AdaBoost is Universally Consistent

Peter Bartlett
Computer Science Division and Department of Statistics
UC Berkeley.

Joint work with Mikhail Traskin.

AdaBoost

Sample, $S_n = ((x_1, y_1), \dots, (x_n, y_n)) \in (X \times \{\pm 1\})^n$

Number of iterations, T

function AdaBoost(S_n, T)

$f_0 := 0$

for t from $1, \dots, T$

$$(\alpha_t, h_t) := \arg \min_{\alpha \in \mathbb{R}, h \in F} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i \left(f_{t-1}(x_i) + \alpha h(x_i)\right)\right)$$

$$f_t := f_{t-1} + \alpha_t h_t$$

return f_T

Universal Consistency

- Assume: i.i.d. data, $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from from $\mathcal{X} \times \mathcal{Y}$ (with $\mathcal{Y} = \{\pm 1\}$).
- Consider a method $f_n = A((X_1, Y_1), \ldots, (X_n, Y_n))$, e.g., $f_n = \texttt{AdaBoost}((X_1, Y_1), \ldots, (X_n, Y_n), t_n)$.

Definition: We say that the method is universally consistent if, for all distributions P,

$$L(f_n) \stackrel{a.s}{\to} L^*,$$

where L is the risk and L^* is the Bayes risk:

$$L(f) = \Pr(Y \neq \text{sign}(f(X))), \qquad L^* = \inf_f L(f).$$

AdaBoost is Universally Consistent

- Previous results
- The key theorem:
  Universal consistency for sublinearly increasing stopping times.
- Idea of proof
- Open questions

Previous results: Regularized versions

AdaBoost greedily minimizes

$$\mathbf{E}_n \exp(-Yf(X)) = \frac{1}{n}\sum_{i=1}^{n} \exp(-Y_i f(X_i))$$

over $f \in \mathrm{span}(F)$.

(Notice that, for many interesting basis classes F, the infimum is zero.)

Instead of AdaBoost, consider a regularized version of its criterion.

Previous results: Regularized versions

1. Minimize
$$\mathbf{E}_n \exp(-Yf(X))$$
over $f \in \gamma_n \mathrm{co}(F)$, the scaled (by γ_n) convex hull of F.

2. Minimize
$$\mathbf{E}_n \exp(-Yf(X)) + \lambda_n \|f\|_*,$$
over $f \in \mathrm{span}(F)$, where $\|f\|_* = \inf\{\gamma : f \in \gamma \mathrm{co}(F)\}$.

For suitable choices of the parameters (γ_n and λ_n), these algorithms are universally consistent.

(Lugosi and Vayatis, 2004), (Zhang, 2004)

Previous results: Bounded step size

```
function AdaBoostwithBoundedStepSize(S_n, T)
    f_0 := 0
    for t from 1, ..., T
        (α_t, h_t) := arg min_{α∈ℝ, h∈F} (1/n) Σ_{i=1}^n exp(−y_i (f_{t−1}(x_i) + α h(x_i)))
        f_t := f_{t−1} + min{α_t, ε} h_t
    return f_T
```

function AdaBoostwithBoundedStepSize(S_n, T)

$f_0 := 0$

for t from $1, \ldots, T$

$$(\alpha_t, h_t) := \arg \min_{\alpha \in \mathbb{R}, h \in F} \frac{1}{n} \sum_{i=1}^{n} \exp\left(-y_i \left(f_{t-1}(x_i) + \alpha h(x_i)\right)\right)$$

$$f_t := f_{t-1} + \min\{\alpha_t, \epsilon\} h_t$$

return f_T

For suitable choices of the parameters ($T = T_n$ and $\epsilon = \epsilon_n$), this algorithm is universally consistent.

(Zhang and Yu, 2005), (Bickel, Ritov, Zakai, 2006)

Previous results about AdaBoost

AdaBoost greedily minimizes

$$\mathbf{E}_n \exp(-Yf(X)) = \frac{1}{n}\sum_{i=1}^{n} \exp(-Y_i f(X_i))$$

over $f \in \mathrm{span}(F)$.

- What is f_n?
  The function returned by AdaBoost after t_n steps.
- What is t_n?
  Note: The infimum is often zero. Don't want t_n too large.

Previous result about AdaBoost: 'Process consistency'

Theorem: [Jiang, 2004] For all probability distributions P satisfying certain smoothness assumptions,
there is a sequence t_n such that $f_n =$`AdaBoost(`S_n, t_n`)` satisfies

$$L(f_n) \stackrel{a.s.}{\to} L^*.$$

- Conditions on the distribution P are unnatural and cannot be checked.
- How should the stopping time t_n grow with sample size n?
  Does it need to depend on the distribution P?
- Rates?

AdaBoost is Universally Consistent

- Previous results
- The key theorem:
  Universal consistency for sublinearly increasing stopping times.
- Idea of proof
- Open questions

The key theorem

- Assume $d_{VC}(F) < \infty$
  Otherwise AdaBoost must stop and fail after one step.
- Assume
  $$\lim_{\lambda \to \infty} \inf \{R(f) : f \in \lambda \text{co}(F)\} = R^*,$$
  where
  $$R(f) = \mathbf{E} \exp(-Y f(X)), \qquad R^* = \inf_f R(f).$$
  That is, the approximation error is zero.
  For example, F is linear threshold functions, or binary trees with axis orthogonal decisions in $\mathbb{R}^d$ and at least $d + 1$ leaves.

The key theorem

Theorem: If

$$
\begin{aligned}
d_{VC}(F) &< \infty, \\
R_\phi^* &= \lim_{\lambda \to \infty} \inf \left\{ R_\phi(f) : f \in \lambda \mathrm{co}(F) \right\}, \\
t_n &\to \infty \\
t_n &= O(n^{1-\alpha}) \qquad \text{for some } \alpha > 0,
\end{aligned}
$$

then AdaBoost is universally consistent.

The key theorem: Idea of proof

We show $R(f_{t_n}) \to R^*$, which implies $L(f_{t_n}) \to L^*$, since the loss function $\alpha \mapsto \exp(-\alpha)$ is classification calibrated.

Step 1. Notice that we can clip f_{t_n}:
If we define $\pi_\lambda(f)$ as $x \mapsto \max\{-\lambda, \min\{\lambda, f(x)\}\}$, then

$$R(\pi_\lambda(f_{t_n})) \to R^* \implies L(\pi_\lambda(f_{t_n})) \to L^* \implies L(f_{t_n}) \to L^*.$$

We will need to relax the clipping ($\lambda_n \to \infty$).

The key theorem: Idea of proof

Step 2. Use VC-theory (for clipped combinations of t functions from F) to show that, with high probability,

$$R(\pi_\lambda(f_t)) \leq R_n(\pi_\lambda(f_t)) + c(\lambda)\sqrt{\frac{d_{VC}(F)t\log t}{n}},$$

where R_n is the empirical version of R,

$$R_n(f) = \mathbf{E}_n \exp(-Yf(X)).$$

The key theorem: Idea of proof

Step 3. The clipping only hurts for small values of the exponential criterion:

$$R_n(\pi_\lambda(f_t)) \leq R_n(f_t) + e^{-\lambda}.$$

The key theorem: Idea of proof

Step 4. Apply numerical convergence result of (Bickel et al, 2006): For any comparison function $\bar{f} \in F_\lambda$,

$$R_n(f_t) \leq R_n(\bar{f}) + \epsilon(\lambda, t).$$

The key theorem: Idea of proof

Step 5. Apply VC-theory again to relate $R_n(\bar{f})$ to $R(\bar{f})$.

Choosing $\lambda_n \to \infty$ suitably slowly, we can choose $\bar{f}_n$ so that $R(\bar{f}_n) \to R^*$ (by assumption), and then for $t = O(n^{1-\alpha})$, we have the result.

Open Problems

- Other loss functions?
  e.g., LogitBoost uses $\alpha \mapsto \log(1+\exp(-2\alpha))$ in place of $\exp(-\alpha)$.
  (The difficulty is the behaviour of the second derivative of R_n in the direction of a basis function. For the numerical convergence results, we want it large whenever R_n is large.)
- Real-valued basis functions?
  (The same issue arises.)
- Rates?
  The bottleneck is the rate of decrease of $R_n(f_t)$. The (Bickel et al, 2006) result ensures it decreases to $\bar{f}$ as $\log^{-1/2} t$. This seems pessimistic.

AdaBoost is Universally Consistent

- Previous results
- The key theorem:
  Universal consistency for sublinearly increasing stopping times.
- Idea of proof
- Open questions

Slides at http://www.cs.berkeley.edu/ bartlett