AdaBoost and other Large Margin Classifiers: Convexity in Classification

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

Joint work with
Mike Jordan, Jon McAuliffe
Mikhail Traskin.

slides at http://www.cs.berkeley.edu/∼bartlett

The Pattern Classification Problem

- i.i.d. $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{X} \times \{\pm 1\}$.
- Use data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose $f_n : \mathcal{X} \to \mathbb{R}$ with small risk,

$$R(f_n) = \Pr\left(\operatorname{sign}(f_n(X)) \neq Y\right) = \mathbf{E}\ell(Y, f(X)).$$

- Natural approach: minimize empirical risk,

$$\hat{R}(f) = \hat{\mathbf{E}}\ell(Y, f(X)) = \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

- Often intractable...
- Replace 0-1 loss, ℓ, with a convex surrogate, ϕ.

Large Margin Algorithms

- Consider the margins, $Yf(X)$.
- Define a margin cost function $\phi : \mathbb{R} \to \mathbb{R}^+$.
- Define the **ϕ-risk** of $f : \mathcal{X} \to \mathbb{R}$ as $R_\phi(f) = \mathbf{E}\phi(Yf(X))$.
- Choose $f \in \mathcal{F}$ to minimize ϕ-risk.
  (e.g., use data, $(X_1, Y_1), \ldots, (X_n, Y_n)$, to minimize **empirical ϕ-risk**,

$$\hat{R}_\phi(f) = \hat{\mathbf{E}}\phi(Yf(X)) = \frac{1}{n}\sum_{i=1}^{n}\phi(Y_i f(X_i)),$$

  or a regularized version.)

Large Margin Algorithms

- Adaboost:
  - $\mathcal{F} = \text{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
  - Minimizes $\hat{R}_\phi(f)$ using greedy basis selection, line search:

$$f_{t+1} = f_t + \alpha_{t+1} g_{t+1},$$

$$\hat{R}_\phi(f_t + \alpha_{t+1} g_{t+1}) = \min_{\alpha \in \mathbb{R}, g \in \mathcal{G}} \hat{R}_\phi(f_t + \alpha g).$$

Large Margin Algorithms

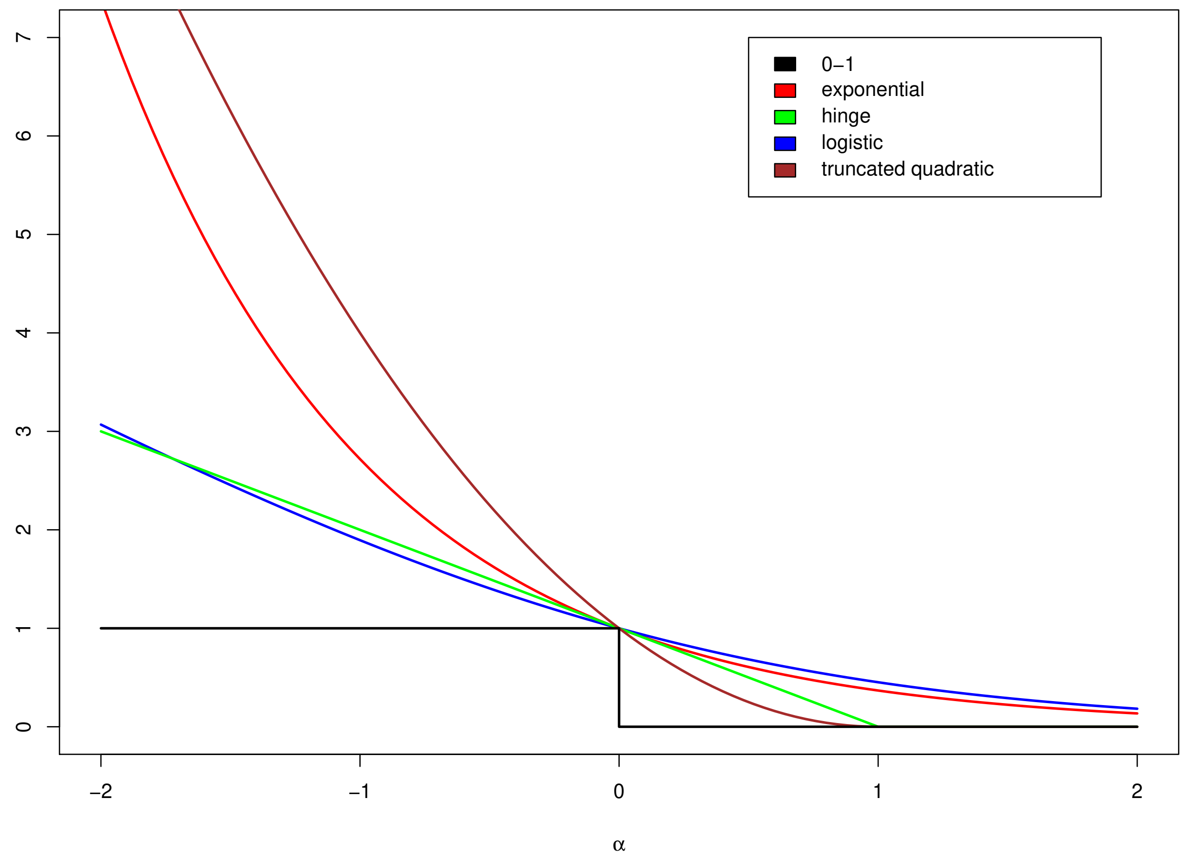
- Many other variants
  - Support vector machines with 1-norm soft margin.
    * $\mathcal{F}$ = ball in reproducing kernel Hilbert space, $\mathcal{H}$.
    * $\phi(\alpha) = \max(0, 1-\alpha)$.
    * Algorithm minimizes $\hat{R}_\phi(f) + \lambda \|f\|_{\mathcal{H}}^2$.
  - Neural net classifiers

    $\phi(\alpha) = \max(0, (0.8-\alpha)^2)$.
  - L2Boost, LS-SVMs

    $\phi(\alpha) = (1-\alpha)^2$.
  - Logistic regression

    $\phi(\alpha) = \log(1 + \exp(-2\alpha))$.

Large Margin Algorithms



Statistical Consequences of Using a Convex Cost

- Is AdaBoost universally consistent? Other ϕ?
  - (Lugosi and Vayatis, 2004), (Mannor, Meir and Zhang, 2002): regularized boosting.
  - (Jiang, 2004): process consistency of AdaBoost, for certain probability distributions.
  - (Zhang, 2004), (Steinwart, 2003): SVM.

Statistical Consequences of Using a Convex Cost

- How is risk related to ϕ-risk?
  - (Lugosi and Vayatis, 2004), (Steinwart, 2003): asymptotic.
  - (Zhang, 2004): comparison theorem.

Overview

- Relating excess risk to excess ϕ-risk.
  - ψ-transform: best possible bound.
  - conditions on ϕ.
- Universal consistency of AdaBoost.

Definitions and Facts

$$R(f) = \Pr\left(\mathrm{sign}(f(X)) \neq Y\right) \qquad R^* = \inf_f R(f) \qquad \text{risk}$$

$$R_\phi(f) = \mathbb{E}\phi(Yf(X)) \qquad R_\phi^* = \inf_f R_\phi(f) \qquad \phi\text{-risk}$$

$$\eta(x) = \Pr(Y = 1|X = x) \qquad \text{conditional probability.}$$

- η defines an optimal classifier: $R^* = R(\mathrm{sign}(\eta(x) - 1/2))$.

Definitions and Facts

$R(f) = \Pr\left(\mathrm{sign}(f(X)) \neq Y\right)$	$R^* = \inf_f R(f)$	risk
$R_\phi(f) = \mathbb{E}\phi(Yf(X))$	$R_\phi^* = \inf_f R_\phi(f)$	ϕ-risk
$\eta(x) = \Pr(Y = 1 \vert X = x)$	conditional probability.	

- η defines an optimal classifier: $R^* = R(\mathrm{sign}(\eta(x) - 1/2))$.

Notice: $R_\phi(f) = \mathbb{E}\left(\mathbb{E}\left[\phi(Yf(X))|X\right]\right)$, and conditional ϕ-risk is:

$$\mathbb{E}\left[\phi(Yf(X))|X = x\right] = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).$$

Definitions

Conditional ϕ-risk:

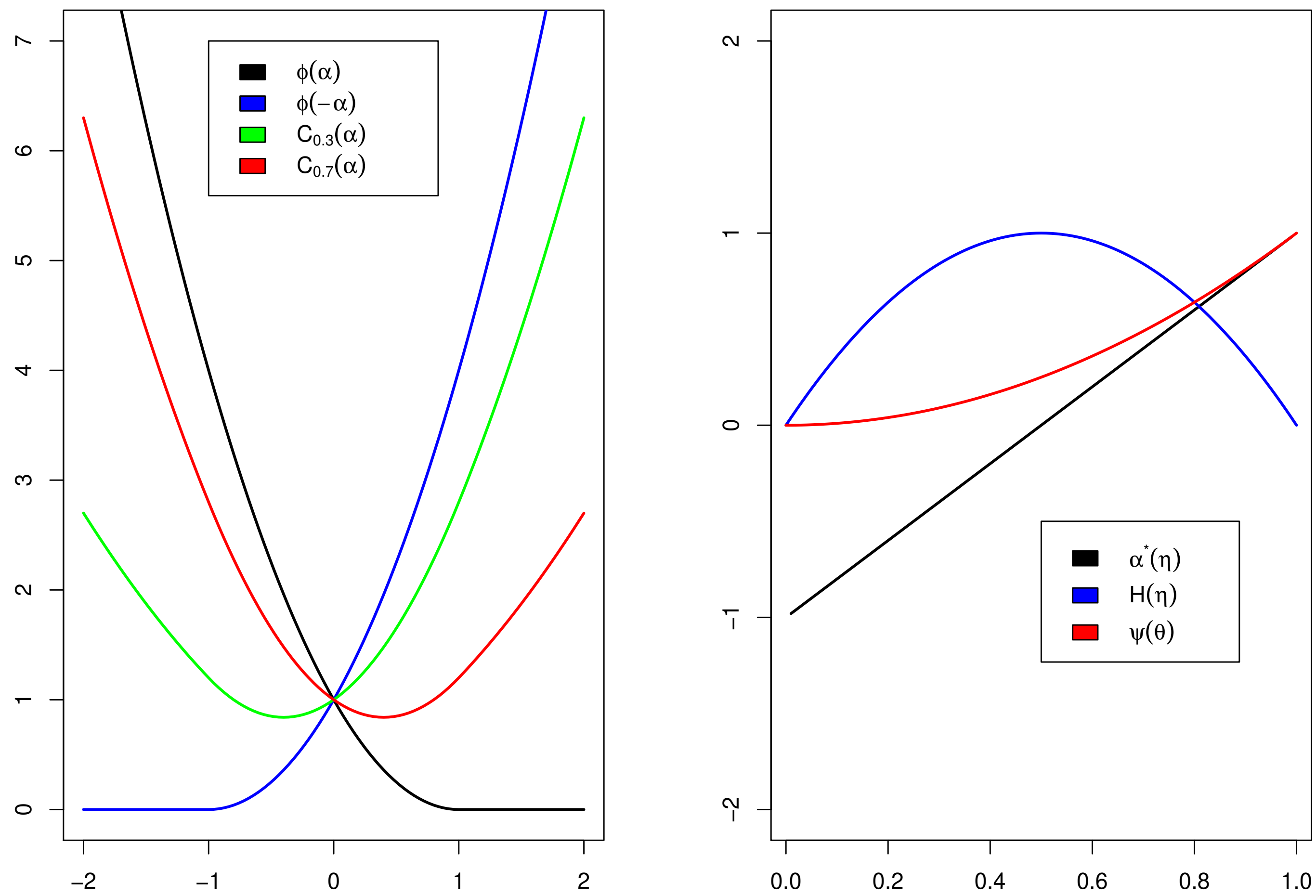
$$\mathbb{E}\left[\phi(Yf(X))|X=x\right]=\eta(x)\phi(f(x))+(1-\eta(x))\phi(-f(x)).$$

Optimal conditional ϕ-risk for $\eta\in[0,1]$:

$$H(\eta)=\inf_{\alpha\in\mathbb{R}}\left(\eta\phi(\alpha)+(1-\eta)\phi(-\alpha)\right).$$

$$R_\phi^*=\mathbb{E}H(\eta(X)).$$

Optimal Conditional ϕ-risk: Example



Definitions

Optimal conditional ϕ-risk for $\eta \in [0,1]$:

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).$$

Optimal conditional ϕ-risk with incorrect sign:

$$H^-(\eta) = \inf_{\alpha : \alpha(2\eta-1) \leq 0} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).$$

Note: $H^-(\eta) \geq H(\eta)$ $\qquad$ $H^-(1/2) = H(1/2)$.

Definitions

$$H(\eta) = \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right)$$

$$H^-(\eta) = \inf_{\alpha : \alpha(2\eta - 1) \leq 0} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).$$

Definition: ϕ is **classification-calibrated** if, for $\eta \neq 1/2$,

$$H^-(\eta) > H(\eta).$$

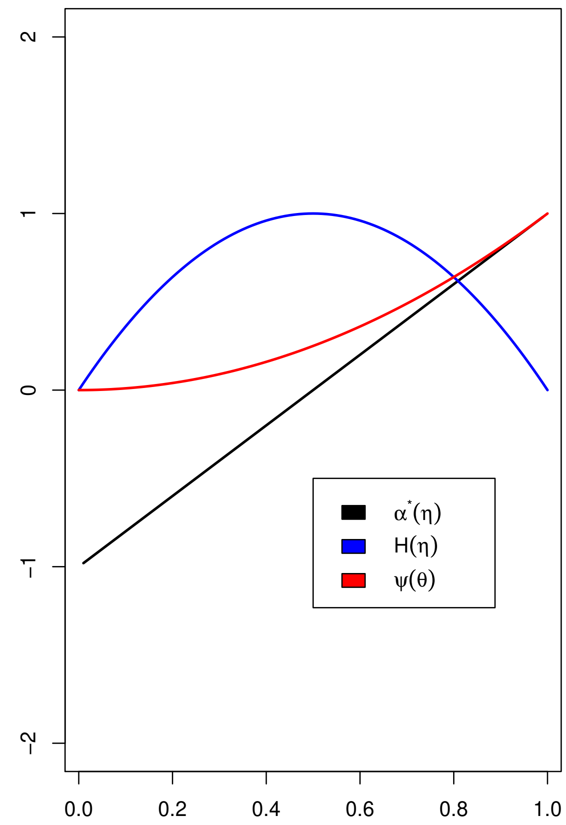
i.e., pointwise optimization of conditional ϕ-risk leads to the correct sign. (c.f. Lin (2001))

The ψ transform

Definition: Given convex ϕ, define $\psi : [0,1] \to [0,\infty)$ by

$$\psi(\theta) = \phi(0) - H\left(\frac{1+\theta}{2}\right).$$

(The definition is a little more involved for non-convex ϕ.)



The Relationship between Excess Risk and Excess ϕ-risk

Theorem:

1. For any P and f, $\quad \psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.
2. For $|\mathcal{X}| \geq 2$, $\epsilon > 0$ and $\theta \in [0, 1]$, there is a P and an f with

$$R(f) - R^* = \theta$$

$$\psi(\theta) \leq R_\phi(f) - R_\phi^* \leq \psi(\theta) + \epsilon.$$

3. The following conditions are equivalent:
   (a) ϕ is classification calibrated.
   (b) $\psi(\theta_i) \to 0$ iff $\theta_i \to 0$.
   (c) $R_\phi(f_i) \to R_\phi^*$ implies $R(f_i) \to R^*$.

Classification-calibrated ϕ

If ϕ is classification-calibrated, then

$$\psi(\theta_i) \to 0 \text{ iff } \theta_i \to 0.$$

Since the function ψ is always convex, in that case it is strictly increasing and so has an inverse.

Thus, we can write

$$R(f) - R^* \leq \psi^{-1}\left(R_\phi(f) - R_\phi^*\right).$$

Excess Risk Bounds: Proof Idea

Facts:

- $H(\eta), H^{-}(\eta)$ are symmetric about $\eta = 1/2$.
- $H(1/2) = H^{-}(1/2)$, hence $\psi(0) = 0$.
- $\psi(\theta)$ is convex.
- $\psi(\theta) = H^{-}\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right)$.

Excess Risk Bounds: Proof Idea

Excess risk of $f : \mathcal{X} \to \mathbb{R}$ is

$$R(f) - R^* = \mathbb{E}\left(\mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right).$$

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) \qquad\qquad (\psi \text{ convex}, \psi(0) = 0)\\
&\leq \mathbb{E}\left(\mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right)\\
&= \mathbb{E}\Big(\mathbf{1}\left[\operatorname{sign}(f(X)) \neq \operatorname{sign}(\eta(X) - 1/2)\right] \big(H^-(\eta(X)) - H(\eta(X))\big)\Big)\\
&\leq \mathbb{E}\left(\phi(Yf(X)) - H(\eta(X))\right)\\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Excess risk of $f : \mathcal{X} \to \mathbb{R}$ is

$$R(f) - R^* = \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right).$$

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) && \text{(definition of } \psi\text{)} \\
&\leq \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right) \\
&= \mathbb{E}\Big(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \big(H^-(\eta(X)) - H(\eta(X))\big)\Big) \\
&\leq \mathbb{E}\left(\phi(Yf(X)) - H(\eta(X))\right) \\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Excess risk of $f : \mathcal{X} \to \mathbb{R}$ is

$$R(f) - R^* = \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right).$$

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) \qquad\qquad \text{(H^- minimizes conditional ϕ-risk)}\\
&\leq \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right)\\
&= \mathbb{E}\Big(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \big(H^-(\eta(X)) - H(\eta(X))\big)\Big)\\
&\leq \mathbb{E}\left(\phi(Yf(X)) - H(\eta(X))\right)\\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Excess Risk Bounds: Proof Idea

Excess risk of $f : \mathcal{X} \to \mathbb{R}$ is

$$R(f) - R^* = \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] |2\eta(X) - 1|\right).$$

Thus,

$$\begin{aligned}
&\psi(R(f) - R^*) \qquad\qquad \text{(definition of } R_\phi\text{)}\\
&\leq \mathbb{E}\left(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \psi\left(|2\eta(X) - 1|\right)\right)\\
&= \mathbb{E}\Big(\mathbf{1}\left[\mathrm{sign}(f(X)) \neq \mathrm{sign}(\eta(X) - 1/2)\right] \big(H^-(\eta(X)) - H(\eta(X))\big)\Big)\\
&\leq \mathbb{E}\left(\phi(Yf(X)) - H(\eta(X))\right)\\
&= R_\phi(f) - R_\phi^*.
\end{aligned}$$

Classification-calibrated ϕ

Theorem: If ϕ is convex,

$$\phi \text{ is classification calibrated} \Leftrightarrow \begin{cases} \phi \text{ is differentiable at } 0 \\ \phi'(0) < 0. \end{cases}$$

Theorem: If ϕ is classification calibrated,

$\exists \gamma > 0, \forall \alpha \in \mathbb{R},$

$$\gamma \phi(\alpha) \geq \mathbf{1}\left[\alpha \leq 0\right].$$

Overview

- Relating excess risk to excess ϕ-risk.
- Universal consistency of AdaBoost.
  - The approximation/estimation decomposition.
  - AdaBoost: Previous results.
  - Universal consistency.

The Approximation/Estimation Decomposition

Algorithm chooses

$$f_n = \arg\min_{f \in \mathcal{F}_n} \hat{R}_\phi(f) + \lambda_n \Omega(f).$$

We can decompose the excess risk estimate as

$$\begin{aligned}\psi\left(R(f_n) - R^*\right) &\le R_\phi(f_n) - R_\phi^* \\ &= \underbrace{R_\phi(f_n) - \inf_{f \in \mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}} .\end{aligned}$$

The Approximation/Estimation Decomposition

$$\psi\left(R(f_n) - R^*\right) \le R_\phi(f_n) - R^*_\phi$$
$$= \underbrace{R_\phi(f_n) - \inf_{f\in\mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f\in\mathcal{F}_n} R_\phi(f) - R^*_\phi}_{\text{approximation error}}.$$

- Approximation and estimation errors are in terms of R_ϕ, not R.
- Like a regression problem.
- With a rich class and appropriate regularization, $R_\phi(f_n) \to R^*_\phi$. (e.g., $\mathcal{F}_n$ gets large slowly, or $\lambda_n \to 0$ slowly.)
- Universal consistency ($R(f_n) \to R^*$) iff ϕ is classification calibrated.

Overview

- Relating excess risk to excess ϕ-risk.
- Universal consistency of AdaBoost.
  - The approximation/estimation decomposition.
  - AdaBoost: Previous results.
  - Universal consistency.

AdaBoost

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

Number of iterations, T

function AdaBoost(S_n, T)

$f_0 := 0$

for t from $1, \ldots, T$

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

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

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,

$$R(f_n) \xrightarrow{a.s} R^*,$$

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

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

Previous results: Regularized versions

AdaBoost greedily minimizes

$$\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n}\exp(-Y_i f(X_i))$$

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 $\hat{R}_\phi(f)$ over $f \in \gamma_n \mathrm{co}(F)$, the scaled convex hull of F.
2. Minimize
$$\hat{R}_\phi(f) + \lambda_n \|f\|_*,$$
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` $1, \ldots, T$

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

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

`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

$$\hat{R}_\phi(f) = \frac{1}{n}\sum_{i=1}^{n}\exp(-Y_i f(X_i))$$

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

- Consider AdaBoost with early stopping:
  f_n is the function returned by AdaBoost after t_n steps.
- How should we choose 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 a (suitable) basis class defined on $\mathbb{R}^d$, and 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

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

- 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?

Overview

- Relating excess risk to excess ϕ-risk.
- Universal consistency of AdaBoost.
  - The approximation/estimation decomposition.
  - AdaBoost: Previous results.
  - Universal consistency.

The key theorem

- Assume $d_{VC}(F) < \infty$
  Otherwise AdaBoost must stop and fail after one step.

- Assume
  $$\lim_{\lambda \to \infty} \inf \{R_\phi(f) : f \in \lambda \mathrm{co}(F)\} = R_\phi^*,$$
  where
  $$R_\phi(f) = \mathbf{E} \exp(-Y f(X)), \qquad R_\phi^* = \inf_f R(f).$$
  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

$$d_{VC}(F) < \infty,$$

$$R_\phi^* = \lim_{\lambda \to \infty} \inf \left\{ R_\phi(f) : f \in \lambda \mathrm{co}(F) \right\},$$

$$t_n \to \infty$$

$$t_n = O(n^{1-\alpha}) \qquad \text{for some } \alpha > 0,$$

then AdaBoost is universally consistent.

The key theorem: Idea of proof

We show $R_\phi(f_{t_n}) \to R_\phi^*$, which implies $R(f_{t_n}) \to R^*$, 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

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

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,

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

where $\hat{R}_\phi$ is the empirical version of R_ϕ,

$$\hat{R}_\phi(f) = \mathbf{E}_n \exp(-Yf(X)).$$

The key theorem: Idea of proof

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

$$\hat{R}_\phi(\pi_\lambda(f_t)) \leq \hat{R}_\phi(f_t) + e^{-\lambda}.$$

The key theorem: Idea of proof

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

$$\hat{R}_\phi(f_t) \leq \hat{R}_\phi(\bar{f}) + \epsilon(\lambda, t).$$

Here, we exploit an attractive property of the exponential loss function and the fact that classifiers are binary-valued:
The second derivative of $\hat{R}_\phi$ in a basis direction is large whenever $\hat{R}_\phi$ is large. This keeps the steps taken by AdaBoost from being too large.

The key theorem: Idea of proof

Step 5. Apply VC-theory again to relate $\hat{R}_\phi(\bar{f})$ to $R_\phi(\bar{f})$.

Choosing $\lambda_n \to \infty$ suitably slowly, we can choose $\bar{f}_n$ so that $R_\phi(\bar{f}_n) \to R_\phi^*$ (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 $\hat{R}_\phi$ in the direction of a basis function. For the numerical convergence results, we want it large whenever $\hat{R}_\phi$ is large.)
- Real-valued basis functions?
  (The same issue arises.)
- Rates?
  The bottleneck is the rate of decrease of $\hat{R}_\phi(f_t)$. The numerical convergence result ensures it decreases to $\bar{f}$ as $\log^{-1/2} t$.
  This seems pessimistic.

Overview

- Relating excess risk to excess ϕ-risk.
- Universal consistency of AdaBoost.

slides at http://www.cs.berkeley.edu/∼bartlett