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 Cons Cod n

Intro: Supervised Learning

Training data ($_i$ $_i$) , $\mathcal{X}$ $\mathcal{Y}$ = $\{\pm 1\}$;

Testing Data ($_i$ $_i$) $\rightsquigarrow$ $(X_i, Y_i) \sim_{i.i.} X\,Y$

Find Machine (Classifier) $\in$

$: \mathcal{X} \to \mathcal{Y}$

Training Error

$$= \frac{1}{}$$

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:

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 ( ) = 1 = 1 to
2. Repeat for = 1 to

   Obtain ( ) from weak learner using weighted training data wrt

   Compute = $_{D_{\mathbf{t}}} 1_{[}$

Convergence and Consistency

lim $\rightarrow$ □ $Y X$ (XY $Y X$ ($B(XY$

lim $\rightarrow$ □ $B($ X :

where F$_B$ n $(Y=$

$(Y=$)

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 $(\, ^{-Y \quad (X)})$

Exponential Criterion

$\ell(Y \quad (X)) = \, ^{-Y \quad (X)} \approx \ell \; (Y \quad (X)) = 1_{[Y \quad (X) \quad]}$

Easier for statisticians then ML approach

Motivate boosting-like algorithm

Closer Look

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

$$_X \ (\ (X)) = \quad _{X\,Y}[\ ^{-Y} \quad ^{(X)}] \approx \quad _{X\,Y} 1_{[Y \quad (X) \quad]}$$

Min (( )). Update ( ) by ( ) + ( ) with ( ) = $\pm$1 $\mathcal{R}$

For fixed and , expand at ( ) = 0

$$_{t+} (\ ) = \quad _t(\ ) + \quad _t \ ($$

Motivating Questions

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

Regularization needed

Questions Solved?

"Statistic View": AdaBoost as a conditional risk minimizer wrt some approximate losses

- AdaBoost can overfit
-

Normal-normal setting

Let $X \sim$ $(\theta$) and $(\theta) \sim$ ( ), w/ known and
Posterior $(\theta \mid$) $\sim$ ($\bar{}$), where

$$= \frac{1}{} \left(\frac{}{} + \frac{}{} \right) = \frac{+}{+}$$

$$= \frac{1}{} + \frac{1}{} = \frac{+}{}$$

And marginal density of X

$$(\ ) = \frac{1}{\sqrt{2}} \exp\left\{ -\frac{(\ -\ )}{2(\ +\ )} \right\}$$

Iterative Bayes $_B$: Derivation

Follow the steps similar to FFFFim(2.197(t)0.1470.197(t526687

Iterative Bayes $_B$: Iteration

$$B_{t+}(\ ) = B_t(\ ) + {}_t(\ )$$

$$= B_t(\ ) + \frac{\Phi(\sqrt{\ }\ ) - \text{PIB,t}(\ )\left[1 - \Phi(\sqrt{\ }\ )\right]}{\Phi(\sqrt{\ }\ ) + \text{PIB,t}(\ )\left[1 - \Phi(\sqrt{\ }\ )\right]}$$

Does B_t converge?

Does B_t to the optimal Bayes procedure wrt ?

Iterative Bayes $_B$: Convergence

Theorem 1. For any initial $\hat{\pi}_B(\ )$, as goes to infinity

$$\hat{\pi}_{B,t}(\ ) \rightarrow \pi(\ ) = \frac{1}{2}\ln\left(\frac{\Phi(\sqrt{\ }\quad)}{1-\Phi(\sqrt{\ }\quad)}\right)$$

Iterative Bayes $_B$: Lemmas

Lemma 1 (Fixed Point Theorem)**.** If

: Derivation

$$\iota(\quad) \leftarrow \qquad \iota(\quad) + \frac{1}{2} \ln \left(\frac{1 - \text{err}}{\text{err}} \right) \quad (\quad) \qquad (1)$$

FHT's AdaBoost: Convergence

By calculation, the iteration becomes

$$
\begin{aligned}
(\quad) \leftarrow & \quad (\quad) + \frac{(\quad)}{2}\left[\ln\left(\frac{\Phi(\sqrt{\quad}\quad)}{1-\Phi(\sqrt{\quad}\quad)}\right) - 2 \quad (\quad)\right] \\
= & \quad \frac{1}{2}\ln\left(\frac{\Phi(\sqrt{\quad}\quad)}{1-\Phi(\sqrt{\quad}\quad)}\right)
\end{aligned}
$$

Remark 1. One-step convergence

Bayes Risk $X\theta\{1[(\theta)\ (X)\ 0]\}$

Difficulty of the problem

Overfitting

$\pi(\theta|\ )(\theta) = 2\Phi(\sqrt{\ }\ ) - 1$ and

$n(\ \pi(\ )) = \ n(\ \pi(\theta|\ )(\theta))$

Let $\ =\ $ 0 and assume $\ = 0$

$$X\theta\{1[\pi(\theta)\ (X)\ ]\} = \int_{-\ }\Phi(\ )\ (\ )\ \ + \int^{\ }[1 - \Phi(\ )]\ (\ )$$

$$= 2\int_{=-\ }^{=}\Phi(\ \ )\ \Phi(\ )$$

where $(\ ) \sim\ \ \left(\frac{\sqrt{\ ^2+\tau^2}}{\tau}\ \ \frac{\tau^2}{2}\right)$

Summary

Concluding Remark