Joint work with Mikhail Traskin.

UC Berkeley
Department of Statistics and Division of Computer Science

Peter Bartlett

Convexity in Pattern Classification
AdaBoost and other Large Margin Classifiers:
The Pattern Classification Problem

\[ \phi \] Replace 0-1 loss \( \varphi \) with a convex surrogate \( \phi \).

Often intractable.

\[ \left( (\mathbf{X})f, \mathbf{y} \right) \mathbb{P} \sum_{i=1}^{n} \frac{1}{n} = \left( (\mathbf{X})f, \mathbf{y} \right) \mathbb{E} = (f)_{\mathbb{P}} \]

Natural approach: minimize empirical risk,

\[ \left( (\mathbf{X})f, \mathbf{y} \right) \mathbb{P} = (\mathbf{y} \neq (\mathbf{X})^u f) \mathbb{P} = (u^f)_{\mathbb{P}} \]

Risk,

\[ \mathbb{R} \leftarrow \mathcal{X} : u \mathbf{f} \text{ to choose } (u \mathbf{X}, u \mathbf{X}) \cdots (u \mathbf{Y}, u \mathbf{X}) \] from \( (u \mathbf{Y}, u \mathbf{X}) \cdots (u \mathbf{Y}, u \mathbf{X}) \), \( (u \mathbf{Y}, u \mathbf{X}) \) i.i.d.

Use data \( \mathbb{R} \) with small

\[ \{1 \pm \} \times \mathcal{X} \] i.i.d.

The Pattern Classification Problem
Consider the margins, $Yf(X)$. 

Define a margin cost function:
\[ \mathcal{R} \to \mathcal{R}^+ \]

Define the empirical risk:
\[ \hat{\mathcal{R}}(f) = \frac{\sum_{i=1}^{n} u_i}{n} \]

Choose $f \in \mathcal{F}$ to minimize empirical risk.

\[ ((X)f\phi) = (f)\phi \]

(\text{e.g., use data, $(X_1, Y_1), \ldots, (X_n, Y_n)$ to minimize empirical risk.)

Define the risk of $f$ as $\mathcal{R} \leftarrow X : f \Rightarrow \phi$.

Define a margin cost function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$.

Consider the margins, $X_f \phi$. 

---

Large Margin Algorithms
\[
\phi_{\mathcal{H}}^* \left( \sum_{t=1}^{T} \alpha_t \phi \right) = \left( \sum_{t=1}^{T+1} \alpha_t \phi \right) \phi_{\mathcal{H}}
\]

Minimizes \( R(f) \) using greedy basis selection, line search:

\[
\langle x - \exp \rangle = \langle x \rangle \phi
\]

For a VC-class \( \delta \)

\[
\text{span}(\delta) = \mathcal{F}
\]

Adaboost
Large Margin Algorithms

Many other variants

Logistic regression

L2Boost, LS-SVMs

Neural net classifiers

Algorithm minimizes

\[ \min_{\mathcal{H}} \| f \|_{\mathcal{Y}} + (f)_{\mathcal{H}} \]

\[ (\alpha - 1, 0, \max (0, 1) = (\alpha)_{\phi} \]

\[ \mathcal{H} = \text{ball in reproducing kernel Hilbert space}, \mathcal{H} \]

Support vector machines:

Future directions: Prediction in adversarial environments
Consistency of AdaBoost
Universal consistency
Convex cost versus risk
\[
((x)f - \phi((x)u - 1)) + ((x)f)\phi(x)u = [x = X|((X)f \Lambda)\phi] E
\]

**Notice:**
\[
R(f) = \inf_{f \in \mathcal{H}} R(f) = \inf_{f \in \mathcal{H}} \mathbb{E}[(x)f(X)]
\]

**Definitions:**

- **Risk:**
  \[(x) \mathbb{P} = (x) R(f) = \mathbb{E}[Yf(X)]
  \]

- **Conditional probability:**
  \[P(x = X|\Lambda = \lambda) = (x) u
  \]

- **Conditional risk:**
  \[\mathbb{E}[Yf(X)|X = x] = (f) R \mathcal{F}(x)
  \]

- **Conditional \(\phi\)-risk:**
  \[((x)f)\phi(X)u = [x = X|((X)f \Lambda)\phi] E
  \]
\[
\cdot \left( \frac{\tau}{\theta + 1} \right) H - \left( \frac{\tau}{\theta + 1} \right)_- H = (\theta) \phi
\]

Difference:

\[
\cdot \left( (\nu - \phi)(\nu - 1) + (\nu)\phi \nu \right) = (\nu) H
\]

with error:

\[
\cdot \left( (\nu - \phi)(\nu - 1) + (\nu)\phi \nu \right) = (\nu) H
\]

Optimal:

Conditional-risk: \( \phi \)
The Relationship between Excess Risk and Excess -risk

**Theorem:**

1. For any $P$ and $f$, \( (f)_{\phi} R \leq (f)_{\phi} (\epsilon) \)

2. This inequality cannot be improved:
   For $j \in \{1, 2\}$, with $P_i$ different, there is a $P$ and an $f$ with $R(f)_{\phi} R = (f)_{\phi} (\epsilon) + (i)_{\phi} (\epsilon)$.

3. The following conditions are equivalent:
   (a) For $i \neq 1 = 2$, $H \leq H - (f)_{\phi}$.
   (b) $(i)_H < (i)_{H} - 2, \epsilon \neq P_i$
   (c) $R(f_i)_{\phi} R \leq (f)_{\phi} (\epsilon)$
   (d) $\theta = (i)_{\phi} (\epsilon) + (i)_{\phi} (\epsilon)$

For $\epsilon > 2, \epsilon \neq 0$, there is a $P$ and $f$ with

2. This inequality cannot be improved:

1. For any $P$ and $f$, \( R(f)_{\phi} R \geq (R(f)_{\phi} R)_{\phi} \)

For $\epsilon > 2, \epsilon \neq 0$, there is a $P$ and $f$ with $R(f)_{\phi} R \geq (R(f)_{\phi} R)_{\phi}$.

Classification calibrated:
Universal Consistency

Assume: i.i.d. data, \((X,Y)\), \((X_1,Y_1)\), \((X_2,Y_2)\), \ldots, \((X_n,Y_n)\) from \(XY\) (with \(Y = f_1g\)).

Consider a method \(f_n = A((X_1,Y_1); \ldots; (X_n,Y_n))\), e.g., \(f_n = \text{AdaBoost}((X_1,Y_1); \ldots; (X_n,Y_n); t_n)\).

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

\[
\mathbf{H} \leftarrow \frac{1}{n} \mathbf{f} \mathbf{H}
\]

Recall that \(R^* \) is the risk and \(\mathbf{H}^*\) is the Bayes risk:

\[
\mathbf{H} \leftarrow \frac{1}{n} \mathbf{f} \mathbf{H}
\]

\[
R^* \leq (uf)H
\]

Consider a method \(f_n = \text{AdaBoost}((u_X^1,u_X^2)\ldots,(u_X^1,u_X^2))\) with \(\mathcal{C}\) and \(\mathcal{X}\) from \(\mathcal{C} \times \mathcal{X}\) (i.i.d. data).

Universal Consistency
The Approximation/Estimation Decomposition

\[ \phi \hat{R} - (u_f) \phi R \geq (\hat{R} - (u_f)R) \phi \]

Approximation and estimation errors are in terms of \( \phi R \), not \( R \).

Like a regression problem.

With a rich class and suitable method, \( R \) is classification calibrated.

Universal consistency \( R \) * if \( (\hat{R} - (u_f)R) \phi R \) changes in terms of \( \phi R \), not \( R \).

Approximation and estimation errors are in terms of \( \phi R \), not \( R \).
Overview

Future directions: Prediction in adversarial environments

Consistency of AdaBoost.

Universal consistency.

Convex cost versus risk.
AdaBoost chooses \( f \) from the linear span of \( \mathcal{G} \).

\[
\text{function AdaBoost}(S, T) :
\]

\[
\begin{align*}
\text{return } \ & f \\
\text{for } t \text{ from } 1 \text{ to } T \\
& f_t := 0 \\
& f_t := f_t + t g_t \\
\end{align*}
\]

\[
\text{for } t \text{ from } 1 \text{ to } T \\
\text{return } f_t \\
\]

\[
\text{class of functions, } \mathcal{G} \text{, number of iterations, } T \\
\text{sample, } u \text{, } S \\
\text{linear span of } \mathcal{G}.
\]
Instead, we could consider a regularized version of AdaBoost:

1. Minimize $\sum_{\mathcal{F}} (f)$ over $\mathcal{F}$, the scaled convex hull of $\mathcal{F}$.

2. Minimize $\sum_{\mathcal{F}} \inf_{f \in \mathcal{F}} \log \|f\|_{\mathcal{F}}$ over span $(\mathcal{F})$, where $\mathcal{F}$ is a family of functions.

3. AdaBoost with step-size bounded: $a_t \leq \frac{1}{t}$.

For suitable choices of the parameters $(n, u)$, the algorithms are universally consistent.

For $X \subset \mathbb{R}^d$, if the log odds ratio, $\log(\frac{1}{1 + x})$, is smooth, then AdaBoost estimates it asymptotically.

For $\mathcal{F} \subset \mathbb{R}$, if the log odds ratio, $\log \|f\|$, is smooth, then

Universal consistency of AdaBoost

Theorem: If $\text{VC}(F) < 1$ and $R = \lim_{n \to \infty} \inf_{f \in \text{co}(F)} R(f) = O(n^{-\alpha})$ for some $\alpha > 0$, then AdaBoost is universally consistent.

\[
\begin{align*}
\text{for some } & \alpha \in (0, 1), \\
(u^{-1} u) O & = u^T
\end{align*}
\]

\[
\infty \leftarrow u^T
\]

\[
\lim_{n \to \infty} \inf f \in (H) \phi H = H_
\]

\[
\infty > (H) \phi H
\]

Universal consistency of AdaBoost
Overview

Convex cost versus risk.

Universal consistency.

Consistency of AdaBoost.

Future directions: Prediction in adversarial environments.

Universal consistency.

Convex cost versus risk.
Prediction Game:
1. sees side information $x_t$
2. makes prediction $y_t$ in $A$
3. sees outcome $y_t$ and incurs loss $\ell(y_t; y_t)$.

Aim: choose $y_t$ so that, for all data sequences, the cumulative regret is not too large:

$$\left(\sum_{t=1}^{\infty} \inf_{f \in F} \sum_{u} \ell(f; y_t) \right) - \left(\sum_{t=1}^{\infty} \ell(y_t; y_t) \right)$$

Future directions: Prediction in adversarial environments.
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 Bayesian methods?
- Performance of large margin classifiers?
Overview

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