdot)$ is a sequence of random functions on Θ such that,

A5: K_n is convex and $\sup\{|K_n(\vartheta) - K(\vartheta)| : \vartheta \in \Theta_m,\ \rho(\vartheta, \vartheta_m) < M\} \xrightarrow{\mathrm{p}} 0$ for all finite m, M, ϑ_0.

In boosting, $K_n(F) = n^{-1} \sum_{i=1}^n W(Y_i F(X_i))$ and A5 corresponds to requiring that $\{W(YF(X)) : F \in \Theta_m, \rho(F, F_0) \le M\}$ is uniformity class for LLN for P, for instance, a VC class. Bühlmann (2003), Zhang and Yu (2003) and Bickel and Ritov (2003) discuss such conditions in different degrees of generality.

The sequence $\{\bar{\vartheta}_m\}$ is the golden chain we try to follow using the obscure information in the sample. Define $\hat{\vartheta}_{m,n}$ by:

(i) $\hat{\vartheta}_{m+1,n} \in \Pi_m(\hat{\vartheta}_{m,n})$.

(ii) If $\vartheta' \in \Pi_m(\hat{\vartheta}_{m,n})$, then $K_n(\hat{\vartheta}_{m+1,n}) \le K_n(\vartheta')$ and, in case of equality, also $\rho^*(\vartheta_{m+1,n}, \vartheta_0) \le \rho^*(\vartheta', \vartheta_0)$ for some metric ρ^* such that $\rho(\vartheta_m, \vartheta_0) \to \infty \;\Rightarrow\; \rho^*(\vartheta_m, \vartheta_0) \to \infty$.

The purpose of introducing ρ^* is to avoid an unnecessarily large norm of the estimate. In boosting ρ^* can be any metric like the $L_2(\mu)$ metric where μ has fatter tails than P.

Theorem 1.1 *Under A1–A5 there exists a sequence* $\{m_n\}$ *such that* $\rho(\vartheta_{m_m,n},\vartheta_\infty)\xrightarrow{\mathrm{p}}0$.

Proof. Consider $\hat\vartheta_{1,n}$. By definition

$$K_n(\hat\vartheta_{1,n})\le\min\bigl\{K_n(\vartheta_0),K_n(\bar\vartheta_1)\bigr\}. \tag{1}$$

However, for large enough M, we get from A3 that $\inf_{\vartheta\in\Theta_1,\rho(\vartheta,\vartheta_0)=M}K(\vartheta)>K(\vartheta_0)$. By A5 we obtain that also

$$P\Bigl(\inf_{\substack{\vartheta\in\Theta_1\\ \rho(\vartheta,\vartheta_0)=M}}K_n(\vartheta)>K(\vartheta_0)\Bigr)\to1. \tag{2}$$

Convexity of K_n, (1) and (2) imply that $\rho(\hat\vartheta_{1,n})$ is bounded. But then strict convexity of K and uniform convergence imply that

$$\rho(\hat\vartheta_{1,n},\bar\vartheta_1)\xrightarrow{\mathrm{p}}0. \tag{3}$$

We continue now to $\hat\vartheta_{2,n}$. Since K is continuous, (3) implies that $\inf_{\vartheta\in\Pi_1(\hat\vartheta_1)}K(\vartheta)\xrightarrow{\mathrm{p}}K(\bar\vartheta_2)$. Applying the same argument as for $\hat\vartheta_{1,n}$, we get $\rho(\hat\vartheta_{2,n},\bar\vartheta_2)$ is bounded, and since K is continuous and strictly convex, we get again that $\rho(\hat\vartheta_{2,n},\hat\vartheta_2)\xrightarrow{\mathrm{p}}0$. By induction, we obtain that $\rho(\hat\vartheta_{m,n},\bar\vartheta_m)\xrightarrow{\mathrm{p}}0$ for every m.

Let $m_n=\sup\{m:P\bigl(\rho(\hat\vartheta_{m,n},\bar\vartheta_m)<m^{-1}\bigr)<m^{-1}\}$. Then $m_n\to\infty$ and $\rho(\hat\vartheta_{m_n,n},\bar\vartheta_{m_n})\xrightarrow{\mathrm{p}}0$. Apply A4 to conclude the proof.

□

Results based on this theorem cannot give an estimate of the speed of convergence of $\hat\vartheta_{m_n,n}$ to $\vartheta\ _\infty$, since the $\{m_n\}$ are not known. As we have mentioned, regularization can yield such rates but in all cases we are left with a sequence $\{\hat\vartheta_{1,n},\hat\vartheta_{2,n},\dots\}$ of procedures for which we need to select a stopping time τ on the basis of the data such that $\hat\vartheta_{\tau,n}$ behaves well. A natural comparison is to the oracle stopping time W such that $EK(\hat\vartheta_{W,n})=\min_m EK(\hat\vartheta_{m,n})$. In the next section we give a general result guaranteeing that $K(\hat\vartheta_{\tau,n})\approx EK(\hat\vartheta_{W,n})$ in the context of classification. We shall show how this result may be applied to the regularized variants of boosting elsewhere.

2 The beauty of the test-bed

The boosting algorithm can be stopped appropriately if there are available good data driven bounds on the sample error. However, it is more practical to use some type of cross-validation. Here is a general result.

Assume that the observations are i.i.d. from $Z = (Y, X_1, X_2, \dots) = (Y, \mathbf{X})$, where $Y \in \{-1, 1\}$. The task is to find a function $\vartheta(\mathbf{X})$, such that $P(Y\vartheta(\mathbf{X}) > 0)$ is maximized. The sample is divided into a main sample, $Z_1, \dots, Z_n$, and a test-bed $Z_1^T, \dots, Z_k^T$. The main sample is used to derive a sequence of classifiers $\hat\vartheta_1, \hat\vartheta_2, \dots$. The test data is used to pick $\hat\vartheta_\tau$ as the classifier to be used, where

$$\tau = \arg\min_{m<M} \sum_{j=1}^{k} \mathbf{1}(Y_j^T \hat\vartheta_m(\mathbf{X}_j^T) > 0).$$

An oracle constrained to use rules of the form sgn $(Y^T\hat\vartheta_m(\mathbf{X}^T))$ would use

$$W = \arg\min_{m<M} P\Big(Y^T\hat\vartheta_m(\mathbf{X}) > 0 \;\Big|\; \hat\vartheta_i\Big).$$

Let $\eta_m = P\Big(Y^T\hat\vartheta_m(\mathbf{X}) > 0 \;\Big|\; \hat\vartheta_m(\cdot)\Big)$, $m = 1, 2, \dots$. The following assumption will be used:

S: Similarity of the good classifiers With probability converging to 1, one of the following holds for every $m < K$, :

1. $(\log M)^{-1}\sqrt{k}(\eta_m - \eta_W) > b_n$, for some $b_n \to \infty$.
2. $P\Big(\hat\vartheta_m(\mathbf{X})\hat\vartheta_W(\mathbf{X}) < 0 \;\Big|\; \hat\vartheta_m(\cdot), \hat\vartheta_W(\cdot)\Big) < a_n$, for some $a_n \to 0$. Moreover, there is a monotone non-decreasing function $\Psi(\cdot)$, $\Psi(0) = 0$ such that

$$\begin{aligned} &E\Big(Y\big(\mathbf{1}(\hat\vartheta_W(\mathbf{X}) > 0) - \mathbf{1}(\hat\vartheta_m(\mathbf{X}) > 0)\big) \;\Big|\; \hat\vartheta_m(\cdot), \hat\vartheta_W(\cdot)\Big) \\ &\qquad \le \Psi\left(\frac{E\Big(Y\big(\mathbf{1}(\hat\vartheta_W(\mathbf{X}) > 0) - \mathbf{1}(\hat\vartheta_m(\mathbf{X}) > 0)\big) \;\Big|\; \hat\vartheta_m(\cdot), \hat\vartheta_W(\cdot)\Big)}{\sqrt{P(\hat\vartheta_m(\mathbf{X})\hat\vartheta_W(\mathbf{X}) > 0 \mid \hat\vartheta_m(\cdot), \hat\vartheta_W(\cdot))}}\right) \end{aligned}$$

We essentially require that all procedures with close to optimal performance are similar.

Theorem 2.1 *Let Assumption S hold. Then* $\eta_\tau = \eta_W + o_p(\Psi(\sqrt{\log M/k}))$

Proof. Let the two sets of indexes postulated in Assumption **S** be S_1 and S_2 respectively. Since the estimates $k^{-1}\sum_{j=1}^{k} \mathbf{1}(Y_j^T\vartheta_m(\mathbf{X}_j^T) > 0)$, $m =$

$1, \dots, M$ have a uniform error bound of $\log(M)/\sqrt{k}$, we have $\tau \notin S_1$. Hence, with probability converging to 1, the test-bed stopping time is minimizing

$$U_m = k^{-1} \sum_{j=1}^{k} \Big(\mathbf{1}(Y_j^T \vartheta_W(\mathbf{X}_j^T) > 0) - \mathbf{1}(Y_j^T \vartheta_m(\mathbf{X}_j^T) > 0) \Big) \tag{4}$$

over $m \in S_2$. But the sum in (4) is of $\{-1, 0, 1\}$ *i.i.d.* random variables, which are 0 with high probability. Let p_m and q_m be the conditional probabilities (conditioned on the main sample) that a given term in the sum is 1 or -1 respectively. Then

$$\begin{aligned} \mathrm{E}\, U_m &= p_m - q_m \\ \operatorname{Var} U_m &= (1 + o(1))(p_m + q_m)/k. \end{aligned}$$

Hence, with probability converging to 1,

$$\begin{aligned} \eta_W - \eta_\tau &= \max\Big\{ p_m - q_m : \; m \in S_2, \sqrt{\frac{k}{\log M}} \frac{p_m - q_m}{\sqrt{p_m + q_m}} < 1 \Big\} \\ &\le \max\big\{ \Psi(\frac{p_m - q_m}{\sqrt{p_m + q_m}}) : \; m \in S_2, \Psi(\frac{p_m - q_m}{\sqrt{p_m + q_m}}) < \Psi\big(\sqrt{\frac{\log M}{k}}\big) \big\} \\ &\le \Psi\Big(\sqrt{\frac{\log M}{k}}\Big) \end{aligned}$$

□

REFERENCES

Bickel, P. J. and Ritov, Y. (2003): Boosting and other iterative procedures, unpublished.

Bühlmann, P. (2002): Consistency for L_2 Boosting and Matching Pursuit with Trees and Tree-type Basis Functions. *Technical Report.*

Baraud, Y. (2002): Model selection for regression on a random design. *ESAIM Probab. Statistiqes*, **5**, 127–146.

Mallat S. and Zhang Z. (1993) Matching pursuit with time frequency dictionaries. *IEEE Transactions on Signal Processing*, **5**,3397-3415

Zhang T. and Yu B. (2003) Boosting with early stopping: Convergence and consistency. *Technical Report*