Convex methods for classification

Peter Bartlett

Department of Statistics and Division of Computer Science

UC Berkeley

Joint work with

Ambuj Tewari, Mikhail Traskin, Marten Wegkamp.

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

The Pattern Classification Problem

- i.i.d. $(X, Y), (X_1, Y_1), \dots, (X_n, Y_n)$ from $\mathcal{X} \times \{\pm 1\}$.
- Use data $(X_1, Y_1), \dots, (X_n, Y_n)$ 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

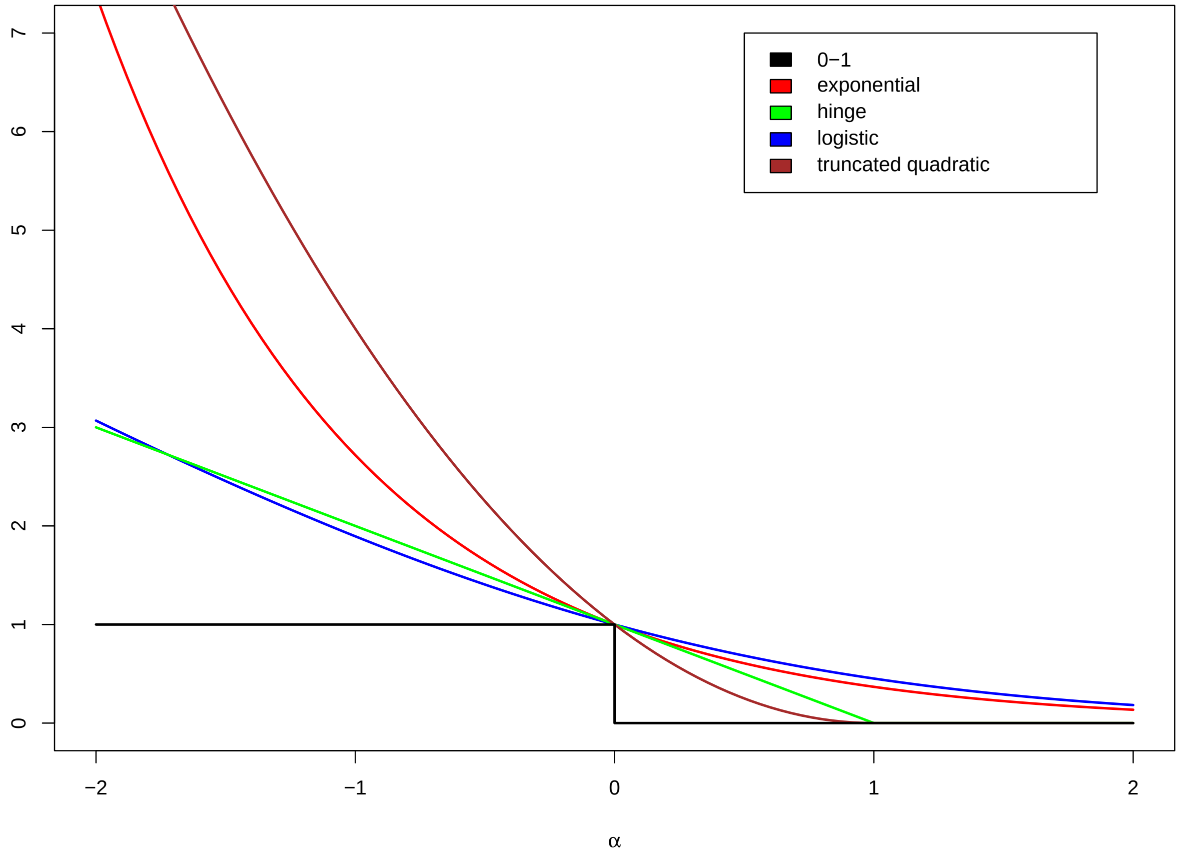
- Adaboost:
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$ for a VC-class $\mathcal{G}$,
  - $\phi(\alpha) = \exp(-\alpha)$,
  - Minimizes $\hat{R}_\phi(f)$ using greedy basis selection, line search:

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

Large Margin Algorithms

- Many other variants
  - Support vector machines:
    - $\mathcal{F}$ = 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

- Review: Convex cost versus risk.
- Universal consistency.
- Classification with a reject option.
- Multiclass generalizations.

Definitions and Facts

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

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

$\eta(x) = \Pr(Y = 1 | X = x)$ conditional probability.

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

$$
\begin{aligned}
H(\eta) &= \inf_{\alpha\in\mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \\
H^{-}(\eta) &= \inf_{\alpha:\alpha(2\eta-1)\leq 0} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right).
\end{aligned}
$$

Definition: We say that 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.

The ψ transform

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

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

(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

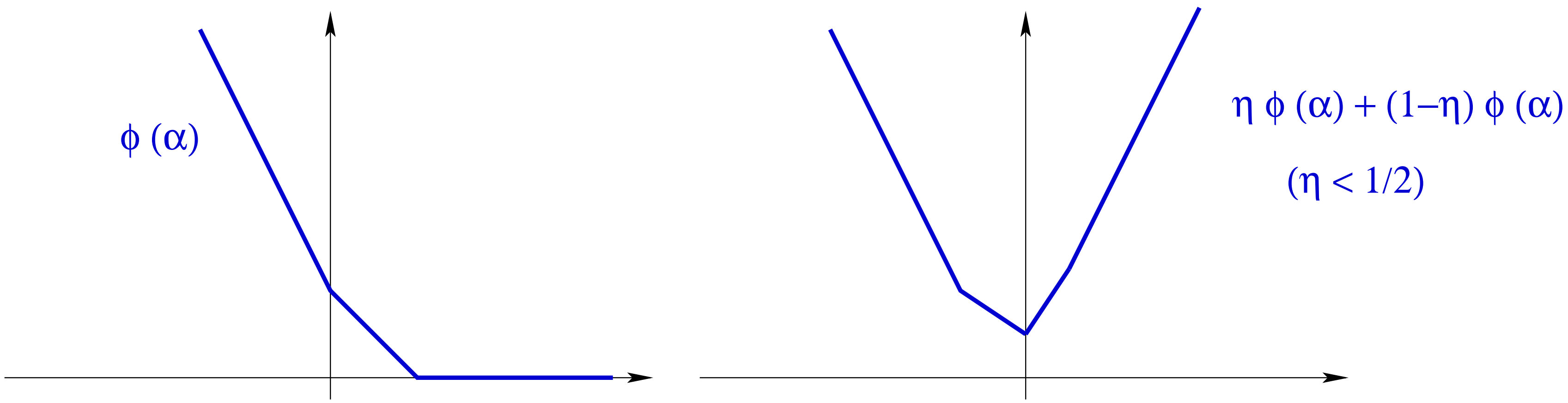
Theorem: If ϕ is convex,

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

Classification-calibrated ϕ

Theorem: If ϕ is convex,

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



Overview

- Review: Convex cost versus risk.
- Universal consistency.
- Classification with a reject option.
- Multiclass generalizations.

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,

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

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

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

The Approximation/Estimation Decomposition

Consider an algorithm that chooses

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

($\hat{R}_\phi(f)$ is empirical ϕ-risk, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}$, and Ω is regularization.)

We can decompose the excess risk estimate as

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

The Approximation/Estimation Decomposition

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

- 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.

Universal Consistency: SVMs

For a Reproducing Kernel Hilbert Space $\mathcal{H}$, choose

$$f_n = \arg\min_{f \in \mathcal{H}_n} \left(\hat{R}_\phi(f) + \lambda_n \|f\|_{\mathcal{H}}^2 \right), \qquad \text{with } \mathcal{H}_n = \{f \in \mathcal{H} : \lambda_n \|f\| \le 1\}.$$

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

If $\mathcal{H}$ is large (for example, a Gaussian kernel on $\mathbb{R}^d$), $\inf_{f \in \mathcal{H}} R_\phi(f) = R_\phi^*$.

For $\lambda_n \to 0$ (suitably slowly), $|\hat{R}_\phi(f_n) - R_\phi(f_n)| \overset{a.s}{\to} 0$.

In that case, $R_\phi(f_n) \overset{a.s}{\to} R_\phi^*$, and universal consistency follows.

Universal Consistency: AdaBoost?

- For SVMs, the regularization term keeps f_n small, which is essential for the uniform convergence result: $|\hat{R}_\phi(f_n) - R_\phi(f_n)| \overset{a.s}{\to} 0$.
- AdaBoost?

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 $f_T$
```

Previous results: Regularized versions

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 span$(\mathcal{G})$, where $\|f\|_* = \inf\{\gamma : f \in \gamma \mathrm{co}(\mathcal{G})\}$.

For suitable choices of the parameters (γ_n and λ_n), these algorithms are universally consistent. (Lugosi and Vayatis, 2004), (Zhang, 2004)

Also bounded step size. (Zhang and Yu, 2005), (Bickel, Ritov, Zakai, 2006)

Previous results: 'Process consistency'

If the log odds ratio, $\log(\eta(x)/(1-\eta(x)))$, is smooth, then it turns out that AdaBoost estimates it in some asymptotic sense:

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

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

Universal consistency of AdaBoost

Theorem: [with Mikhail Traskin]

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.

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.

Idea of proof:

Uniform convergence of clipped t_n-combinations. Clipping does not greatly increase $\hat{R}_\phi$. Then $\hat{R}_\phi(f_{t_n})$ approaches best in an ℓ_*-ball. Then uniform convergence over ℓ_*-balls.

Overview

- Review: Convex cost versus risk.
- Universal consistency.
- Classification with a reject option.
- Multiclass generalizations.

Classification with a reject option

(with Marten Wegkamp)

- Classifier can predict $\{-1, 0, 1\}$. The loss incurred in predicting $\hat{y}$ is

$$\ell(\hat{y}, y) = \begin{cases} 1 & \text{if } \hat{y} \in \{-1, 1\}, \hat{y} \neq y, \\ d & \text{if } \hat{y} = 0, \\ 0 & \text{otherwise.} \end{cases} \qquad \longleftarrow 0 < d \leq 1/2.$$

- Risk of $f : \mathcal{X} \to \{-1, 0, 1\}$ is

$$R(f) = \mathbf{E}\ell(f(X), Y).$$

Classification with a reject option

- Optimal decision rule:

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1 - d, \\ -1 & \text{if } \eta(x) < d, \\ 0 & \text{otherwise.} \end{cases}$$

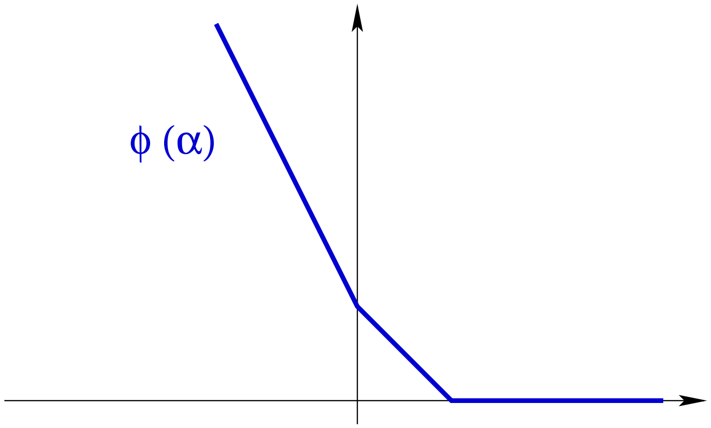
- $d = 1/2$ is the usual binary classification problem.

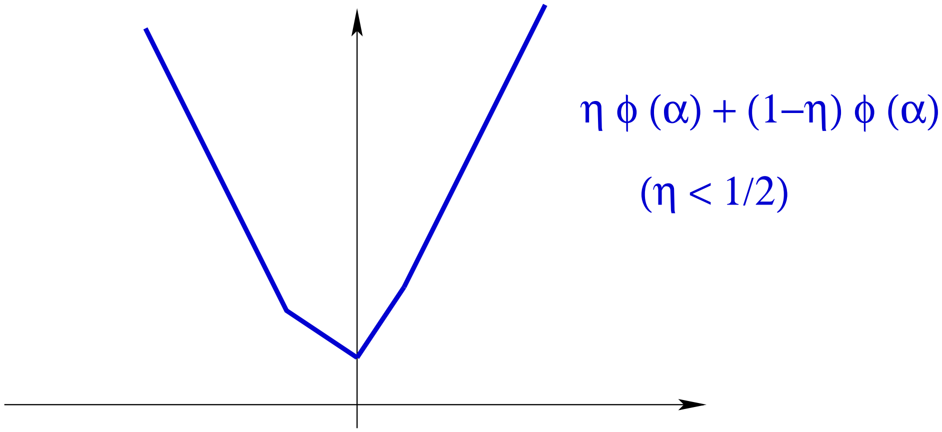
Classification with a reject option

(Herbei and Wegkamp, 2005): Empirical minimization.

We consider a convex alternative that is a generalization of the hinge loss:

$$\phi(\alpha) = \begin{cases} 1 - \alpha(1-d)/d & \text{if } \alpha \le 0, \\ 1 - \alpha & \text{if } 0 < \alpha \le 1, \\ 0 & \text{otherwise.} \end{cases}$$





Classification with a reject option

$$\phi(\alpha) = \begin{cases} 1 - \alpha(1-d)/d & \text{if } \alpha \le 0, \\ 1 - \alpha & \text{if } 0 < \alpha \le 1, \\ 0 & \text{otherwise.} \end{cases}$$

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

Predict using $$g(f_n(x)) = \begin{cases} -1 & \text{if } f_n(x) < -1/2, \\ 1 & \text{if } f_n(x) > 1/2, \\ 0 & \text{otherwise.} \end{cases}$$

Classification with a reject option

Comparison theorem:

$$R(f_n) - R^* \leq 2d(R_\phi(f_n) - R_\phi^*)$$

This generalizes the corresponding result for binary classification with hinge loss ($d = 1/2$):

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

$$R(f_n) - R^* \leq R_\phi(f_n) - R_\phi^*.$$

Classification with a reject option

Comparison theorem:

$$R(f_n) - R^* \le 2d(R_\phi(f_n) - R_\phi^*)$$

Thus, minimizing $R_\phi(\cdot)$ makes sense. If, for example, we choose

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

then the usual arguments show that, for a suitably rich class $\mathcal{F}$ and slowly decreasing regularization coefficient λ_n,

$$R(f_n) \to R^*.$$

Low Noise

The difficulty of a binary classification problem (for example, convergence rate) is determined by the probability that $\eta(X)$ is near $1/2$.

Most favorable case: for some $c > 0$, $\Pr\left(0 < |2\eta(X) - 1| < c\right) = 0$.

Analogous condition here:
$\Pr\left(|\eta(X) - d| < c\right) = \Pr\left(|\eta(X) - (1 - d)| < c\right) = 0$.

Low Noise

Definition: [Tsybakov] The distribution P on $\mathcal{X} \times \{\pm 1\}$ has *noise exponent* $0 \le \alpha < \infty$ if there is a $c > 0$ such that

$$\Pr\left(0 < |2\eta(X) - 1| < \epsilon\right) \le c\epsilon^{\alpha}.$$

- Equivalently, there is a c such that for every $f : \mathcal{X} \to \{\pm 1\}$,

$$\Pr\left(f(X)(\eta(X) - 1/2) < 0\right) \le c\left(R(f) - R^*\right)^{\beta},$$

where $\beta = \dfrac{\alpha}{1+\alpha}$.

- $\alpha = \infty$: for some $c > 0$, $\Pr\left(0 < |2\eta(X) - 1| < c\right) = 0$.

Low Noise

- Tsybakov considered empirical risk minimization in binary classification.
- With the noise assumption, and the Bayes classifier in the function class Tsybakov showed that the empirical risk minimizer has (true) risk converging suprisingly quickly to the minimum.

(More recently, similar results for plug-in methods.)

Risk Bounds with Low Noise: Convex Losses

A similar result is true for strictly convex losses, such as AdaBoost's loss. In these cases, we can improve the comparison inequality:

$$c\left(R(f)-R^*\right)^{\beta}\,\psi\left(\frac{\left(R(f)-R^*\right)^{1-\beta}}{2c}\right) \le R_\phi(f)-R_\phi^*,$$

where $\beta = \dfrac{\alpha}{1+\alpha} \in [0,1]$. (Consider, for example, $\alpha = \infty$.)

Theorem: [Bartlett, Jordan, McAuliffe, 2006] If ϕ has quadratic modulus of convexity, $\Pr\left(0 < |2\eta(X)-1| < c\right) = 0$, and f_n minimizes $\hat{R}_\phi$ over a finite-dimensional function class $\mathcal{F}$, then

$$\mathbf{E}R(f_n) - R^* \le C\left(\inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^* + \frac{\log n}{n}\right).$$

Modified Hinge Loss for Classification with Rejects

A similar result applies for classification with a reject option. **(and SVMs)**
Recall that the critical probabilities in the optimal decision rule are d, $1-d$:

$$f^*(x) = \begin{cases} 1 & \text{if } \eta(x) > 1-d, \\ -1 & \text{if } \eta(x) < d, \\ 0 & \text{otherwise.} \end{cases}$$

Theorem: If ϕ is the modified hinge loss, $\Pr\left(|\eta(X) - d| < c\right) = \Pr\left(|\eta(X) - (1-d)| < c\right) = 0$, and f_n minimizes $\hat{R}_\phi$ over a finite-dimensional function class $\mathcal{F}$, then

$$\mathbf{E}R(f_n) - R^* \leq C\left(\inf_{f\in\mathcal{F}} R_\phi(f) - R_\phi^* + \frac{\log n}{n}\right).$$

Overview

- Review: Convex cost versus risk.
- Universal consistency.
- Classification with a reject option.
- Multiclass generalizations.

Multiclass large margin methods $(|\mathcal{Y}| > 2)$

(with Ambuj Tewari)

Two broad categories of methods:

- Combine several binary classifiers,
- Minimize a cost function defined on a vector space.

We will focus on methods in the second category.

Think of a classifier as a vector valued function $\mathbf{f} : \mathcal{X} \mapsto \mathbb{R}^K$.

For a suitable loss function $L : \mathcal{Y} \times \mathbb{R}^K \to \mathbb{R}_+$, pick $\hat{\mathbf{f}}_n$ by minimizing

$$\frac{1}{n} \sum_{i=1}^{n} L(y_i, \mathbf{f}(x_i)) + \Omega_n(\mathbf{f}) \ .$$

Multiclass large margin methods

A few methods of this kind from the literature:

$(x_+ = \max\{0, x\})$

	$L(y_i, \mathbf{f}(x_i))$
Vapnik; Weston and Watkins; Bredensteiner and Bennett	$\sum_{y' \neq y_i} (f_{y'}(x_i) - f_{y_i}(x_i) + 1)_+$
Crammer and Singer; Taskar et al	$\max_{y' \neq y_i} (f_{y'}(x_i) - f_{y_i}(x_i) + 1)_+$
Lee, Lin and Wahba	$\sum_{y' \neq y_i} (1 + f_{y'}(x_i))_+$ with sum-to-zero constraint, $\sum_y f_y(x) = 0$

All predict label using $\arg\max_{y \in \mathcal{Y}} f_y(x) = \arg\min_{y \in \mathcal{Y}} L(y, f(x))$.

Different behaviors

- For $K = 2$, all methods are equivalent and universally consistent.
- But they have different behaviors for $K > 2$.
  - Lee, Lin and Wahba's is consistent.
  - The other two are not.
- This led us to investigate consistency of a general class of methods of which all of these are special cases.

General Framework

- $L(y, \mathbf{f}(x)) = \Psi_y(\mathbf{f}(x))$, $\Psi_y : \mathbb{R}^K \mapsto \mathbb{R}_+$.
- Pointwise constraint on $\mathbf{f}$, $\forall x, \mathbf{f}(x) \in \mathcal{C}$ for some $\mathcal{C} \subseteq \mathbb{R}^K$.

$\Psi_y(\mathbf{f})$:	$\mathcal{C}$:
$\sum_{y' \neq y} \phi(f_y - f_{y'})$	$\mathbb{R}^K$
$\max_{y' \neq y} \phi(f_y - f_{y'})$	$\mathbb{R}^K$
$\sum_{y' \neq y} \phi(-f_{y'})$	$\{\mathbf{z} \in \mathbb{R}^K : \sum_{i=1}^K z_i = 0\}$

- $\phi(x) = (1 - x)_+$ gives us our three example methods but we can think of using other ϕ as well.

Ψ-risk

Fix a class $\mathcal{F} = \{\mathbf{f} : \forall x, \mathbf{f}(x) \in \mathcal{C}\}$ of vector functions.

Ψ-risk: $$R_\Psi(\mathbf{f}) = \mathbf{E}\Psi_y(\mathbf{f}(x)) \; ,$$

optimal Ψ-risk: $$R_\Psi^* = \inf_{\mathbf{f}\in\mathcal{F}} R_\Psi(\mathbf{f}) = \mathbf{E}_x \left[\inf_{\mathbf{f}(x)\in\mathcal{C}} \sum_y p_y(x)\Psi_y(\mathbf{f}(x)) \right]$$

where $p_y(x) = P(Y = y | X = x)$.

Since $\mathbf{f}$ enters into the Ψ-risk definition only through Ψ, we assume that we predict labels using

$$\mathrm{pred}(\Psi_1(\mathbf{f}(x)), \ldots, \Psi_K(\mathbf{f}(x)))$$

for some $\mathrm{pred} : \mathbb{R}^K \mapsto \mathcal{Y}$.

Consistency

Here, consistency means that for all probability distributions and all sequences $\{\mathbf{f}^{(n)}\}$,

$$R_\Psi(\mathbf{f}^{(n)}) \to R_\Psi^* \quad \Longrightarrow \quad R(\mathbf{f}^{(n)}) \to R^*.$$

$$R_\Psi^* = \mathbf{E}_x \left[\inf_{\mathbf{f}(x)\in\mathcal{C}} \sum_y p_y(x)\Psi_y(\mathbf{f}(x)) \right]$$

- To minimize the inner sum for a given x, we have to minimize:

$$\langle \mathbf{p}(x), \mathbf{z} \rangle$$

for $\mathbf{z} \in \mathcal{S}$, where $\mathcal{S} = \mathrm{conv}\{(\Psi_1(\mathbf{f}), \ldots, \Psi_K(\mathbf{f})) : \mathbf{f} \in \mathcal{C}\}$.

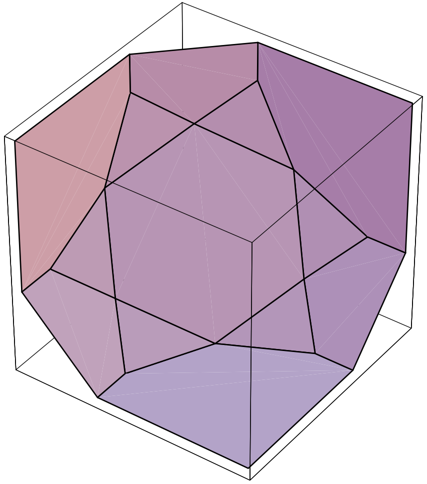
Consistency

- Consider an (informal) game where:
  - The opponent chooses a $\mathbf{p} \in \Delta_K$ and reveals to us a sequence $\mathbf{z}^{(n)}$ with $\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z} \in \mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle$
  - We output the sequence $l_n = \mathrm{pred}(\mathbf{z}^{(n)})$.

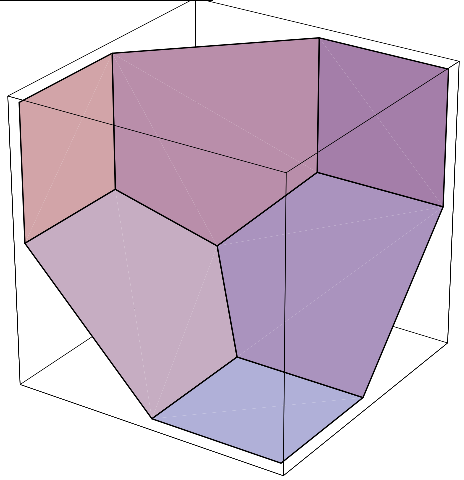
  We win if $p_{l_n} = \max_y p_y$ ultimately.

- For consistency, there should be a pred such that we win irrespective of the choice of the opponent.

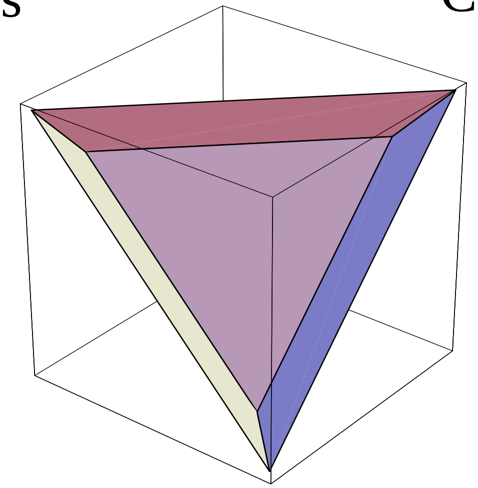
Pictures of boundary of $\mathcal{S}$



Weston & Watkins



Crammer & Singer



Lee, Lin & Wahba

Classification Calibration

Definition: $\mathcal{S} \subseteq \mathbb{R}_+^K$ is CC iff $\exists\,\mathrm{pred}$ such that $\forall \mathbf{p} \in \Delta_K$ and all $\{\mathbf{z}^{(n)}\}$ in $\mathcal{S}$,

$$\langle \mathbf{p}, \mathbf{z}^{(n)} \rangle \to \inf_{\mathbf{z}\in\mathcal{S}} \langle \mathbf{p}, \mathbf{z} \rangle \ ,$$

implies

$$p_{\mathrm{pred}(\mathbf{z}^{(n)})} = \max_y p_y$$

ultimately.

- Assume that the set $\mathcal{S}$ is convex and symmetric (symmetry means that all K classes are treated equally).
- The definition is useful because we can show that it is equivalent to:

$$\forall \{\mathbf{f}^{(n)}\} \text{ in } \mathcal{F}, \quad R_\Psi(\mathbf{f}^{(n)}) \to R_\Psi^* \quad \Rightarrow \quad R(\mathbf{f}^{(n)}) \to R^* \ .$$

Admissibility

- If any pred works then so will one satisfying $z_{\mathrm{pred}(\mathbf{z})} = \min_y z_y$, which motivates the definition below.

Definition: $\mathcal{S}$ is admissible if $\forall \mathbf{z} \in \partial\mathcal{S}$, $\forall \mathbf{p} \in \mathcal{N}(\mathbf{z})$, we have

$$\arg\min_y(z_y) \subseteq \arg\max_y(p_y) \ .$$

where $\mathcal{N}(\mathbf{z})$ is the set of non-negative normals (to $\mathcal{S}$) at $\mathbf{z}$.

- For admissibility, it seems that we have to check all points $\mathbf{z}$ on the boundary of $\mathcal{S}$, but it turns out that we can ignore many points (like those with singleton normal sets or those which have a unique minimum coordinate).

Necessary and sufficient condition

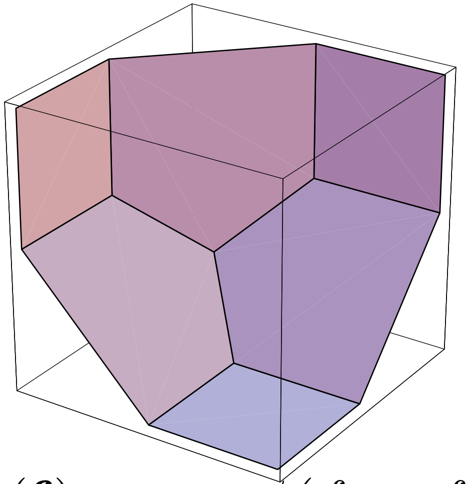
- Admissibility *weaker* than classification calibration.
- It is equivalent to the CC definition with the additional assumption of *boundedness* of the sequence $\{\mathbf{z}^{(n)}\}$.
- Necessary and sufficient condition is given by:

Theorem Let $\mathcal{S} \subseteq \mathbb{R}_+^K$ be a symmetric convex set. Define the sets

$$\mathcal{S}^{(i)} = \{(z_1, \ldots, z_i) : \mathbf{z} \in \mathcal{S}\}$$

for $i \in \{2, \ldots, K\}$. Then $\mathcal{S}$ is classification calibrated iff each $\mathcal{S}^{(i)}$ is admissible.

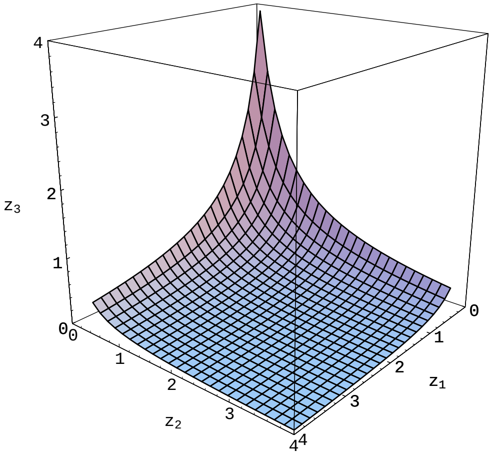
Example 1: Crammer and Singer



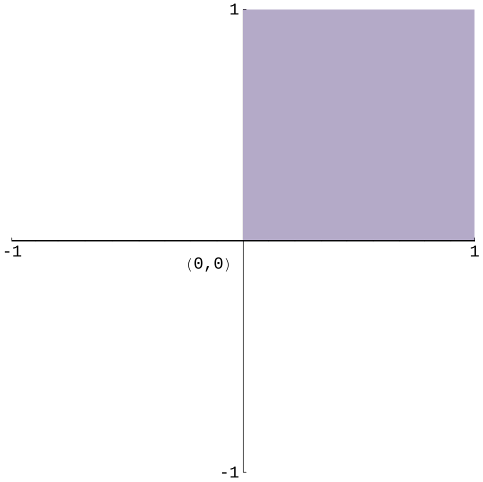
$$\Psi_y(\mathbf{f}) = \max_{y' \neq y} \phi(f_y - f_{y'})$$

- For all ϕ differentiable at 0, the set of normals at $\mathbf{z} = (\phi(0), \phi(0), \phi(0))$ includes $(0,1,1)$, $(1,0,1)$ and $(1,1,0)$. Since $\arg\min_y(z_y) = \{1,2,3\}$ and $\arg\max_y((0,1,1)) = \{2,3\}$, admissibility is violated.

Example 2: A smooth loss function



Boundary of the set $\mathcal{S} = \mathcal{S}^{(3)}$



The set $\mathcal{S}^{(2)}$

- $\Psi_y(\mathbf{f}) = \exp(-f_y)$ with $K = 3$ and $\sum_y f_y = 0$ gives $\mathcal{S} = \{\mathbf{z} \in \mathbb{R}_+ : z_1 z_2 z_3 \geq 1\}$.

- $\mathcal{S}$ is admissible, $\mathcal{S}^{(2)}$ is not (origin has $(0, 1)$ and $(1, 0)$ as normals).

Overview

- Review: Convex cost versus risk.
- Universal consistency.
- Classification with a reject option.
- Multiclass generalizations.

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