Bayes Consistency of Boosting: Population versus Sample

W. Drago Chen and C. Andy Tsao

Department of Applied Math, National Dong Hwa University

Conference on MDT, Statistical Inference and Applications

In Honor of Deng-Yuan Huang

Outline

- Introduction
- Convergence and Consistency
- FHT's statistical view
- A decision theoretical approach
- Results
- Conclusion

Intro: Supervised Learning

- Training data $(x_i, y_i)_1^N$, $x \in \mathcal{X}, y \in \mathcal{Y} = \{\pm 1\}$.;
  Testing Data $(x'_j, y'_j)_1^M \rightsquigarrow (X_i, Y_i) \sim_{iid} P_{X,Y}$.
- Find *Machine* (Classifier) $F \in \mathcal{F}$
  $F : \mathcal{X} \to \mathcal{Y}$
- Training Error

$$TE = \frac{1}{N} \sum_i 1_{[y_i \neq F(x_i)]}$$

- Generalization/Testing Error

$$\widehat{GE} = \frac{1}{M} \sum_j 1_{[y'_j \neq F(x'_j)]}; \qquad GE = E_{Y,X} 1_{[Y \neq F(X)]}$$

Intro: Boosting

Ensemble classifiers.

- *Weak (base) learner*
- Sequentially applying it to reweighted version of the training data
  - Higher weights on the previous misclassified cases
  - Boosting iteration: T
- Weighted majority vote

Schapire (1990), Freund and Schapire (1997), Friedman, Hastie and Tibshirani (2000).
Breiman (2004), Jiang (2004), Meir and Rätsch (2003)

Intro: Discrete AdaBoost

1. Start with weights $D_t(i) = 1/N, i = 1$ to N.
2. Repeat for $t = 1$ to T
   - Obtain $h_t(x)$ from weak learner h using weighted training data wrt D_t
   - Compute $\epsilon_t = E_{D_t} 1_{[yh_t(x)<0]}, \alpha_t = \log \frac{1-\epsilon_t}{\epsilon_t}$.
   - Update $i = 1$ to N,

$$D_{t+1}(i) = \frac{1}{Z_t} D_t(i) \exp\left(\alpha_t 1_{[y_i h_t(x_i)<0]}\right),$$

   where Z_t is the normalizer.

3. Output the classifier $F_T(x) = sgn\left[\sum_{t=1}^{T} \alpha_t h_t(x)\right]$

Freund and Schapire (1997)

Convergence and Consistency

- $\lim_{T\to\infty} E_{Y,X} L(F_T(X), Y) = E_{Y,X} L(F_B(X), Y)$
- $\lim_{T\to\infty} F_T(x) = F_B(x)$, for all $x \in \mathcal{X}$.

where $F_B(x) = sgn(\log(\frac{P(Y=1|x)}{P(Y=-1|x)}))$ and $L(F(X), Y) = 1_{[YF(X)<0]}$.

Intro: Theories

- Bayes consistent
  (Population Version, Breiman (2004)).
  Process Consistent
  (Sample Version, Jiang (2004))
- Regularization needed, say, early stopping, restriction on base learners,particularly for noise data.

On the other hand

- Relatively immune to overfitting in practical apps.
- Mease and Wyner (2007, JMLR). Evidence Contradictory to Statistical View.
  - Relatively immune to overfitting (Convergence)
  - No regularization needed for some noisy data sets

"Statistical View": FHT's Insights

Friedman, Hastie and Tibishirani (2000).

- The Discrete AdaBoost (population version) builds an additive logistic regression model via Newton-like updates for minimizing $E(e^{-YF(X)})$.
- Exponential Criterion
  $L(Y, F(X)) = e^{-YF(X)} \approx L_0(Y, F(X)) = 1_{[YF(X)<0]}$.
- Easier for statisticians then ML approach
- Motivate boosting-like algorithm

Closer Look

Goal: Predicting $Y \in \{\pm 1\}$ by the sign of estimated F. $F : \mathcal{X} \rightarrow \mathcal{R}$.

- $E_X J(F(X)) = E_{X,Y}[e^{-YF(X)}] \approx E_{X,Y} 1_{[YF(X)<0]}$
- Min $J(F(x))$. Update $F(x)$ by $F(x) + cf(x)$ with $f(x) = \pm 1, c \in \mathcal{R}$
- For fixed c and x, expand at $f(x) = 0$

$$F_{t+1}(x) = F_t(x) + \alpha_t \; sgn(E_{w_t}(Y|x)) \text{ where}$$

$$\alpha_t = \log(\frac{1-\epsilon_t}{\epsilon_t}), \qquad \epsilon_t = E_{w_t} 1_{[y \; sgn(E_{w_t}(Y|x))<0]}$$

$$w_t(x,y) = \exp(-yF_t(x)).$$

Motivating Questions

- Convergence: Whether this iterative update converge?
- Consistency: Does it converge to the optimal Bayes with respect to $L_0(Y, F(X)) = 1_{[YF(X)<0]}$?
- Mease and Wyner (2007). Evidence Contradictory to Statistical View. of Boosting

Questions Solved?

- “Statistic View”: AdaBoost as a conditional risk minimizer wrt some approximate losses
  - AdaBoost can overfit
  - Regularization needed
- process-consistent or consistent under conditions on base learners
- Mease and Wyner (2007): Simulation experiments
  - AdaBoost relatively immunes to overfitting
  - No regularization needed

A decision theoretical approach

Find $F(x)$ minimizing $E_{Y|x}L(Y,F(x))$

- $Y = g(\theta) = sgn(\theta)$ and $X \sim P_\theta(x)$
- $\theta \sim \pi(\theta)$, prior
- Statistical problem

Objective: Find a classifier F minimizing

$$J(F) = E_{\pi(\theta|x)}L(g(\theta),F(x)) \approx E_{\pi(\theta|x)}\tilde{L}(g(\theta),F(x))$$

Loss

$$L(g(\theta),F(x)) = e^{-g(\theta)F(x)} \approx L_0(g(\theta),F(x)) = 1_{[g(\theta)F(x)<0]}.$$

Normal-normal setting

Let $X \sim N(\theta, \sigma^2)$ and $\pi(\theta) \sim N(\mu, \tau^2)$, w/ known μ and τ^2
Posterior $\pi(\theta|x) \sim N(\mu_x, \rho^{-1})$, where

$$\mu_x = \frac{1}{\rho}\left(\frac{\mu}{\tau^2} + \frac{x}{\sigma^2}\right) = \frac{\sigma^2\mu + \tau^2 x}{\sigma^2 + \tau^2}$$

$$\rho = \frac{1}{\tau^2} + \frac{1}{\sigma^2} = \frac{\sigma^2 + \tau^2}{\sigma^2\tau^2}$$

And marginal density of X

$$m(x) = \frac{1}{\sqrt{2\pi\rho}\sigma\tau} \exp\left\{-\frac{(\mu - x)^2}{2(\sigma^2 + \tau^2)}\right\}.$$

Iterative Bayes F_{PIB}: Derivation

Follow the steps similar to FHT (2000)

$$J(F+f) = E_{\pi(\theta|x)}\left\{e^{-g(\theta)[F(x)+f(x)]}\right\}$$

$$\approx \tilde{J}(F+f) = E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}[1-g(\theta)f(x)+f^2(x)/2]\right\}.$$

The minimizer f can then be found by differentiation

$$f(x) = \frac{E_{\pi(\theta|x)}\left\{g(\theta)e^{-g(\theta)F(x)}\right\}}{E_{\pi(\theta|x)}\left\{e^{-g(\theta)F(x)}\right\}}$$

$$= \frac{e^{-F(x)}\Phi(\sqrt{\rho}\mu_x) - e^{F(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}{e^{-F(x)}\Phi(\sqrt{\rho}\mu_x) + e^{F(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}.$$

Iterative Bayes F_{PIB}: Iteration

$$
\begin{aligned}
F_{PIB,t+1}(x) &= F_{PIB,t}(x) + f_t(x) \\
&= F_{PIB,t}(x) + \frac{\Phi(\sqrt{\rho}\mu_x) - e^{2F_{PIB,t}(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}{\Phi(\sqrt{\rho}\mu_x) + e^{2F_{PIB,t}(x)}\left[1 - \Phi(\sqrt{\rho}\mu_x)\right]}.
\end{aligned}
$$

- Does $F_{PIB,t}$ converge?
- Does $F_{PIB,t}$ to the optimal Bayes procedure wrt L_0?

Iterative Bayes F_{PIB}: Convergence

Theorem 1. *For any initial $F_{PIB,1}(x)$, as t goes to infinity*

$$F_{PIB,t}(x) \rightarrow F_\pi(x) = \frac{1}{2} \ln \left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1 - \Phi(\sqrt{\rho}\mu_x)} \right).$$

Iterative Bayes F_{PIB}: Lemmas

Lemma 1 (Fixed Point Theorem)**.** *If φ is a contraction of $\Re \to \Re$, that is, there exists $\alpha \in (0,1)$ such that $|\varphi(x) - \varphi(y)| < \alpha|x - y|$ for all $x, y \in \Re$, then there exists one and only one $x \in \Re$ such that $\varphi(x) = x$.*

Lemma 2 (Cauchy-Schwartz Inequality)**.** *For any real $a_i, b_i, i = 1, 2, \cdots, n$,*

$$\left(\sum_{i=1}^{n} a_i^2\right)\left(\sum_{i=1}^{n} b_i^2\right) \geq \left(\sum_{i=1}^{n} a_i b_i\right)^2 .$$

Lemma 3. *For all $x \neq 0$ $\frac{2(e^x-1)}{x(e^x+1)} < 1.$*

F_{FHT}: Derivation

$$F_{FHT,t}(x) \leftarrow F_{FHT,t}(x) + \frac{1}{2}\ln\left(\frac{1-\text{err}}{\text{err}}\right)s(x) \qquad (1)$$

where $s(x) = sgn(f(x))$ and

$$\text{err} = \frac{E_{\pi(\theta|x)}\{1_{[s(x)g(\theta)<0]}e^{-g(\theta)F_{FHT}(x)}\}}{E_{\pi(\theta|x)}\{e^{-g(\theta)F_{FHT}(x)}\}}$$

$$f(x) = \frac{e^{-F_{FHT}(x)}\Phi(\sqrt{\rho}\mu_x) - e^{F_{FHT}(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}{e^{-F_{FHT}(x)}\Phi(\sqrt{\rho}\mu_x) + e^{F_{FHT}(x)}[1-\Phi(\sqrt{\rho}\mu_x)]}.$$

FHT's AdaBoost: Convergence

By calculation, the iteration becomes

$$\begin{aligned} F_{FHT}(x) &\leftarrow F_{FHT}(x) + \frac{s^2(x)}{2}\left[\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}\right) - 2F_{FHT}(x)\right] \\ &= \frac{1}{2}\ln\left(\frac{\Phi(\sqrt{\rho}\mu_x)}{1-\Phi(\sqrt{\rho}\mu_x)}\right). \end{aligned}$$

Remark 1. *One-step convergence*

Bayes Risk $E_{X,\theta}\{1_{[g(\theta)F(X)<0]}\}$

- Difficulty of the problem
- Overfitting

$E_{\pi(\theta|x)}g(\theta) = 2\Phi(\sqrt{\rho}\mu_x) - 1$ and
$sgn(F_\pi(x)) = sgn(E_{\pi(\theta|x)}g(\theta))$.
Let $r = \tau/\sigma > 0$ and assume $\mu = 0$

$$E_{X,\theta}\left\{1_{[g(\theta)F(X)<0]}\right\} = \int_{-\infty}^{0} \Phi(t)\eta(t)dt + \int_{0}^{\infty} [1 - \Phi(t)]\eta(t)dt$$
$$= 2\int_{u=-\infty}^{u=0} \Phi(ru)d\Phi(u)$$

where $\eta(t) \sim N\left(\frac{\sqrt{\sigma^2+\tau^2}}{\sigma\tau}\mu, \frac{\tau^2}{\sigma^2}\right)$.

Bayes Risk: $h(r)$

Define

$$h(r) = 2\int_{u=-\infty}^{u=0} \Phi(ru)d\Phi(u).$$

Since $h(1) = \int_0^{1/2} \Phi d\Phi + \int_{1/2}^{1}(1-\Phi)d\Phi = \frac{1}{4}$ and $h'(r) = -[\pi(1+r^2)]^{-1}$.
Thus

$$E_{X,\theta}\left\{\mathbf{1}_{[g(\theta)F(X)<0]}\right\} = h(r) = \frac{1}{2} - \frac{1}{\pi}\tan^{-1} r. \qquad (2)$$

Summary

- Consistency of F_{PIB} and F_{FHT}
- AdaBoost F_{FHT} yields a highly effective one-step convergence under our distributional assumption
- Bayes Risk

Concluding Remark

- For the classification problems we formulated, our population version results suggest AdaBoost is extremely effective and no regularization needed.
- Contrast with current “statistical view” of boosting: $+$distribution assumption; $-$base learner/target learner assumptions.
- Distribution modelling can provide alternative “statistical view” to the boosting.

Thanks for your attention!