AdaBoost and other Large Margin Classifiers: Convexity in Pattern Classification

Peter Bartlett

Department of Statistics and Division of Computer Science

UC Berkeley

Joint work with Mikhail Traskin.

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,

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

- 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

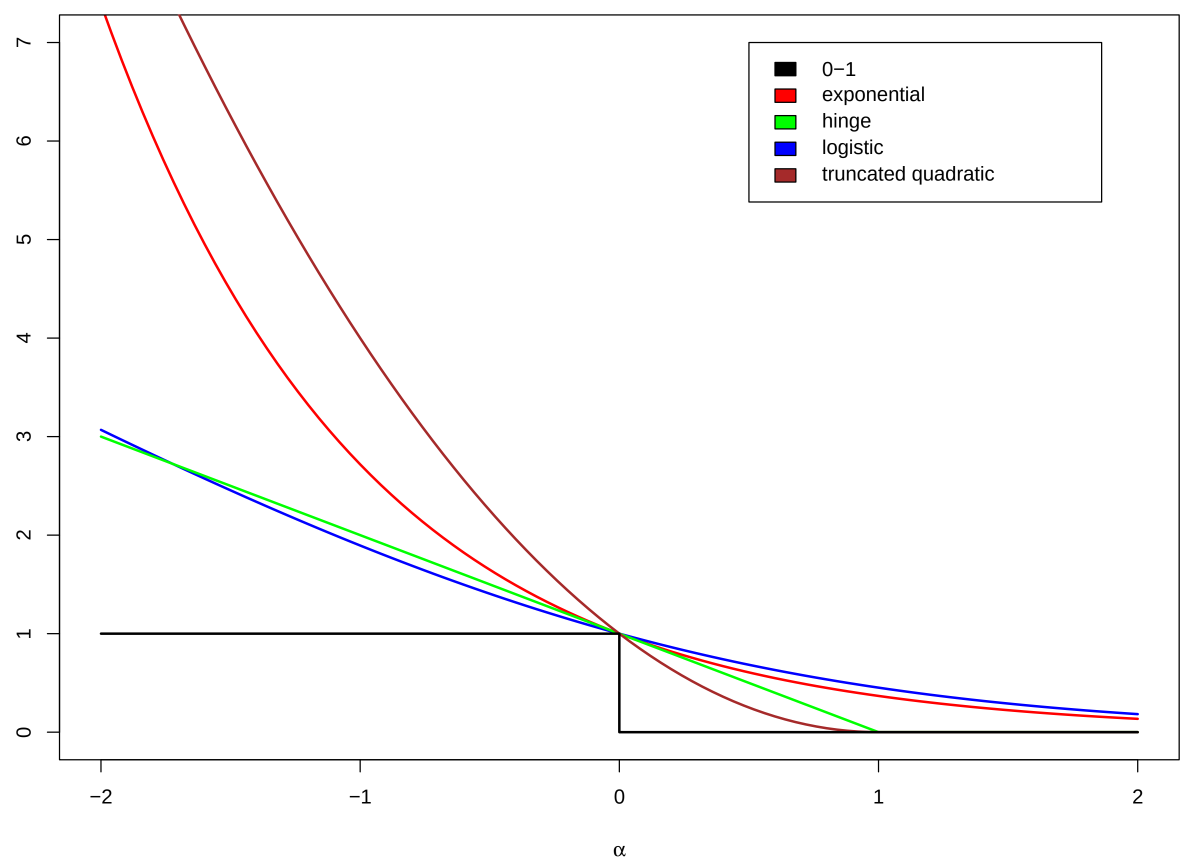
- Adaboost:
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
  - Minimizes $\hat{R}_\phi(f)$ using greedy basis selection, line search:

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

Large Margin Algorithms

- Many other variants
  - Support vector machines:
    - $\mathcal{F}$ = 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
  - L2Boost, LS-SVMs
  - Logistic regression

Large Margin Algorithms



Overview

- Convex cost versus risk.
- Universal consistency.
- Consistency of AdaBoost.
- Future directions: Prediction in adversarial environments

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.	

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

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

Definitions

Conditional ϕ-risk:

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

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

Difference: $$\psi(\theta) = H^-\left(\frac{1+\theta}{2}\right) - H\left(\frac{1+\theta}{2}\right).$$

The Relationship between Excess Risk and Excess ϕ-risk

Theorem:

1. For any P and f, $\psi(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$.
2. This inequality cannot be improved:
   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) For $\eta \neq 1/2$, $H^-(\eta) > H(\eta)$. *'classification calibrated'*
   (b) $\psi(\theta_i) \to 0$ iff $\theta_i \to 0$.
   (c) $R_\phi(f_i) \to R_\phi^*$ implies $R(f_i) \to R^*$.

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., $f_n = \texttt{AdaBoost}((X_1, Y_1), \ldots, (X_n, Y_n), t_n)$.

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

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

Recall that R is the risk and R^* is the Bayes risk:

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

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 suitable method, $R_\phi(f_n) \to R_\phi^*$.
- Universal consistency ($R(f_n) \to R^*$) iff ϕ is *classification calibrated*.

Overview

- Convex cost versus risk.
- Universal consistency.
- Consistency of AdaBoost.
- Future directions: Prediction in adversarial environments

AdaBoost

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

Number of iterations, T

Class of basis functions, $\mathcal{G}$

function AdaBoost(S_n, T):

$f_0 := 0$

for t from $1, \ldots, T$

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

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

return f_T

AdaBoost chooses f_T from the *linear span* of $\mathcal{G}$.

Previous results

Instead, we could consider a regularized version of AdaBoost:

1. Minimize $\hat{R}_\phi(f)$ over $\mathcal{F}_n = \gamma_n \mathrm{co}(\mathcal{G})$, the scaled convex hull of $\mathcal{G}$.
2. Minimize $\hat{R}_\phi(f) + \lambda_n \|f\|_*$, over $\mathrm{span}(\mathcal{G})$, where $\|f\|_* = \inf\{\gamma : f \in \gamma \mathrm{co}(\mathcal{G})\}$.
3. AdaBoost with step-size bounded: $\alpha_t \leq \beta_{n,t}$.

For suitable choices of the parameters (γ_n, λ_n, β_n), these algorithms are universally consistent. (Lugosi and Vayatis, 2004), (Zhang, 2004), (Zhang and Yu, 2005), (Bickel, Ritov, Zakai, 2006)

For $\mathcal{X} \subset \mathbb{R}^d$, if log odds ratio, $\log(\eta(x)/(1-\eta(x)))$, is smooth, then AdaBoost estimates it asymptotically. (Jiang, 2004).

Universal consistency of AdaBoost

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.

Overview

- Convex cost versus risk.
- Universal consistency.
- Consistency of AdaBoost.
- Future directions: Prediction in adversarial environments

Future directions: Prediction in adversarial environments

Prediction game:

1. see side information $x_t \in \mathcal{X}$,
2. make prediction $\hat{y}_t \in \mathcal{A}$,
3. see outcome $y_t \in \mathcal{Y}$ and incur loss $\ell(y_t, \hat{y}_t)$.

Aim: choose $\hat{y}_t$ so that, for all data sequences, the *cumulative regret*

$$\sum_{t=1}^{n} \ell(y_t, \hat{y}_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} \ell(y_t, f(x_t))$$

is not too large.

Future directions: Prediction in adversarial environments

Applications:

1. Computer security
   - detection of anomalous network traffic
   - virus detection
   - spam filtering
2. Internet search
3. Financial portfolio optimization

Future directions: Prediction in adversarial environments

Adversary: controls some data, and benefits if predictions are incorrect.

- Performance of large margin classifiers?
- Performance of Bayesian methods?

Overview

- Convex cost versus risk.
- Universal consistency.
- Consistency of AdaBoost.
- Future directions: Prediction in adversarial environments