Regression Methods for Pattern Classification: Statistical Properties of Large Margin Classifiers

Peter Bartlett

Computer Science Division and Department of Statistics

UC Berkeley

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

The Pattern Classification Problem

- i.i.d. $(X,Y),(X_1,Y_1),\ldots,(X_n,Y_n)$ from $\mathcal{X}\times\mathcal{Y}$, $|\mathcal{Y}|<\infty$, for example, $\mathcal{Y}=\{\pm 1\}$.
- Use data $(X_1,Y_1),\ldots,(X_n,Y_n)$ to choose $f_n:\mathcal{X}\to\mathcal{Y}$ with small risk, $R(f_n)=\Pr\left(f_n(X)\neq Y\right)=\mathbb{E}\ell(Y,f_n(X))$.
- Natural approach: minimize empirical risk,

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

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

Large Margin Algorithms: Two Class Case

- Suppose $Y \in \{\pm 1\}$, $f_n : \mathcal{X} \to \mathbb{R}$. Define

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

- Consider the margins, $Y f_n(X)$.
- Define a margin cost function $\phi : \mathbb{R} \to \mathbb{R}^+$.
- Define the **ϕ-risk** of $f : \mathcal{X} \to \mathbb{R}$ as $R_\phi(f) = \mathbb{E}\phi(Y f(X))$.
- 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**,

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

or a regularized version.)

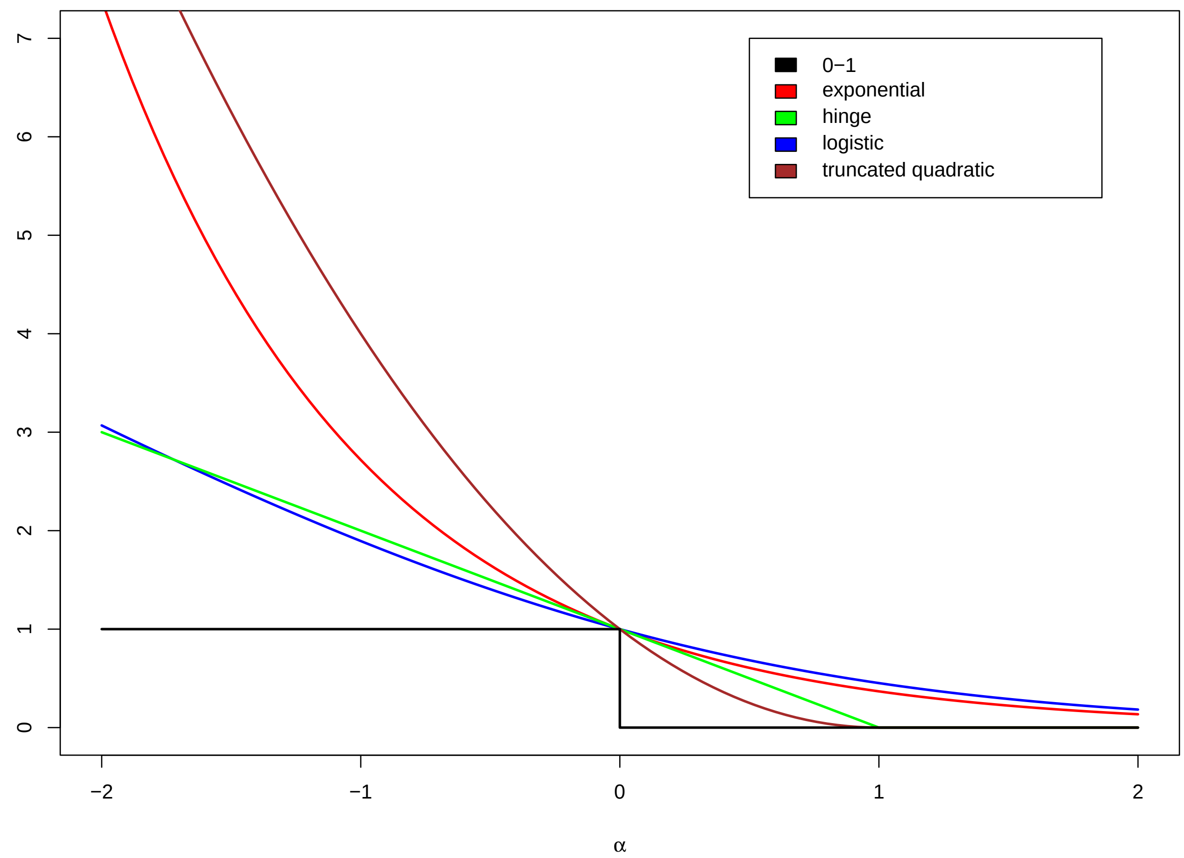
Large Margin Algorithms

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

Large Margin Algorithms

- Many other variants
  - Neural net classifiers
    $\phi(\alpha) = \max(0, (0.8 - \alpha)^2)$.
  - Support vector machines with 1-norm soft margin
    $\phi(\alpha) = \max(0, 1 - \alpha)$.
  - 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

- Universal consistency? For which ϕ?
- How is risk related to ϕ-risk?
- Model selection. Oracle inequalities.
- Does minimizing ϕ-risk correspond to estimating a model of $Y|X$?
- Similarly for multiclass.

Statistical Consequences of Using a Convex Cost

Sources:

Lin, 2004: Loss functions.

Zhang, 2004: SVMs, regularized boosting.

Lugosi and Vayatis, 2004: Regularized boosting methods.

Steinwart, 2003, 2004: Support vector machines.

Jiang, 2004: Process consistency of boosting.

Koltchinskii and Panchenko, 2000: Boosting.

Blanchard, Lugosi and Vayatis, 2003: Regularized boosting methods.

Shen, Tseng, Zhang and Wong, 2003: ψ-learning.

Bickel and Ritov, 2004: Boosting.

Buhlmann and Yu, 2002: L2 boosting.

Bartlett, Jordan, McAuliffe, 2005: Convex loss functions.

Tewari and Bartlett, 2005: Multiclass.

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition, universal consistency, and oracle inequalities.
- ϕ-risk and probability models.
- Multiclass classification: Universal consistency.

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

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

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

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

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

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:

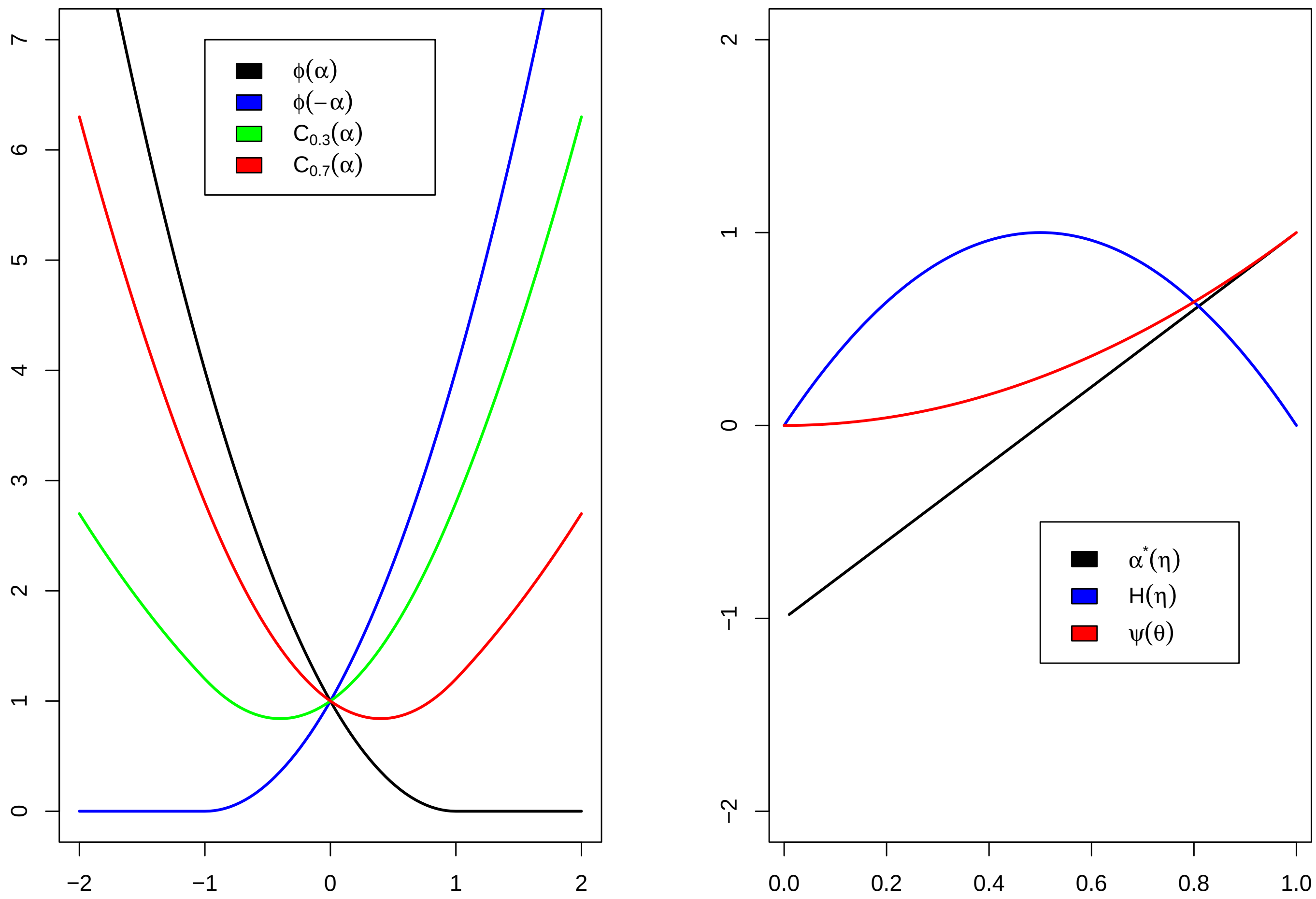
$$\mathbb{E}\left[\phi(Yf(X))|X=x\right] = \eta(x)\phi(f(x)) + (1-\eta(x))\phi(-f(x)).$$

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 $\eta \neq 1/2$,

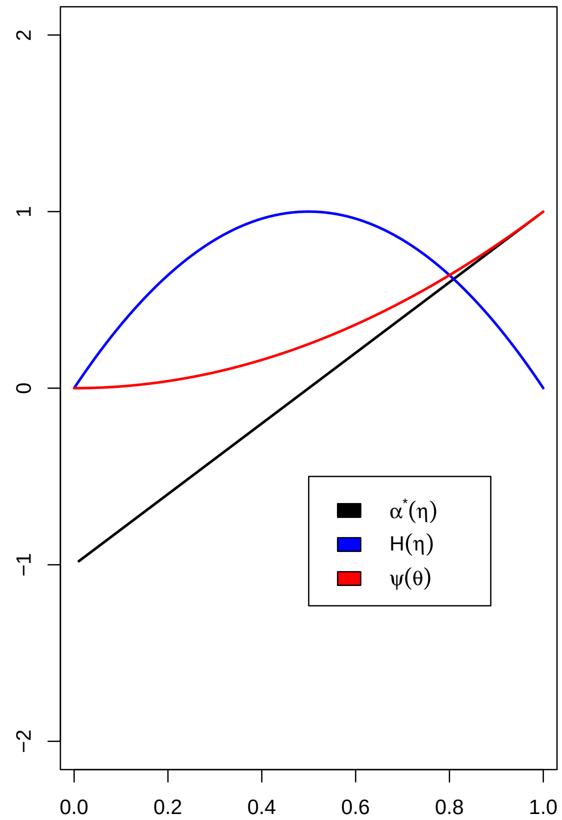
$$H^{-}(\eta) > H(\eta).$$

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

$$R(f) - R^* = \theta$$
$$\psi(\theta) \le R_\phi(f) - R_\phi^* \le \psi(\theta) + \epsilon.$$

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

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

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.
- The approximation/estimation decomposition, universal consistency, and oracle inequalities.
- ϕ-risk and probability models.
- Multiclass classification: Universal consistency.

Method of sieves/Regularized empirical risk

$$f_n = \hat{f}_{k_n} \qquad \hat{f}_k = \arg\min_{f \in \mathcal{F}_k} \hat{R}_\phi(f), \qquad \mathcal{F} = \bigcup_k \mathcal{F}_k,$$

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

Examples:

- Adaboost:
  - $\mathcal{F}_k = \mathrm{span}_k(\mathcal{G}) = \left\{ \sum_{i=1}^k \alpha_i g_i : g_i \in \mathcal{G} \right\}$, $\mathcal{G}$ is a VC-class, or
  - $\mathcal{F}_k = k\,\mathrm{co}(\mathcal{G})$, or
  - $\mathcal{F} = \mathrm{span}(\mathcal{G})$, $\Omega(f) = \sum_i |\alpha_i|$.

Method of sieves/Regularized empirical risk

$$f_n = \hat{f}_{k_n} \qquad \hat{f}_k = \arg\min_{f \in \mathcal{F}_k} \hat{R}_\phi(f), \qquad \mathcal{F} = \bigcup_k \mathcal{F}_k,$$

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

Examples:

- Support vector machines:
  - $\mathcal{F} = \mathcal{H}$, reproducing kernel Hilbert space, $\Omega(f) = \|f\|_{\mathcal{H}}$, or
  - $\mathcal{F}_k = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq k\}$.

The Approximation/Estimation Decomposition

We can decompose the excess risk estimate as

$$
\begin{aligned}
R(f_n) - R^* &\le \psi^{-1}\left(R_\phi(f_n) - R_\phi^*\right) \\
&= \psi^{-1}\Bigg(\underbrace{R_\phi(f_n) - \inf_{f \in \mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f \in \mathcal{F}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}}\Bigg).
\end{aligned}
$$

- Approximation and estimation errors are in terms of R_ϕ, not R.
- Like a regression problem.

The Approximation/Estimation Decomposition

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

$$= \psi^{-1}\left(\underbrace{R_\phi(f_n) - \inf_{f\in\mathcal{F}_n} R_\phi(f)}_{\text{estimation error}} + \underbrace{\inf_{f\in\mathcal{F}_n} R_\phi(f) - R_\phi^*}_{\text{approximation error}}\right).$$

- If the class is suitably rich (so that $\inf_{f\in\mathcal{F}} R_\phi(f) = R_\phi^*$), and the regularization is relaxed suitably slowly (e.g., $k_n \to \infty$ slowly, or $\lambda_n \to 0$ slowly),

$$R_\phi(f_n) \xrightarrow{P} R_\phi^*.$$

- Universal consistency ($R(f_n) \to R^*$) follows iff ϕ is classification calibrated.

Oracle Inequalities

For $\hat{f}_k = \arg\min_{f \in \mathcal{F}_k} \hat{R}_\phi(f)$,

$$f_n = \hat{f}_{\hat{k}} \qquad \text{with} \quad \hat{k} = \arg\min_k \left(\hat{R}_\phi(\hat{f}_k) + p_k \right),$$

for some penalty p_k (that might depend on n),
we are interested in *oracle inequalities* of the form

$$R_\phi(f_n) - R_\phi^* \leq \inf_k \left(\inf_{f \in \mathcal{F}_k} R_\phi(f) - R_\phi^* + c p_k \right).$$

This would imply

$$R(f_n) - R^* \leq \inf_k \psi^{-1} \left(\inf_{f \in \mathcal{F}_k} R_\phi(f) - R_\phi^* + c p_k \right).$$

Oracle Inequalities: Uniform Convergence Suffices

Define

empirical risk minimizer in $\mathcal{F}_k$: $\hat{f}_k = \arg\min_{f \in \mathcal{F}_k} \hat{R}_\phi(f)$,

penalized ERM in $\mathcal{F}$: $f_n = \hat{f}_{\hat{k}}$,

class with best penalized emp. risk: $\hat{k} = \arg\min_k \left(\hat{R}_\phi(\hat{f}_k) + p_k \right)$,

risk minimizer in $\mathcal{F}_k$: $f_k^* = \arg\min_{f \in \mathcal{F}_k} R_\phi(f)$,

class with best penalized risk: $k^* = \arg\min_k \left(R_\phi(f_k^*) + 2p_k \right)$.

Oracle Inequalities: Uniform Convergence Suffices

If

$$\sup_k \left(\sup_{f \in \mathcal{F}_k} \left| R_\phi(f) - \hat{R}_\phi(f) \right| - p_k \right) \leq 0, \qquad (*)$$

then

$$\begin{aligned}
R_\phi(f_n) &\leq \hat{R}_\phi(\hat{f}_{\hat{k}}) + p_{\hat{k}} && \text{(by (*) and definition of } f_n\text{)} \\
&\leq \hat{R}_\phi(\hat{f}_{k^*}) + p_{k^*} && \text{(definition of } \hat{k}\text{)} \\
&\leq \hat{R}_\phi(f^*_{k^*}) + p_{k^*} && \text{(definition of } \hat{f}_{k^*}\text{)} \\
&\leq R_\phi(f^*_{k^*}) + 2p_{k^*} && \text{(by (*) again)} \\
&= \inf_k \inf_{f \in \mathcal{F}_k} \left(R_\phi(f) + 2p_k \right).
\end{aligned}$$

So *uniform convergence* of empirical ϕ-risks to ϕ-risks suffices.

Oracle Inequalities: Ratio Inequalities Suffice

But this approach can be improved. For example, if ϕ is quadratic and $\mathcal{F}_k$ is convex, finite dimensional, and uniformly bounded, then the rate of uniform convergence over $\mathcal{F}_k$ is $\Omega(n^{-1/2})$, but with high probability

$$\underbrace{R_\phi(f) - R_\phi(f_k^*)}_{\text{excess risk}} \leq 2 \underbrace{\left(\hat{R}_\phi(f) - \hat{R}_\phi(f_k^*)\right)}_{\text{difference of empirical risks}} + O\left(\frac{1}{n}\right).$$

Since $\hat{R}_\phi(\hat{f}_k) \leq \hat{R}_\phi(f_k^*)$, this implies $\mathbb{E}\left(R_\phi(\hat{f}_k) - R_\phi(f_k^*)\right) = O(1/n)$.

The key property is the relationship

$$\mathbb{E}\left(\phi(Yf(X)) - \phi(Yf_k^*(X))\right)^2 \leq c\left(\mathbb{E}\left(\phi(Yf(X)) - \phi(Yf_k^*(X))\right)\right)^2,$$

which follows from ϕ being Lipschitz and uniformly convex.

Oracle Inequalities: Ratio Inequalities Suffice

It turns out that such inequalities suffice for oracle inequalities, provided the $\mathcal{F}_k$ are ordered by inclusion.

Theorem: Suppose $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots$ and $\bigcup_k \mathcal{F}_k = \mathcal{F}$. If

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(R_\phi(f) - R_\phi(f_k^*) - 2\left(\hat{R}_\phi(f) - \hat{R}_\phi(f_k^*) \right) - \epsilon_k \right) \leq 0,$$

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(\hat{R}_\phi(f) - \hat{R}_\phi(f_k^*) - 2\left(R_\phi(f) - R_\phi(f_k^*) \right) - \epsilon_k \right) \leq 0,$$

then with $p_k = 7\epsilon_k/2$, we have

$$R_\phi(f_n) \leq \inf_k \left(R_\phi(f_k^*) + 9\epsilon_k \right).$$

Oracle Inequalities: Ratio Inequalities Suffice

For example, for $\phi(\alpha) = \exp(-\alpha)$ and $\mathcal{F}_k = \ln(k)\,\mathrm{co}(\mathcal{G})$, with probability at least $1-\delta$, we can choose

$$\epsilon_k = c\left(\frac{k\ln k}{n^{(d+2)/(2d+2)}} + \frac{k^3\ln(k/\delta)}{n}\right),$$

where $d = \mathrm{VCdim}(\mathcal{G})$.

Choosing f_n to minimize $\hat{R}_\phi(\hat{f}_k) + c_1\epsilon_k$ gives

$$R_\phi(f_n) - R_\phi^* \le \inf_k\left(\inf_{f\in\mathcal{F}_k} R_\phi(f) - R_\phi^* + c_2\epsilon_k\right).$$

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition, universal consistency, and oracle inequalities.
- ϕ-risk and probability models.
- Multiclass classification: Universal consistency.

Estimating Conditional Probabilities

Does a large margin classifier, f_n, correspond to a model for the conditional probability $\eta(x) = \Pr(Y = 1|X = x)$?

For what ϕ?

Estimating Conditional Probabilities

If ϕ is convex, we can write

$$\begin{aligned} H(\eta) &= \inf_{\alpha \in \mathbb{R}} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \\ &= \eta\phi(\alpha^*(\eta)) + (1-\eta)\phi(-\alpha^*(\eta)), \end{aligned}$$

where $\alpha^*(\eta) = \arg\min_{\alpha} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \subset \mathbb{R} \cup \{\pm\infty\}$.

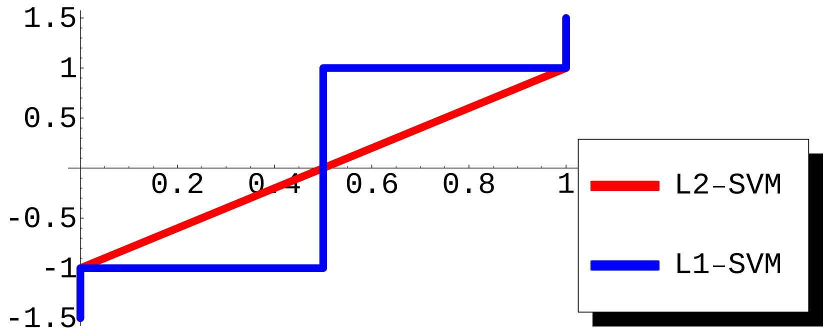
Recall:

$$\begin{aligned} R_\phi^* &= \mathbb{E}H(\eta(X)) = \mathbb{E}\phi(Y\alpha^*(\eta(X))) \\ \eta(x) &= \Pr(Y = 1 | X = x). \end{aligned}$$

Estimating Conditional Probabilities

$$\alpha^*(\eta) = \arg\min_{\alpha} \left(\eta\phi(\alpha) + (1-\eta)\phi(-\alpha)\right) \subset \mathbb{R} \cup \{\pm\infty\}.$$

Examples of $\alpha^*(\eta)$ versus $\eta \in [0,1]$:

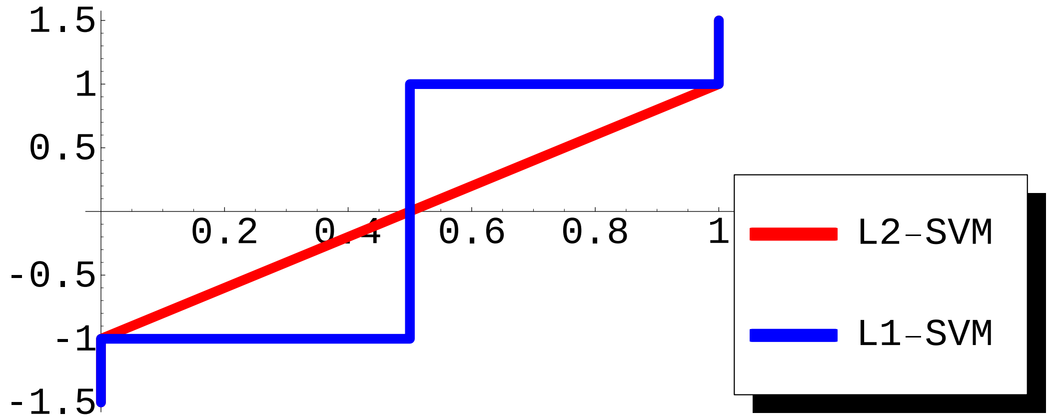


L2-SVM: $\phi(\alpha) = ((1-\alpha)_+)^2$

L1-SVM: $\phi(\alpha) = (1-\alpha)_+$,

where $(x)_+ = \max\{0, x\}$.

Estimating Conditional Probabilities



We say that α^* is invertible at η if, for all $\eta_1 \neq \eta$, $\alpha^*(\eta) \cap \alpha^*(\eta_1) = \emptyset$.

If α^* is invertible, then for any f satisfying $R_\phi(f) = R_\phi^*$, we can write η as a monotone function of f.

If α^* is not invertible, we cannot use f_n with $R_\phi(f_n) \xrightarrow{P} R_\phi^*$ to estimate η.

Estimating Conditional Probabilities

Theorem: There is a $\beta \in [0, 1/2]$ such that

1. α^* is invertible on an interval $(\beta, 1-\beta)$.

2. If $\beta > 0$, α^* is constant on $[\beta', \beta]$ and on $[1-\beta, 1-\beta']$, for some $\beta' \in [0, \beta)$.

3. $\beta \geq \gamma$.

4. Every point $\alpha \in [-\alpha_0, \alpha_0]$ of non-differentiability of ϕ corresponds to a set $[\eta_1, \eta_2] \cup [1-\eta_2, 1-\eta_1]$ where α^* is constant.

$$\text{where} \quad \gamma = \frac{\phi'_-(\alpha_0)}{\phi'_-(\alpha_0) + \phi'_+(-\alpha_0)},$$

$$\alpha_0 = \inf\{\alpha : 0 \in \partial\phi(\alpha)\},$$

$$\partial\phi(\alpha) = [\phi'_-(\alpha), \phi'_+(\alpha)] \quad \text{(subgradient of } \phi \text{ at } \alpha\text{)}.$$

Estimating Conditional Probabilities

(Zhang, 2004)

If α^* is invertible and H is differentiable (it suffices, for example, for ϕ, α^* to be differentiable), then we can view minimization of ϕ-risk as estimation of a probability model:

$$R_\phi(\alpha^*(\hat{\eta})) - R_\phi^* = \mathbb{E} d_H(\hat{\eta}(X), \eta(X)),$$

where d_H is the *Bregman divergence* with respect to H,

$$d_H(\hat{\eta}, \eta) = H(\hat{\eta}) + H'(\hat{\eta})(\eta - \hat{\eta}) - H(\eta).$$

(d_H is non-negative and zero only when its arguments are equal).

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition, universal consistency, and oracle inequalities.
- ϕ-risk and probability models.
- Multiclass classification: Universal consistency.

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

Two broad categories:

- 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 $\arg\max_{y \in \mathcal{Y}} f_y(x)$.

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: $\quad R_\Psi(\mathbf{f}) = \mathbb{E}\Psi_y(\mathbf{f}(x))$,

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

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 $\{\mathbf{f}^{(n)}\}$,

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

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

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

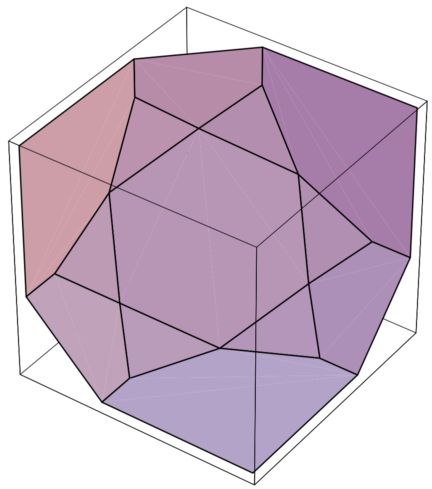
Consistency

- 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 $l_n = \mathrm{pred}(\mathbf{z}^{(n)})$.

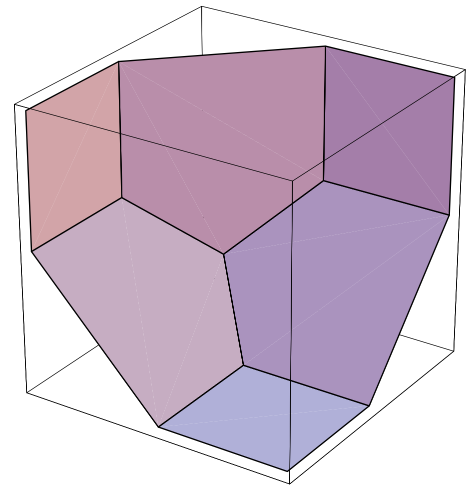
  We win if $p_{l_n} = \max_y p_y$ ultimately.

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

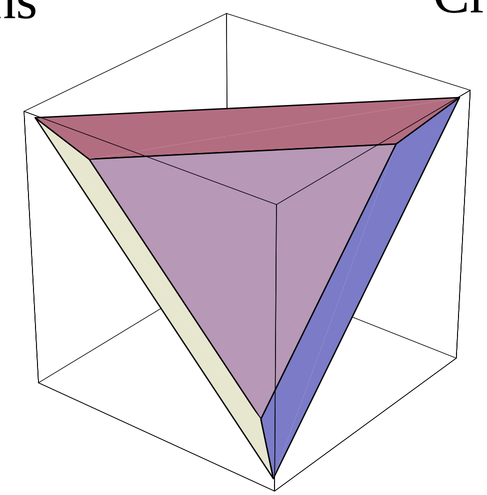
Pictures of boundary of $\mathcal{S}$



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

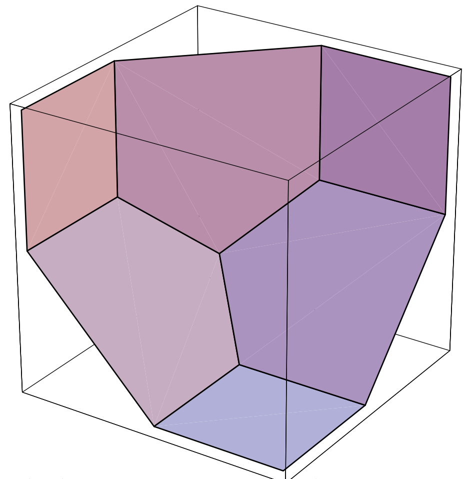
- Admissibility *weaker* than classification calibration.
- It is equivalent to the CC definition with the additional assumption of *boundedness* of the sequence $\{\mathbf{z}^{(n)}\}$.
- Necessary and sufficient condition is given by:

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.

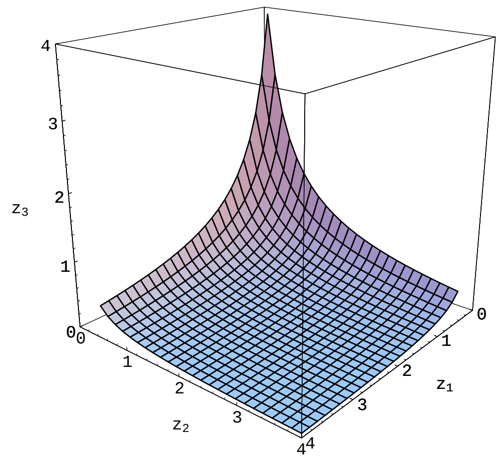
Example 1: Crammer and Singer



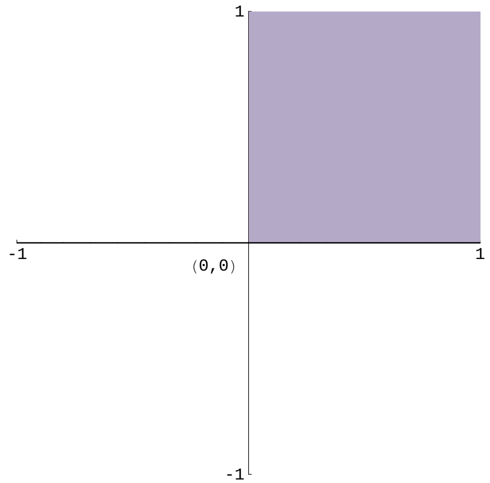
$$\Psi_y(\mathbf{f}) = \max_{y' \neq y} \phi(f_y - f_{y'})$$

- 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).
- A differentiable loss function yields an inconsistent method: something that cannot happen for binary classification.

Statistical Consequences of Using a Convex Cost

- The relationship between excess risk and excess ϕ-risk.
- The approximation/estimation decomposition, universal consistency, and oracle inequalities.
- ϕ-risk and probability models.
- Multiclass classification: Universal consistency.