Statistical Properties of Large Margin Classifiers

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

Joint work with
Mike Jordan, Jon McAuliffe, Ambuj Tewari.

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 $\mathcal{X} \times \{\pm 1\}$.
- 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.)

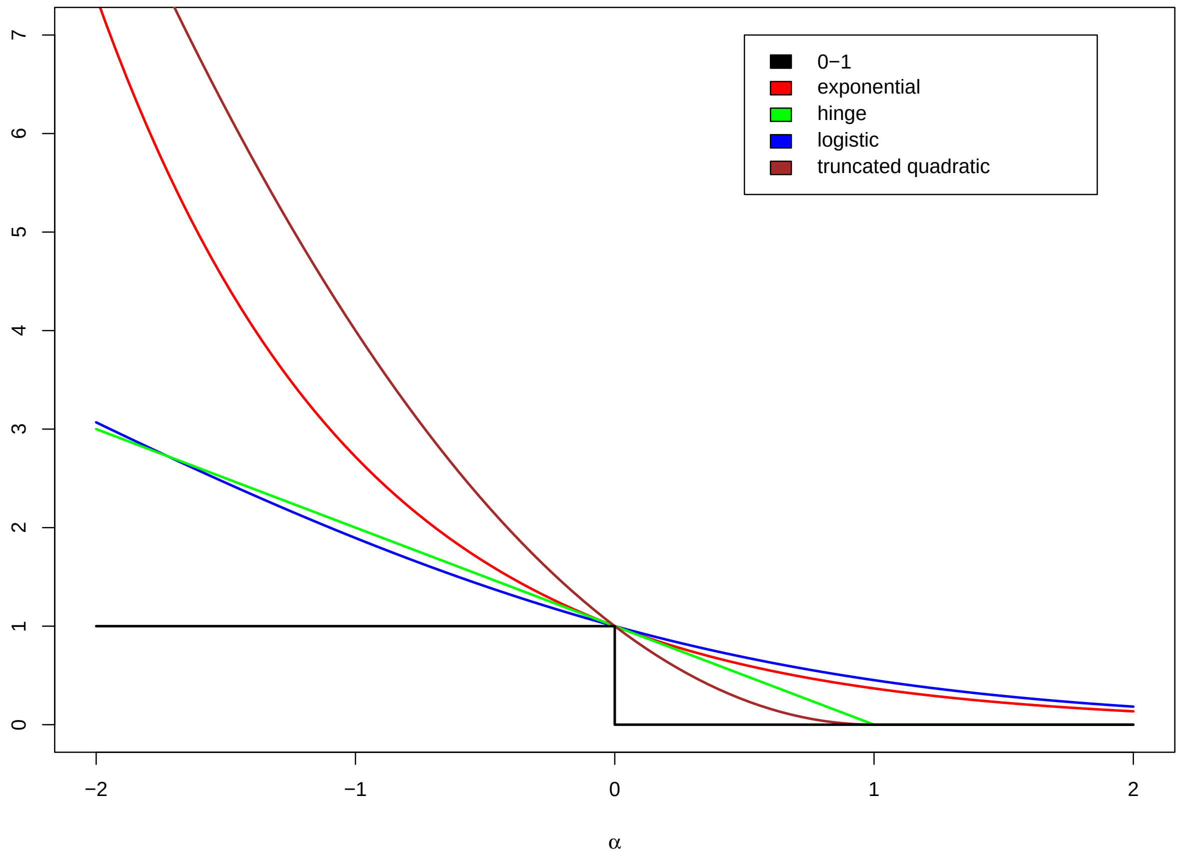
Large Margin Algorithms

- 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.
- 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

- Bayes risk consistency? For which ϕ?
  - (Lugosi and Vayatis, 2004), (Mannor, Meir and Zhang, 2002): regularized boosting.
  - (Zhang, 2004), (Steinwart, 2003): SVM.
  - (Jiang, 2004): boosting with early stopping.

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.
- Convergence rates?
- Estimating conditional probabilities?

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Kernel classifiers: sparseness versus probability estimation.

Definitions and Facts

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

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

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

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

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:

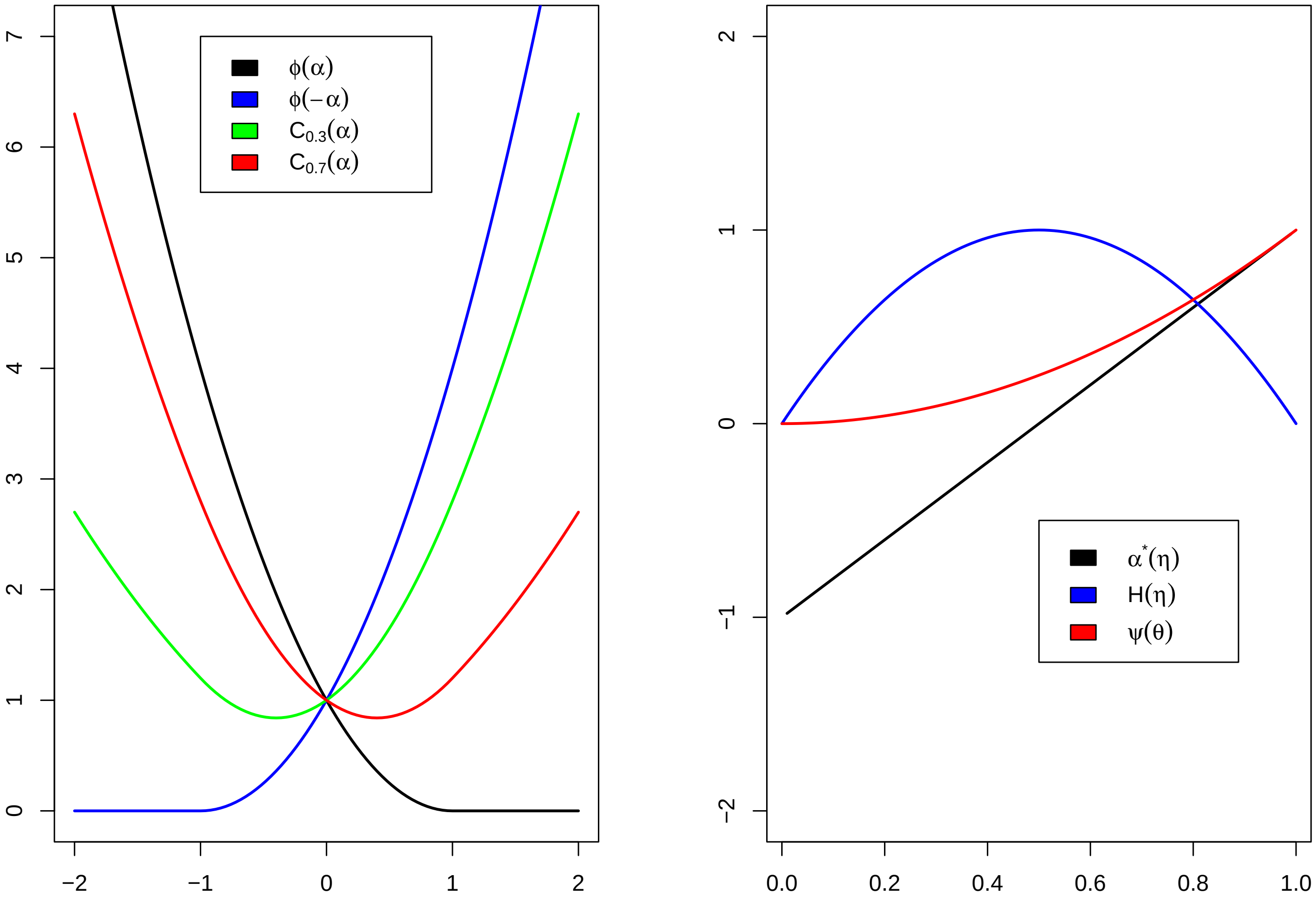
$$\mathbb{E}\left[\phi(Yf(X))|X = x\right] = \eta(x)\phi(f(x)) + (1 - \eta(x))\phi(-f(x)).$$

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))

Definitions

Definition: Given ϕ, define $\psi : [0,1] \to [0,\infty)$ by $\psi = \tilde{\psi}^{**}$, where

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

Here, g^{**} is the Fenchel-Legendre biconjugate of g,

$$\mathrm{epi}(g^{**}) = \overline{\mathrm{co}}(\mathrm{epi}(g)),$$

$$\mathrm{epi}(g) = \left\{(x,y) : x \in [0,1],\ g(x) \le y\right\}.$$

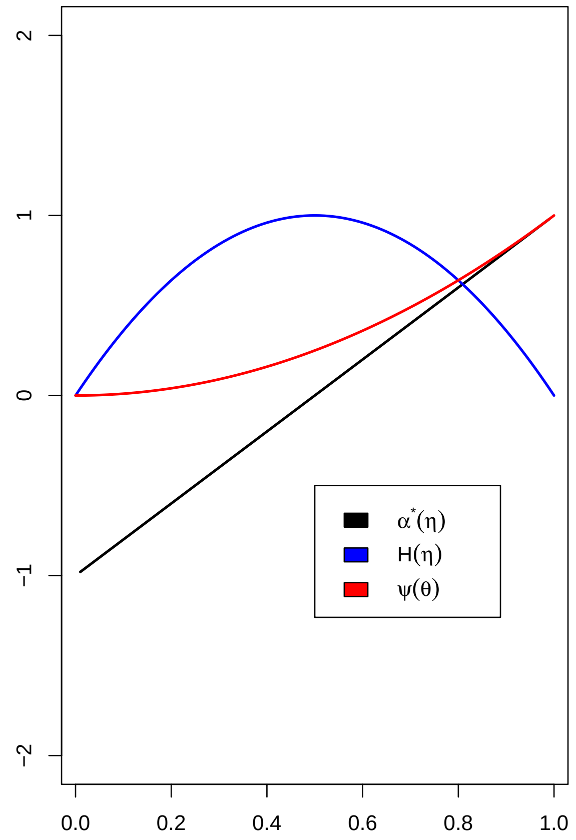
ψ-transform: Example

- ψ is the best convex lower bound on $\tilde{\psi}(\theta) = H^-((1+\theta)/2) - H((1+\theta)/2)$, the excess conditional ϕ-risk when the sign is incorrect.
- $\psi = \tilde{\psi}^{**}$ is the biconjugate of $\tilde{\psi}$,

$$\mathrm{epi}(\psi) = \overline{\mathrm{co}}(\mathrm{epi}(\tilde{\psi})),$$

$$\mathrm{epi}(\psi) = \{(\alpha, t) : \alpha \in [0,1],\ \psi(\alpha) \leq t\}.$$

- ψ is the functional convex hull of $\tilde{\psi}$.



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 bound cannot be improved.
3. Near-minimal ϕ-risk implies near-minimal risk precisely when ϕ is classification-calibrated.

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. This bound 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. Near-minimal ϕ-risk implies near-minimal risk
   precisely when ϕ is classification-calibrated.

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. This bound cannot be improved.
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 $R(f_i) \to R^*$.

Proof involves Jensen's inequality.

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 and universal consistency.
- Kernel classifiers: sparseness versus probability estimation.

The Approximation/Estimation Decomposition

Algorithm chooses

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

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

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

- 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.
- The approximation/estimation decomposition and universal consistency.
- Kernel classifiers: sparseness versus probability estimation.

Estimating Conditional Probabilities

Does a large margin classifier, f_n, allow estimates of the conditional probability $\eta(x) = \Pr(Y = 1|X = x)$, say, asymptotically?

- Confidence-rated predictions are of interest for many decision problems.
- Probabilities are useful for combining decisions.

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

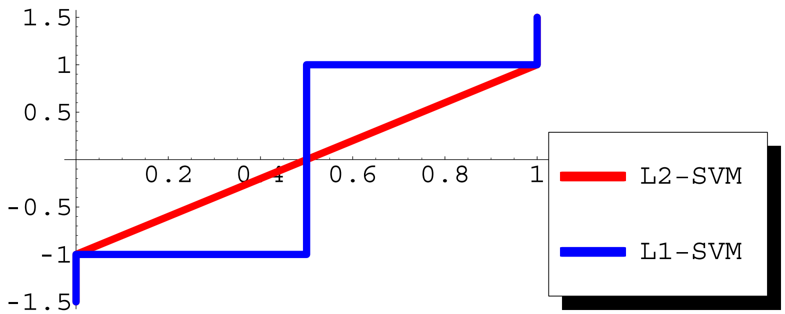
Recall:

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

Estimating Conditional Probabilities

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

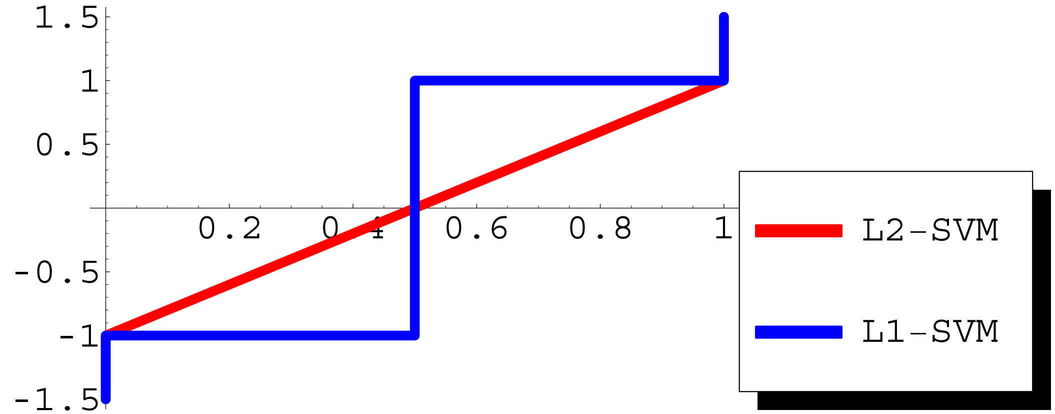
Examples of $\alpha^*(\eta)$ versus $\eta \in [0,1]$:



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

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

Estimating Conditional Probabilities



If $\alpha^*(\eta)$ is not invertible, that is, there are $\eta_1 \neq \eta_2$ with

$$\alpha^*(\eta_1) \cap \alpha^*(\eta_2) \neq \emptyset,$$

then there are distributions P and functions f_n with $R_\phi(f_n) \to R_\phi^*$ but $f_n(x)$ cannot be used to estimate $\eta(x)$.
e.g., $f_n(x) \to \alpha^*(\eta_1) \cap \alpha^*(\eta_2)$. Is $\eta(x) = \eta_1$ or $\eta(x) = \eta_2$?

Kernel classifiers and sparseness

- Kernel classification methods:

$$f_n = \arg\min_{f \in \mathcal{H}} \left(\hat{E}\phi(Yf(X)) + \lambda_n \|f\|^2 \right),$$

  where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS), with norm $\|\cdot\|$, and $\lambda_n > 0$ is a regularization parameter.

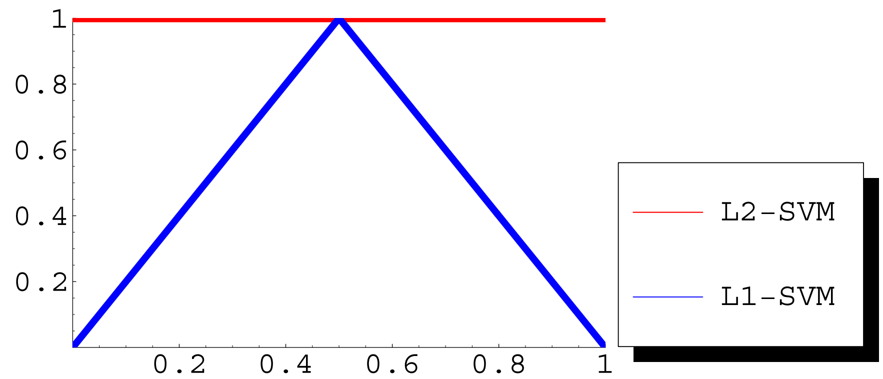
- Representer theorem: solution of optimization problem can be represented as:

$$f_n(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i) \, .$$

- Data x_i with $\alpha_i \neq 0$ are called *support vectors* (SV's).
- Sparseness (number of support vectors $\ll n$) means faster evaluation of the classifier.

Sparseness: Steinwart's results

- For L1 and L2-SVM, Steinwart proved that the asymptotic fraction of SV's is $\mathbb{E}G(\eta(X))$ (under some technical assumptions).
- The function $G(\eta)$ depends on the loss function used:



- L2-SVM doesn't produce sparse solutions (asymptotically) while L1-SVM does.
- Recall: L2-SVM can estimate η while L1-SVM cannot.

Sparseness versus Estimating Conditional Probabilities

The ability to estimate conditional probabilities always causes loss of sparseness:

- Lower bound of the asymptotic fraction of data that become SV's can be written as $\mathbb{E}G(\eta(X))$.
- $G(\eta)$ is 1 throughout the region where probabilities can be estimated.
- The region where $G(\eta) = 1$ is an interval centered at $1/2$.

Asymptotically Sharp Result

For loss functions of the form:

$$\phi(t) = h((t_0 - t)_+)$$

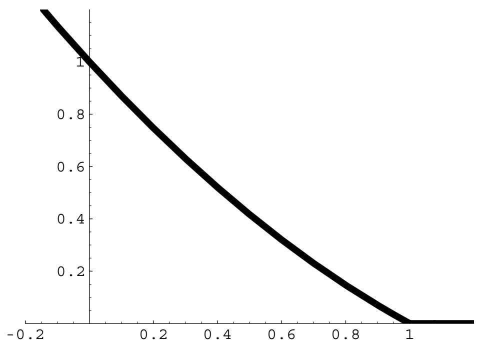
where h is convex, differentiable and $h'(0) > 0$, if the kernel k is *analytic* and *universal* (and the underlying P_X is continuous and non-trivial), then for a regularization sequence $\lambda_n \to 0$ sufficiently slowly:

$$\frac{|\{i : \alpha_i \neq 0\}|}{n} \xrightarrow{P} \mathbb{E}G(\eta(X))$$

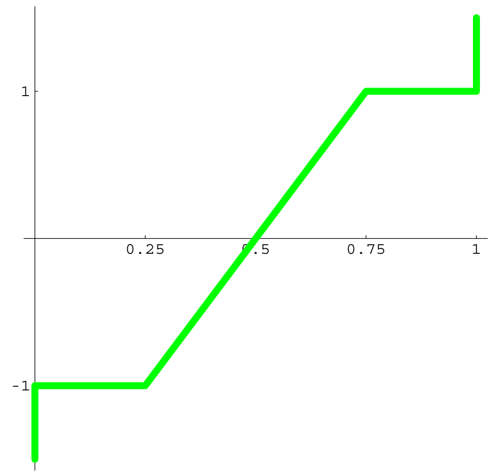
where

$$G(\eta) = \begin{cases} \eta/\gamma & 0 \leq \eta \leq \gamma \\ 1 & \gamma < \eta < 1 - \gamma \\ (1-\eta)/\gamma & 1 - \gamma \leq \eta \leq 1 \end{cases}$$

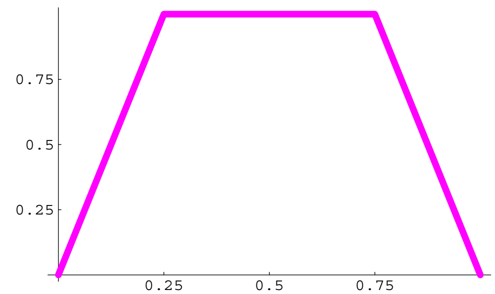
Example



$\frac{1}{3}((1-t)_+)^2 + \frac{2}{3}(1-t)_+$



$\alpha^*(\eta)$ vs. η



$G(\eta)$ vs. η

Overview

- Relating excess risk to excess ϕ-risk.
- The approximation/estimation decomposition and universal consistency.
- Kernel classifiers
  - No sparseness where $\alpha^*(\eta)$ is invertible.
  - Can design ϕ to trade off sparseness and probability estimation.

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