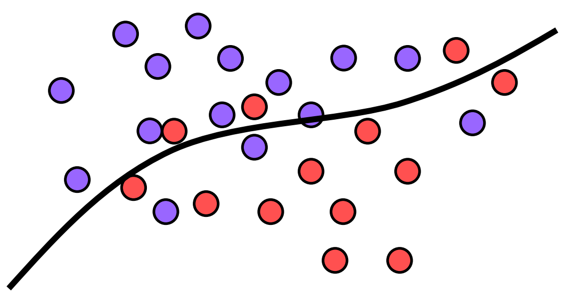
Surrogate Losses for Cost-Sensitive Classification

Clayton Scott

EECS and Statistics
University of Michigan

Cost-Sensitive Classification

- Random triple (X, Y, C)
  - X = pattern, feature vector
  - $Y \in \{-1, 1\}$, class label
  - $C \geq 0$, misclassification cost



- Example-dependent costs:
  - C may depend on X and Y

Example

- Mailing to request charitable contributions
  - γ = cost of mailing a donation request
  - X = feature vector associated to potential recipient
  - Z = amount donated

$$Y = \text{sign}(Z - \gamma)$$

$$C = |Z - \gamma|$$

- Equivalent variables: Given, Y and C
  - Fix any $\gamma \in \mathbb{R}$
  - Set

$$Z = \begin{cases} \gamma + C, & \text{if } Y = +1 \\ \gamma - C, & \text{if } Y = -1 \end{cases}$$

Formal Setting

- Given

$$(X, Z) \in \mathcal{X} \times \mathbb{R}$$

$$(X, Z) \sim P$$

$$\gamma \in \mathbb{R}$$

- Decision function

$$f : \mathcal{X} \mapsto \mathbb{R}$$

- Target risk

$$R(f) = \mathbb{E}_{(X,Z)\sim P}\left[\underbrace{|Z - \gamma| \mathbf{1}_{\{f(X)(Z-\gamma)\leq 0\}}}_{\text{loss}}\right]$$

Surrogate Losses

- Surrogate loss

$$L_\phi(z,t) = |z-\gamma|\phi(\mathrm{sign}(z-\gamma)t)$$

- Surrogate risk

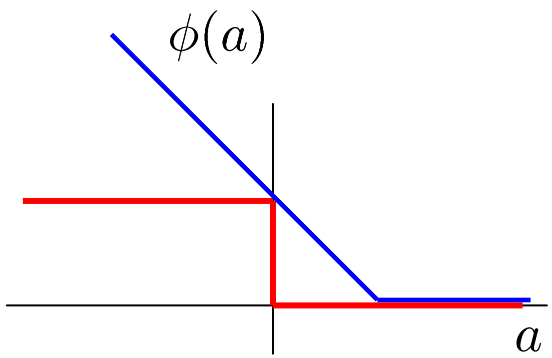
$$R_\phi(f) = \mathbb{E}_{(X,Z)\sim P}\left[L_\phi(Z, f(X))\right]$$

- Example: SVM for example-dependent costs

$$\phi(a) = \max(0, 1-a)$$

$$f(x) = \langle w, x\rangle$$

data $(x_1, z_1), \ldots, (x_n, z_n)$



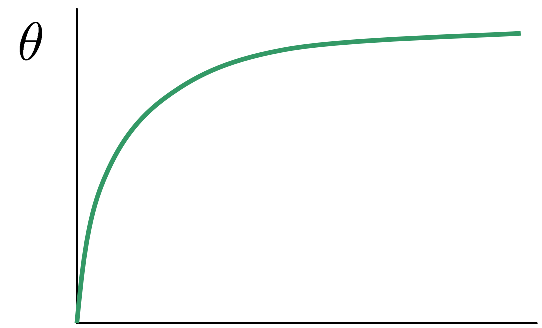
$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} L_\phi(z_i, \langle w, x_i\rangle)$$

(Zadrozny et al., Brefeld et al., 2003)

Surrogate Regret (Excess Risk) Bounds

For all $f : \mathcal{X} \mapsto \mathbb{R}$,

$$R(f) - R^* \leq \theta(R_\phi(f) - R^*_\phi)$$



Questions:

- When do such bounds exist?
- What assumptions are needed on the distribution, if any?

Previous work: Zhang (2004a,b), Bartlett et al. (2006), Steinwart (2007), Tewari and Bartlett (2007), Reid and Williamson (2011), ...

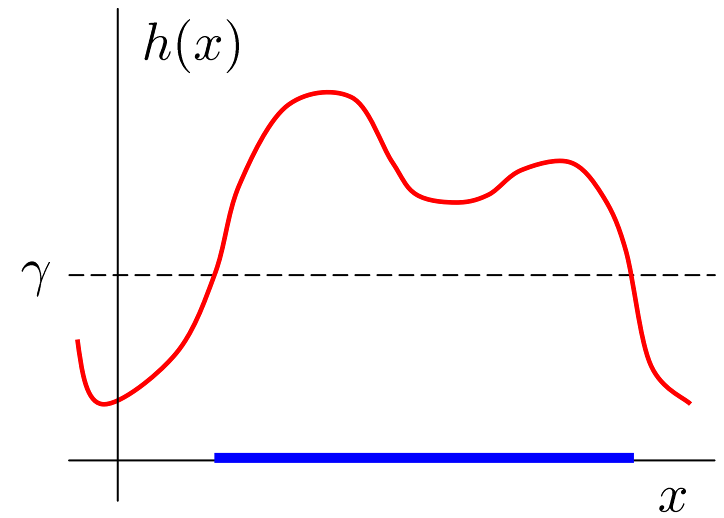
Target Regret (based on 0/1 loss)

Lemma: For all $f : \mathcal{X} \mapsto \mathbb{R}$,

$$R(f) - R^* = \mathbb{E}_X \left[\mathbf{1}_{\{\mathrm{sign}(f(X)) \neq \mathrm{sign}(h(X)-\gamma)\}} |h(X) - \gamma| \right]$$

where

$$h(x) := \mathbb{E}[Z \mid X = x]$$



- Optimal rule: $f(x) = h(x) - \gamma$
- Special case: If $Z \in \{0, 1\}$ and $\gamma = \frac{1}{2}$, then

$$h(x) = \eta(x) := P(Z = 1 | X = x)$$

$\Longrightarrow$ cost-**insensitive** classification

Basic Bound

$$
\begin{aligned}
R_\phi(f) &= \mathbb{E}_X \mathbb{E}_{Z|X}\left[(Z-\gamma)\mathbf{1}_{\{Z\geq\gamma\}}\phi(f(X)) + (\gamma - Z)\mathbf{1}_{\{Z<\gamma\}}\phi(-f(X))\right] \\
&= \mathbb{E}_X\left[C_\phi(X, f(X))\right]
\end{aligned}
$$

where

$$
\begin{aligned}
C_\phi(x,t) &:= \eta_1(x)\phi(t) + \eta_{-1}(x)\phi(-t) \\
\eta_1(x) &:= \mathbb{E}_{Z|X=x}[(Z-\gamma)\mathbf{1}_{\{Z\geq\gamma\}}] \\
\eta_{-1}(x) &:= \mathbb{E}_{Z|X=x}[(\gamma-Z)\mathbf{1}_{\{Z<\gamma\}}]
\end{aligned}
$$

conditional risk

$$H_\phi(x) = \inf_{t:t(h(x)-\gamma)\leq 0} C_\phi(x,t) - \inf_t C_\phi(x,t)$$

$$\psi_\phi(\epsilon) = \inf_{x:|h(x)-\gamma|\geq\epsilon} H_\phi(x)$$

Theorem: $\psi_\phi^{**}(R(f) - R^*) \leq R_\phi(f) - R_\phi^*$

Extension of
Bartlett et al. (2006)

Inverting the Basic Bound: Hinge Loss

Theorem: If $\phi(a) = \max(0, 1 - a)$, then

$$\psi_\phi^{**}(\epsilon) = \epsilon$$

and

$$R(f) - R^* \leq R_\phi(f) - R_\phi^*$$

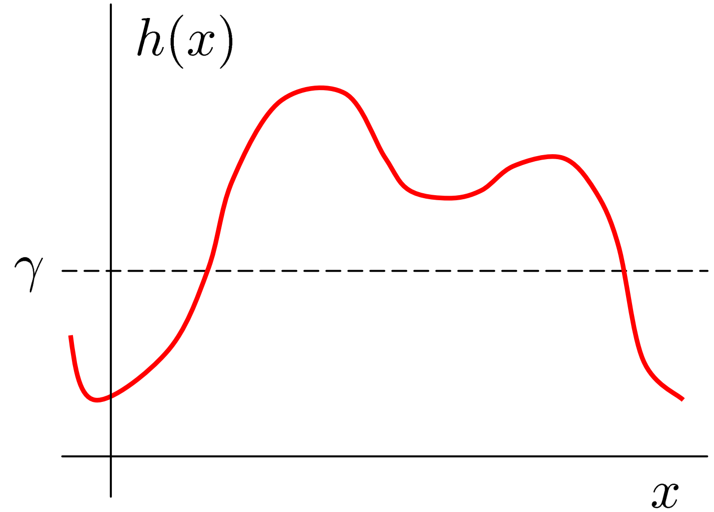
for all f and all distributions of (X, Z)

Inverting the Basic Bound: Other Losses

For other losses, need to control spread of $Z \mid X = x$

(A) $Z \mid X = x$ has variance $\sigma_x^2 \leq \sigma^2 < \infty$

(B) $Z \mid X = x$ is subGaussian



Theorem: For exponential, logistic, and squared error losses,

- Under **(A)**, $\psi_\phi^{**}(\epsilon) \geq c'\epsilon^3$ for ϵ small
- Under **(B)**, $\psi_\phi^{**}(\epsilon) \geq c'\epsilon^2$ for ϵ small

Details

Lemma:

$$H_\phi^{\text{cs}}(x) = (\eta_1(x) + \eta_{-1}(x)) H_\phi^{\text{ci}}\left(\frac{\eta_1(x)}{\eta_1(x) + \eta_{-1}(x)}\right)$$

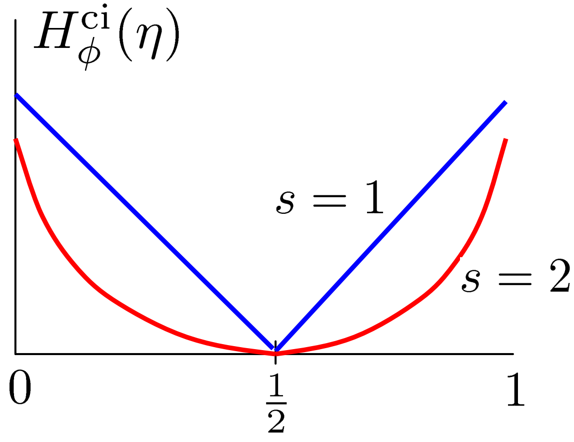
cost-**sensitive** cost-**insensitive**

where

$$\eta_1(x) := \mathbb{E}_{Z|X=x}[(Z-\gamma)\mathbf{1}_{\{Z \geq \gamma\}}]$$
$$\eta_{-1}(x) := \mathbb{E}_{Z|X=x}[(\gamma - Z)\mathbf{1}_{\{Z < \gamma\}}]$$

Condition on ϕ (Zhang, 2004): For some $s \geq 1, c > 0$

$$\forall \eta \in [0,1], \quad \left|\eta - \frac{1}{2}\right|^s \leq c^s H_\phi^{\text{ci}}(\eta)$$



Examples

- $(s = 1)$ hinge
- $(s = 2)$ exponential, logistic, squared error

Conclusions

- Hinge loss:

$$R(f) - R^* \leq R_\phi(f) - R_\phi^*$$

- Exponential, logistic, squared error ..., depends on noise

$$\begin{aligned} \text{Bounded variance:} \quad R(f) - R^* &\leq C(R_\phi(f) - R_\phi^*)^{1/3} \\ \text{subGaussian:} \quad R(f) - R^* &\leq C(R_\phi(f) - R_\phi^*)^{1/2} \end{aligned}$$

- Extension to asymmetric (non-margin) losses
- Faster rates under Polonik's low noise assumption
- Suggests new algorithms, e.g., boosting, based on different losses
- Future work: Consistency, rates