Surrogate Losses and Regret Bounds for Cost-Sensitive Classification with Example-Dependent Costs

Clayton Scott CSCOTT@EECS.UMICH.EDU
Dept. of Electrical Engineering and Computer Science, and of Statistics
University of Michigan, 1301 Beal Ave., Ann Arbor, MI 48105 USA

Abstract

We study surrogate losses in the context of cost-sensitive classification with example-dependent costs, a problem also known as regression level set estimation. We give sufficient conditions on the surrogate loss for the existence of a surrogate regret bound. Such bounds imply that as the surrogate risk tends to its optimal value, so too does the expected misclassification cost. Our sufficient conditions encompass example-dependent versions of the hinge, exponential, and other common losses. These results provide theoretical justification for some previously proposed surrogate-based algorithms, and suggests others that have not yet been developed.

1. Introduction

In traditional binary classification, there is a jointly distributed pair $(X, Y) \in \mathcal{X} \times \{-1, 1\}$, where X is a pattern and Y the corresponding class label. Training data $(x_i, y_i)_{i=1}^n$ are given, and the problem is to design a classifier $x \mapsto \text{sign}(f(x))$, where $f : \mathcal{X} \to \mathbb{R}$ is called a decision function. In *cost-insensitive* (CI) classification, the goal is to find f such that $E_{X,Y}[1_{\{\text{sign}(f(X)) \neq Y\}}]$ is minimized.

We study a generalization of the above called *cost-sensitive* (CS) classification with *example-dependent* (ED) costs (Zadrozny & Elkan, 2001; Zadrozny et al., 2003). There is now a random pair $(X, Z) \in \mathcal{X} \times \mathbb{R}$, and a threshold $\gamma \in \mathbb{R}$. Training data $(x_i, z_i)_{i=1}^n$ are given, and the problem is to correctly predict the sign of $Z - \gamma$ from X, with errors incurring a cost of $|Z - \gamma|$. The performance of the decision function $f : \mathcal{X} \to \mathbb{R}$ is assessed by the risk $R_\gamma(f) := E_{X,Z}[|Z - \gamma| 1_{\{\text{sign}(f(X)) \neq \text{sign}(Z-\gamma)\}}]$. This formulation of CS classification with ED costs is equivalent to, or specializes to, other formulations that have appeared in the literature. These connections are discussed in the next section.

As an exemplary application, consider the problem posed for the 1998 KDD Cup. The dataset is a collection of (x_i, z_i) where i indexes people who may have donated to a particular charity, x_i is a feature vector associated to that person, and z_i is the amount donated by that person (possibly zero). The cost of mailing a donation request is $\gamma = \$0.32$, and the goal is to predict who should receive a mailing, so that overall costs are minimized. (The related problem of maximizing profit is discussed below.)

Since the loss $(z, f(x)) \mapsto |z - \gamma| 1_{\{\text{sign}(f(x)) \neq \text{sign}(z-\gamma)\}}$ is neither convex nor differentiable in its second argument, it is natural to explore the use of surrogate losses. For example the support vector machine (SVM), extended to ED costs, has been considered by Zadrozny et al. (2003); Brefeld et al. (2003). In the linear case where $f(x) = w^T x$, this SVM minimizes

$$\frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^n L_\gamma(z_i, w^T x_i)$$

with respect to w, where $L_\gamma(z, t) = |z - \gamma| \max(0, 1 - \text{sign}(z - \gamma)t)$ is a generalization of the hinge loss, and $\lambda > 0$ is a regularization parameter.

Our contribution is to establish surrogate regret bounds for a class of surrogate losses that include the generalized hinge loss just described, as well as analogous generalizations of the exponential, logistic, and other common losses. Given a surrogate loss $L_\gamma : \mathbb{R} \times \mathbb{R} \mapsto [0, \infty)$, define $R_{L_\gamma}(f) = E_{X,Z}[L_\gamma(Z, f(X))]$. Define R_γ^* and $R_{L_\gamma}^*$ to be the respective infima of $R_\gamma(f)$ and $R_{L_\gamma}(f)$ over all decision functions f. We refer to $R_\gamma(f) - R_\gamma^*$ and $R_{L_\gamma}(f) - R_{L_\gamma}^*$ as the target and surrogate *regret* (or excess risk), respectively. A surrogate

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

regret bound is a function θ with $\theta(0) = 0$ that is strictly increasing, continuous, and satisfies

$$R_\gamma(f) - R^*_\gamma \leq \theta(R_{L_\gamma}(f) - R^*_{L_\gamma})$$

for all f and all distributions of (X, Z). Such bounds imply that consistency of an algorithm with respect to the surrogate risk implies consistency with respect to the target risk. These kinds of bounds are not only natural requirements of the surrogate loss, but have also emerged in recent years as critical tools when proving consistency of algorithms based on surrogate losses (Mannor et al., 2003; Blanchard et al., 2003; Zhang, 2004a;b; Lugosi & Vayatis, 2004; Steinwart, 2005; Tewari & Bartlett, 2007). Surrogate regret bounds for cost-sensitive classification with *label dependent* costs have also recently been developed by Scott (2010) and Reid & Williamson (2011).

Surrogate regret bounds were established for CI classification by Zhang (2004a) and Bartlett et al. (2006), and for other learning problems by Steinwart (2007). Our work builds on ideas from these three papers. The primary technical contributions are Theorems 2 and 3, and associated lemmas. In particular, Theorem 3 addresses issues that arise from the fact that, unlike in CI classification, with ED costs the first argument of the loss is potentially unbounded.

Another important aspect of our analysis is our decision to represent the data in terms of the pair (X, Z). Perhaps a more common representation for CS classification with ED costs is in terms of a random triple $(X, Y, C) \in \mathcal{X} \times \{-1, 1\} \times [0, \infty)$. This is equivalent to the (X, Z) representation. Given Z and γ, we may take $Y = \text{sign}(Z - \gamma)$ and $C = |Z - \gamma|$. Conversely, given Y and C, let $\gamma \in \mathbb{R}$ be arbitrary, and set

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

We have found that the (X, Z) representation leads to some natural generalizations of previous work on CI classification. In particular, the regression of Z on X turns out to be a key quantity, and generalizes the regression of the label Y on X from CI classification (i.e., the posterior label probability).

The next section relates our problem to some other supervised learning problems. Main results, concluding remarks, and some proofs are presented in Sections 3, 4, and 5, respectively. Remaining proofs may be found in Scott (2011).

2. Related Problems

Regression level set estimation. We show in Lemma 1 that for any f,

$$R_\gamma(f) - R^*_\gamma = E_X[1_{\{\text{sign}(f(X)) \neq \text{sign}(h(X) - \gamma)\}} |h(X) - \gamma|]$$

where $h(x) := E[Z \,|\, X = x]$ is the regression of Z on X. From this it is obvious that $f(x) = h(x) - \gamma$ is an optimal decision function. Therefore the optimal classifier predicts 1 on the level set $\{x : h(x) > \gamma\}$. Thus, the problem has been referred to as regression level set estimation (Cavalier, 1997; Polonik & Wang, 2005; Willett & Nowak, 2007; Scott & Davenport, 2007).

Deterministic and label-dependent costs. Our framework is quite general in the sense that given X and $Y = \text{sign}(Z - \gamma)$, the cost $C = |Z - \gamma|$ is potentially random. The special case of deterministic costs has also received attention. As yet a further specialization of the case of deterministic costs, a large body of literature has addressed the case where cost is a deterministic function of the label only (Elkan, 2001). In our notation, a typical setup has $Z \in \{0, 1\}$ and $\gamma \in (0, 1)$. Then false positives cost $1 - \gamma$ and false negatives cost γ. An optimal decision function is $h(x) - \gamma = P(Z = 1 \,|\, X = x) - \gamma$. Taking $\gamma = \frac{1}{2}$ recovers the CI classification problem.

Maximizing profit. Another framework is to not only penalize incorrect decisions, but also reward correct ones. The risk here is

$$\begin{aligned} \tilde{R_\gamma}(f) &= E_{X,Z}[|Z - \gamma| 1_{\{\text{sign}(f(X)) \neq \text{sign}(Z-\gamma)\}} \\ &\quad - |Z - \gamma| 1_{\{\text{sign}(f(X)) = \text{sign}(Z-\gamma)\}}]. \end{aligned}$$

Using $\tilde{R_\gamma}(f) = R_\gamma(f) - R_\gamma(-f)$, it can easily be shown that $\tilde{R_\gamma}(f) = 2R_\gamma(f) - E_{X,Z}[|Z - \gamma|]$, and hence $\tilde{R_\gamma}(f) - \tilde{R_\gamma}^* = 2(R_\gamma(f) - R^*_\gamma)$. Therefore, the inclusion of rewards presents no additional difficulties from the perspective of risk analysis.

3. Surrogate Regret Bounds

We first introduce the notions of label-dependent and example-dependent losses, and then outline the nature of our results before proceeding to technical details.

A measurable function $L : \{-1, 1\} \times \mathbb{R} \to [0, \infty)$ will be referred to as a *label-dependent* (LD) loss. Such losses are employed in CI classification and in CS classification with LD costs. Any LD loss can be written

$$L(y, t) = 1_{\{y=1\}} L_1(t) + 1_{\{y=-1\}} L_{-1}(t),$$

for some L_1 and L_{-1}.

A measurable function $M : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ will be referred to as an *example-dependent* (ED) loss. These are the natural losses for cost-sensitive classification with example-dependent costs. For every LD loss and $\gamma \in \mathbb{R}$, there is a corresponding ED loss. In particular, given an LD loss L and $\gamma \in \mathbb{R}$, let $L_\gamma : \mathbb{R} \times \mathbb{R} \to [0, \infty)$ be the ED loss

$$L_\gamma(z, t) := (z - \gamma)1_{\{z>\gamma\}} L_1(t) + (\gamma - z)1_{\{z\le\gamma\}} L_{-1}(t).$$

For example, if $L(y, t) = \max(0, 1 - yt)$ is the hinge loss, then

$$\begin{aligned} L_\gamma(z, t) &= (z - \gamma)1_{\{z>\gamma\}} \max(0, 1 - t) \\ &\quad +(\gamma - z)1_{\{z\le g\}} \max(0, 1 + t) \\ &= |z - \gamma| \max(0, 1 - \operatorname{sign}(z - \gamma)t). \end{aligned}$$

The goal of this research is to provide sufficient conditions on L, and possibly also on the distribution of (X, Z), for the existence of a surrogate regret bound.

Before introducing the bounds, we first need some additional notation. Let $\mathcal{X}$ be a measurable space. A decision function is any measurable $f : \mathcal{X} \to \mathbb{R}$. We adopt the convention $\operatorname{sign}(0) = -1$, although this choice is not important.

Given a LD loss and a random pair $(X, Y) \in \mathcal{X} \times \{-1, 1\}$, define the risk $R_L(f) = E_{X,Y}[L(Y, f(X))]$, and let R^*_L be the infimum of $R_L(f)$ over all decision functions f. For $\eta \in [0, 1]$ and $t \in \mathbb{R}$, the conditional risk is defined to be

$$C_L(\eta, t) := \eta L_1(t) + (1 - \eta)L_{-1}(t),$$

and for $\eta \in [0, 1]$, the optimal conditional risk is

$$C^*_L(\eta) := \inf_{t\in\mathbb{R}} C_L(\eta, t).$$

If $\eta(x) := P(Y = 1|X = x)$, and f is a decision function, then $R_L(f) = E_X[C_L(\eta(X), f(X))]$ and $R^*_L = E_X[C^*_L(\eta(X))]$.

Now define $H_L(\eta) = C^-_L(\eta) - C^*_L(\eta)$, for $\eta \in [0, 1]$, where

$$C^-_L(\eta) := \inf_{t\in\mathbb{R}:t(2\eta-1)\le 0} C_L(\eta, t).$$

Notice that by definition, $H_L(\eta) \ge 0$ for all η, with equality when $\eta = \frac{1}{2}$. Bartlett et al. showed that surrogate regret bounds exist for CI classification, in the case of margin losses where $L(y, t) = \phi(yt)$, iff $H_L(\eta) > 0 \; \forall\eta \ne \frac{1}{2}$.

We require extensions of the above definitions to the case of ED costs.

If $(X, Z) \in \mathcal{X} \times \mathbb{R}$ are jointly distributed and f is a decision function, the L_γ-risk of f is $R_{L_\gamma}(f) := E_{X,Z}[L_\gamma(Z, f(X))]$, and the optimal L_γ-risk is $R^*_{L_\gamma} = \inf_f R_{L_\gamma}(f)$.

In analogy to the label-dependent case, for $x \in \mathcal{X}$ and $t \in \mathbb{R}$, define

$$C_{L,\gamma}(x, t) := h_{1,\gamma}(x)L_1(t) + h_{-1,\gamma}(x)L_{-1}(t),$$

where

$$h_{1,\gamma}(x) := E_{Z|X=x}[(Z - \gamma)1_{\{Z>\gamma\}}]$$

and

$$h_{-1,\gamma}(x) := E_{Z|X=x}[(\gamma - Z)1_{\{Z\le\gamma\}}].$$

In addition, define

$$C^*_{L,\gamma}(x) = \inf_{t\in\mathbb{R}} C_{L,\gamma}(x, t).$$

With these definitions, it follows that $R_{L_\gamma}(f) = E_X[C_{L,\gamma}(X, f(X))]$ and $R^*_{L_\gamma} = E_X[C^*_{L,\gamma}(X)]$. Finally, for $x \in \mathcal{X}$, set

$$H_{L,\gamma}(x) := C^-_{L,\gamma}(x) - C^*_{L,\gamma}(x)$$

where

$$C^-_{L,\gamma}(x) := \inf_{t\in\mathbb{R}:t(h(x)-\gamma)\le 0} C_{L,\gamma}(x, t).$$

A connection between $H_{L,\gamma}$ and H_L is given in Lemma 3.

Note that if $L(y, t) = 1_{\{y\ne\operatorname{sign}(t)\}}$ is the 0/1 loss, then $L_\gamma(z, t) = |z - \gamma|1_{\{\operatorname{sign}(t)\ne\operatorname{sign}(z-\gamma)\}}$. In this case, as indicated in the introduction, we write $R_\gamma(f)$ and R^*_γ instead of $R_{L_\gamma}(f)$ and $R^*_{L_\gamma}$. We also write $C_\gamma(x, t)$ and $C^*_\gamma(x)$ instead of $C_{L,\gamma}(x, t)$ and $C^*_{L,\gamma}(x)$. Basic properties of these quantities are given in Lemma 1.

We are now ready to define the surrogate regret bound. Let $B_\gamma := \sup_{x\in\mathcal{X}} |h(x) - \gamma|$, where recall that $h(x) = E_{Z|X=x}[Z]$. B_γ need not be finite. We may assume $B_\gamma > 0$, for if $B_\gamma = 0$, then by Lemma 1, every decision function has the same excess risk, namely zero. For $\epsilon \in [0, B_\gamma)$, define

$$\mu_{L,\gamma}(\epsilon) = \begin{cases} \inf_{x\in\mathcal{X}:|h(x)-\gamma|\ge\epsilon} H_{L,\gamma}(x), & 0 < \epsilon < B_\gamma, \\ 0, & \epsilon = 0. \end{cases}$$

Note that $\{x : |h(x) - \gamma| \ge \epsilon\}$ is nonempty because $\epsilon < B_\gamma$. Now set $\psi_{L,\gamma}(\epsilon) = \mu^{**}_{L,\gamma}(\epsilon)$ for $\epsilon \in [0, B_\gamma)$, where g^{**} denotes the Fenchel-Legendre biconjugate of g. The biconjugate of g is the largest convex lower semi-continuous function that is $\le g$, and is defined by

$$\operatorname{epi} g^{**} = \overline{\operatorname{co} \operatorname{epi} g},$$

where $\operatorname{epi} g = \{(r, s) : g(r) \le s\}$ is the epigraph of g, co denotes the convex hull, and the bar indicates set closure.

Theorem 1. *Let L be a label-dependent loss and $\gamma \in \mathbb{R}$. For any decision function f and any distribution of (X, Z),*

$$\psi_{L,\gamma}(R_\gamma(f) - R^*_\gamma) \leq R_{L_\gamma}(f) - R^*_{L_\gamma}.$$

A proof is given in Section 5.1, and another in Scott (2011). The former generalizes the argument of Bartlett et al. (2006), while the latter applies ideas from Steinwart (2007).

We show in Lemma 2 that $\psi_{L,\gamma}(0) = 0$ and that $\psi_{L,\gamma}$ is nondecreasing and continuous. For the above bound to be a valid surrogate regret bound, we need for $\psi_{L,\gamma}$ to be strictly increasing and therefore invertible. Thus, we need to find conditions on L and possibly on the distribution of (X, Z) that are sufficient for $\psi_{L,\gamma}$ to be strictly increasing. We adopt the following assumption on L:

(A) There exist $c > 0$, $s \geq 1$ such that

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

This condition was employed by Zhang (2004a) in the context of cost-insensitive classification. He showed that it is satisfied for several common margin losses, i.e., losses having the form $L(y, t) = \phi(yt)$ for some ϕ, including the hinge ($s = 1$), exponential, least squares, truncated least squares, and logistic ($s = 2$) losses. The condition was also employed by Blanchard et al. (2003); Mannor et al. (2003); Lugosi & Vayatis (2004) to analyze certain boosting and greedy algorithms.

Lemma 3 shows that $H_{L,\gamma}(x)$

$$= (h_{1,\gamma}(x) + h_{-1,\gamma}(x))H_L\left(\frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)}\right).$$

When (A) holds with $s = 1$, the $h_{1,\gamma}(x) + h_{-1,\gamma}(x)$ terms cancel out, and one can obtain a lower bound on $H_{L,\gamma}(x)$ in terms of $|h(x) - \gamma|$, leading to the desired lower bound on $\psi_{L,\gamma}$. This is the essence of the proof of following result (see Sec. 5).

Theorem 2. *Let L be a label-dependent loss and $\gamma \in \mathbb{R}$. Assume (A) holds with $s = 1$ and $c > 0$. Then $\psi_{L,\gamma}(\epsilon) \geq \frac{1}{2c}\epsilon$. Furthermore, if $L(y, t) = \max(0, 1 - yt)$ is the hinge loss, then $\psi_{L,\gamma}(\epsilon) = \epsilon$.*

By this result and the following corollary, the modified SVM discussed in the introduction is now justified from the perspective of the surrogate regret.

Corollary 1. *If L is a LD loss, $\gamma \in \mathbb{R}$, and (A) holds with $s = 1$ and $c > 0$, then*

$$R_\gamma(f) - R^*_\gamma \leq 2c(R_{L_\gamma}(f) - R^*_{L_\gamma})$$

for all measurable $f : \mathcal{X} \to \mathbb{R}$. If L is the hinge loss, then

$$R_\gamma(f) - R^*_\gamma \leq R_{L_\gamma}(f) - R^*_{L_\gamma}$$

for all measurable $f : \mathcal{X} \to \mathbb{R}$.

When (A) holds with $s > 1$, to obtain a lower bound on $H_{L,\gamma}(x)$ in terms of $|h(x) - \gamma|$ it is now necessary to upper bound $h_{1,\gamma}(x) + h_{-1,\gamma}(x)$ in terms of $|h(x) - \gamma|$ (please see the proof of Theorem 3 to understand this precisely). We present two sufficient conditions on the distribution of (X, Z), either of which make this possible. The conditions require that the conditional distribution of Z given $X = x$ is not too heavy tailed.

(B) $\exists C > 0, \beta \geq 1$ such that $\forall x \in \mathcal{X}$,

$$P_{Z|X=x}(|Z - h(x)| \geq t) \leq Ct^{-\beta}, \ \forall t > 0.$$

(C) $\exists C, C' > 0$ such that $\forall x \in \mathcal{X}$,

$$P_{Z|X=x}(|Z - h(x)| \geq t) \leq Ce^{-C't^2}, \ \forall t > 0.$$

By Chebyshev's inequality, condition (B) holds provided $Z|X = x$ has uniformly bounded variance. In particular, if $\mathrm{Var}(Z|X = x) \leq \sigma^2 \ \forall x$, then (B) holds with $\beta = 2$ and $C = \sigma^2$.

Condition (C) holds when $Z|X = x$ is subGaussian with bounded variance. For example, (C) holds if $Z|X = x \sim \mathcal{N}(h(x), \sigma_x^2)$ with σ_x^2 bounded. Alternatively, (C) holds if $Z|X = x$ has bounded support $\subseteq [a, b]$, where a and b do not depend on x.

Theorem 3. *Let L be a label-dependent loss and $\gamma \in \mathbb{R}$. Assume (A) holds with exponent $s \geq 1$. If (B) holds with exponent $\beta > 1$, then there exist $c_1, c_2, \epsilon_0 > 0$ such that for all $\epsilon \in [0, B_\gamma)$,*

$$\psi_{L,\gamma}(\epsilon) \geq \begin{cases} c_1\epsilon^{s+(\beta-1)(s-1)}, & \epsilon \leq \epsilon_0 \\ c_2(\epsilon - \epsilon_0), & \epsilon > \epsilon_0. \end{cases}$$

If (C) holds, then there exist $c_1, c_2, \epsilon_0 > 0$ such that for all $\epsilon \in [0, B_\gamma)$

$$\psi_{L,\gamma}(\epsilon) \geq \begin{cases} c_1\epsilon^{s}, & \epsilon \leq \epsilon_0 \\ c_2(\epsilon - \epsilon_0), & \epsilon > \epsilon_0. \end{cases}$$

In both cases, the lower bounds are convex on $[0, B_\gamma)$.

To reiterate the significance of these results: Since $\psi_{L,\gamma}(0) = 0$ and $\psi_{L,\gamma}$ is nondecreasing, continuous, and convex, Theorems 2 and 3 imply that $\psi_{L,\gamma}$ is invertible, leading to the surrogate regret bound $R_\gamma(f) - R^*_\gamma \leq \psi^{-1}_{L,\gamma}(R_{L_\gamma}(f) - R^*_{L_\gamma})$.

Corollary 2. *Let L be a LD loss and $\gamma \in \mathbb{R}$. Assume (A) holds with exponent $s \geq 1$. If (B) holds with exponent $\beta > 1$, then there exist $K_1, K_2 > 0$ such that*

$$R_\gamma(f) - R_\gamma^* \leq K_1(R_{L_\gamma}(f) - R_{L_\gamma}^*)^{1/(s+(\beta-1)(s-1))}$$

for all measurable f with $R_{L_\gamma}(f) - R_{L_\gamma}^* \leq K_2$. If (C) holds, then there exist $K_1, K_2 > 0$ such that

$$R_\gamma(f) - R_\gamma^* \leq K_1(R_{L_\gamma}(f) - R_{L_\gamma}^*)^{1/s}$$

for all measurable f with $R_{L_\gamma}(f) - R_{L_\gamma}^* \leq K_2$.

It is possible to improve the rate when $s > 1$ provided the following distributional assumption is valid.

(D) There exists $\alpha \in (0, 1]$ and $c > 0$ such that for all measurable $f : \mathcal{X} \to \mathbb{R}$,

$$P(\text{sign}(f(X)) \neq \text{sign}(h(X)-\gamma)) \leq c(R_\gamma(f)-R_\gamma^*)^\alpha.$$

We refer to α as the *noise exponent*. This condition generalizes a condition for CI classification introduced by Tsybakov (2004) and subsequently adopted by several authors. Some insight into the condition is offered by the following result.

Proposition 1. *(D) is satisfied with $\alpha \in (0, 1)$ if there exists $B > 0$ such that for all $t \geq 0$,*

$$P(|h(X) - \gamma| \leq t) \leq Bt^{\frac{\alpha}{1-\alpha}}.$$

(D) is satisfied with $\alpha = 1$ if there exists $t_0 > 0$ such that

$$P(|h(X) - \gamma| \geq t_0) = 1.$$

Similar conditions have been adopted in previous work on level set estimation (Polonik, 1995) and CI classification (Mammen & Tsybakov, 1999). From the proposition we see that for larger α, there is less noise in the sense of less uncertainty near the optimal decision boundary.

Theorem 4. *Let L be a LD loss and $\gamma \in \mathbb{R}$. Assume (A) holds with $s > 1$, and (D) holds with noise exponent $\alpha \in (0, 1]$. If (B) holds with exponent $\beta > 1$, then there exist $K_1, K_2 > 0$ such that*

$$\begin{aligned} R_\gamma(f) - R_\gamma^* & \\ \leq \quad & K_1(R_{L_\gamma}(f) - R_{L_\gamma}^*)^{1/(s+(\beta-1)(s-1)-\alpha\beta(s-1))} \end{aligned}$$

for all measurable $f : \mathcal{X} \to \mathbb{R}$ with $R_{L_\gamma}(f) - R_{L_\gamma}^* \leq K_2$. If (C) holds, then there exist $K_1, K_2 > 0$ such that

$$R_\gamma(f) - R_\gamma^* \leq K_1(R_{L_\gamma}(f) - R_{L_\gamma}^*)^{1/(s-\alpha(s-1))}$$

for all measurable $f : \mathcal{X} \to \mathbb{R}$ with $R_{L_\gamma}(f) - R_{L_\gamma}^* \leq K_2$.

The proof of this result combines Theorem 3 with an argument presented in Bartlett et al. (2006).

4. Conclusions

This work gives theoretical justification to the cost-sensitive SVM with example-dependent costs, described in the introduction. It also suggests principled design criteria for new algorithms based on other specific losses. For example, consider the surrogate loss $L_\gamma(z, t)$ based on the label-dependent loss $L(y, t) = e^{-yt}$. To minimize the empirical risk $\frac{1}{n}\sum_{i=1}^n L_\gamma(z_i, f(x_i))$ over a class of linear combinations of some base class, a functional gradient descent approach may be employed, giving rise to a natural kind of boosting algorithm in this setting. Since the loss here differs from the loss in cost-insensitive boosting by scalar factors only, similar computational procedures are feasible. Obviously, similar statements apply to other losses, such as the logistic loss and logistic regression type algorithms.

Another natural next step is to prove consistency for specific algorithms. Surrogate regret bounds have been used for this purpose in the context of cost-insensitive classification by (Mannor et al., 2003; Blanchard et al., 2003; Zhang, 2004a; Lugosi & Vayatis, 2004; Steinwart, 2005). These proofs typically require two additional ingredients in addition to surrogate regret bounds: a class of classifiers with the universal approximation property (to achieve universal consistency), together with a uniform bound on the deviation of the empirical surrogate risk from its expected value. We anticipate that such proof strategies can be extended to CS classification with ED costs.

Finally, we remark that our theory applies to some of the special cases mentioned in Section 2. For example, for CS classification with LD costs, condition (C) holds, and we get surrogate regret bounds for this case (see also Scott (2010)).

5. Proofs

We begin with some lemmas. The first lemma presents some basic properties of the target risk and conditional risk. These extend known results for CI classification (Devroye et al., 1996).

Lemma 1. *Let $\gamma \in \mathbb{R}$. (1) $\forall x \in \mathcal{X}$,*

$$h(x) - \gamma = h_{1,\gamma}(x) - h_{-1,\gamma}(x).$$

(2) $\forall x \in \mathcal{X}, t \in \mathbb{R}$,

$$C_\gamma(x, t) - C_\gamma^*(x) = 1_{\{\text{sign}(t) \neq \text{sign}(h(x)-\gamma)\}}|h(x) - \gamma|.$$

(3) For any measurable $f : \mathcal{X} \to \mathbb{R}$,

$$R_\gamma(f) - R_\gamma^* = E_X[1_{\{\text{sign}(f(X)) \neq \text{sign}(h(X)-\gamma)\}}|h(X) - \gamma|].$$

Proof. (1) $h_{1,\gamma}(x) - h_{-1,\gamma}(x) = E_{Z|X=x}[(Z-\gamma)1_{\{Z>\gamma\}} - (\gamma-Z)1_{\{Z\leq\gamma\}}] = E_{Z|X=x}[Z-\gamma] = h(x)-\gamma$.

(2) Since $C_\gamma(x,t) = h_{1,\gamma}(x)1_{\{t\leq 0\}} + h_{-1,\gamma}(x)1_{\{t>0\}}$, a value of t minimizing this quantity (for fixed x) must satisfy $\text{sign}(t) = \text{sign}(h_{1,\gamma}(x) - h_{1,\gamma}(x)) = \text{sign}(h(x) - \gamma)$. Therefore, $\forall x \in \mathcal{X}, t \in \mathbb{R}$, $C_\gamma(x,t) - C^*_\gamma(x) = [h_{1,\gamma}(x)1_{\{t\leq 0\}} + h_{-1,\gamma}(x)1_{\{t>0\}}] - [h_{1,\gamma}(x)1_{\{h(x)\leq\gamma\}} + h_{-1,\gamma}(x)1_{\{h(x)>\gamma\}}] = 1_{\{\text{sign}(t)\neq\text{sign}(h(x)-\gamma)\}}|h_{1,\gamma}(x) - h_{-1,\gamma}(x)| = 1_{\{\text{sign}(t)\neq\text{sign}(h(x)-\gamma)\}}|h(x)-\gamma|$.

(3) now follows from (2). □

The next lemma records some basic properties of $\psi_{L,\gamma}$.

Lemma 2. *Let L be any LD loss and $\gamma \in \mathbb{R}$. Then (1) $\psi_{L,\gamma}(0) = 0$. (2) $\psi_{L,\gamma}$ is nondecreasing. (3) $\psi_{L,\gamma}$ is continuous on $[0, B_\gamma)$.*

Proof. From the definition of $\mu_{L,\gamma}$, $\mu_{L,\gamma}(0) = 0$ and $\mu_{L,\gamma}$ is nondecreasing. (1) and (2) now follow. Since $\text{epi}\,\psi_{L,\gamma}$ is closed by definition, $\psi_{L,\gamma}$ is lower semi-continuous. Since $\psi_{L,\gamma}$ is convex on the simplical domain $[0, B_\gamma)$, it is upper semi-continuous by Theorem 10.2 of Rockafellar (1970). □

The next lemma is needed for Theorems 2 and 3. An analogous identity was presented by Steinwart (2007) for label-dependent margin losses.

Lemma 3. *For any LD loss L and $\gamma \in \mathbb{R}$, and for all $x \in \mathcal{X}$ such that $h_{1,\gamma}(x) + h_{-1,\gamma}(x) > 0$,*

$$
\begin{aligned}
&H_{L,\gamma}(x) \\
&\quad = \quad (h_{1,\gamma}(x) + h_{-1,\gamma}(x)) H_L\left(\frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)}\right).
\end{aligned}
$$

Proof. Introduce $w_\gamma(x) = h_{1,\gamma}(x) + h_{-1,\gamma}(x)$ and $\vartheta_\gamma(x) = h_{1,\gamma}(x)/(h_{1,\gamma}(x) + h_{-1,\gamma}(x))$. If $w_\gamma(x) > 0$, then

$$
\begin{aligned}
C_{L,\gamma}(x,t) &= h_{1,\gamma}(x)L_1(t) + h_{-1,\gamma}(x)L_{-1}(t) \\
&= w_\gamma(x)[\vartheta_\gamma(x)L_1(t) + (1-\vartheta_\gamma(x))L_{-1}(t)] \\
&= w_\gamma(x)C_L(\vartheta_\gamma(x), t).
\end{aligned}
$$

By Lemma 1, $h(x) - \gamma = w_\gamma(x)(2\vartheta_\gamma(x) - 1)$. Since $w_\gamma(x) > 0$, $h(x) - \gamma$ and $2\vartheta_\gamma(x) - 1$ have the same sign. Therefore

$$
\begin{aligned}
C^-_{L,\gamma}(x) &= \inf_{t:t(h(x)-\gamma)\leq 0} C_{L,\gamma}(x,t) \\
&= w_\gamma(x) \inf_{t:t(h(x)-\gamma)\leq 0} C_L(\vartheta_\gamma(x), t) \\
&= w_\gamma(x) \inf_{t:t(2\vartheta_\gamma(x)-1)\leq 0} C_L(\vartheta_\gamma(x), t) \\
&= w_\gamma(x) C^-_L(\vartheta_\gamma(x)).
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
C^*_{L,\gamma}(x) &= w_\gamma(x) \inf_{t\in\mathbb{R}} C_L(\vartheta_\gamma(x), t) \\
&= w_\gamma(x) C^*_L(\vartheta_\gamma(x)).
\end{aligned}
$$

Thus

$$
\begin{aligned}
H_{L,\gamma}(x) &= C^-_{L,\gamma}(x) - C^*_{L,\gamma}(x) \\
&= w_\gamma(x)[C^-_L(\vartheta_\gamma(x)) - C^*_L(\vartheta_\gamma(x))] \\
&= w_\gamma(x) H_L(\vartheta_\gamma(x)).
\end{aligned}
$$

□

5.1. Proof of Theorem 1

By Jensen's inequality and Lemma 1,

$$
\begin{aligned}
&\psi_{L,\gamma}(R_\gamma(f) - R^*_\gamma) \\
&\quad\leq E_X[\psi_{L,\gamma}(1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}}|h(X)-\gamma|)] \\
&\quad= E_X[1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}}\psi_{L,\gamma}(|h(X)-\gamma|)] \\
&\quad\leq E_X[1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}}\mu_{L,\gamma}(|h(X)-\gamma|)] \\
&\quad= E_X\Bigg[1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}} \cdot \\
&\qquad\qquad \inf_{x\in\mathcal{X}:|h(x)-\gamma|\geq|h(X)-\gamma|} H_{L,\gamma}(x)\Bigg] \\
&\quad\leq E_X[1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}} H_{L,\gamma}(X)] \\
&\quad= E_X\Bigg[1_{\{\text{sign}(f(X))\neq\text{sign}(h(X)-\gamma)\}} \cdot \\
&\qquad\qquad \left(\inf_{t:t(h(X)-\gamma)\leq 0} C_{L,\gamma}(X,t) - C^*_{L,\gamma}(X)\right)\Bigg] \\
&\quad\leq E_X[C_{L,\gamma}(X, f(X)) - C^*_{L,\gamma}(X)] \\
&\quad= R_{L_\gamma}(f) - R^*_{L_\gamma}.
\end{aligned}
$$

□

5.2. Proof of Theorem 2

Let $\epsilon > 0$ and $x \in \mathcal{X}$ such that $|h(x) - \gamma| \geq \epsilon$. It is necessary that $h_{1,\gamma}(x) + h_{-1,\gamma}(x) > 0$. This is because $h_{1,\gamma}(x) + h_{-1,\gamma}(x) = E_{Z|X=x}[|Z-\gamma|]$, and if this is 0, then $Z = \gamma$ almost surely ($P_{Z|X=x}$). But then $h(x) = \gamma$, contradicting $|h(x) - \gamma| \geq \epsilon > 0$. Therefore we may apply Lemma 3 and condition (A) with $s = 1$ to obtain

$$
\begin{aligned}
&H_{L,\gamma}(x) \\
&\quad= (h_{1,\gamma}(x) + h_{-1,\gamma}(x)) H_L\left(\frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)}\right) \\
&\quad\geq (h_{1,\gamma}(x) + h_{-1,\gamma}(x)) \frac{1}{c}\left|\frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)} - \frac{1}{2}\right| \\
&\quad= (h_{1,\gamma}(x) + h_{-1,\gamma}(x)) \frac{1}{2c}\left|\frac{h_{1,\gamma}(x) - h_{-1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)}\right| \\
&\quad= \frac{1}{2c}|h(x) - \gamma| \geq \frac{1}{2c}\epsilon,
\end{aligned}
$$

where in the next to last step we applied Lemma 1. Therefore $\mu_{L,\gamma}(\epsilon) \geq \frac{1}{2c}$. The result now follows. □

5.3. Proof of Theorem 3

Assume (A) and (B) hold. If $s = 1$ the result follows by Theorem 2, so let's assume $s > 1$. Let $\epsilon > 0$ and $x \in \mathcal{X}$ such that $|h(x) - \gamma| \geq \epsilon$. As in the proof of Theorem 2, we have

$$\begin{aligned} & H_{L,\gamma}(x) \\ &= (h_{1,\gamma}(x) + h_{-1,\gamma}(x)) H_L \left(\frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)} \right) \\ &\geq \frac{(h_{1,\gamma}(x) + h_{-1,\gamma}(x))}{c^s} \left| \frac{h_{1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)} - \frac{1}{2} \right|^s \\ &= \frac{(h_{1,\gamma}(x) + h_{-1,\gamma}(x))}{(2c)^s} \left| \frac{h_{1,\gamma}(x) - h_{-1,\gamma}(x)}{h_{1,\gamma}(x) + h_{-1,\gamma}(x)} \right|^s \\ &= \frac{1}{(2c)^s} \frac{|h(x) - \gamma|^s}{(h_{1,\gamma}(x) + h_{-1,\gamma}(x))^{s-1}}. \end{aligned}$$

The next step is to find an upper bound on $w_\gamma(x) = h_{1,\gamma}(x) + h_{-1,\gamma}(x)$ in terms of $|h(x) - \gamma|$, which will give a lower bound on $H_{L,\gamma}(x)$ in terms of $|h(x) - \gamma|$.

For now, assume $h_{1,\gamma}(x) < h_{-1,\gamma}(x)$. Then $w_\gamma(x) = 2h_{1,\gamma}(x) + |h(x) - \gamma|$. Let us write $h_{1,\gamma}(x) = E_{W|X=x}[W]$ where $W = (Z - \gamma)1_{\{Z > \gamma\}} \geq 0$. Then $h_{1,\gamma}(x) = \int_0^\infty P_{W|X=x}(W \geq w) dw$. Now

$$\begin{aligned} & P_{W|X=x}(W \geq w) = P_{Z|X=x}(Z - \gamma \geq w) \\ &\quad = P_{Z|X=x}(Z - h(x) + h(x) - \gamma \geq w) \\ &\quad = P_{Z|X=x}(Z - h(x) \geq w + |h(x) - \gamma|) \\ &\quad \leq P_{Z|X=x}(|Z - h(x)| \geq w + |h(x) - \gamma|) \\ &\quad \leq C(w + |h(x) - \gamma|)^{-\beta} \end{aligned}$$

by (B). Then

$$\begin{aligned} h_{1,\gamma}(x) &\leq \int_0^\infty C(w + |h(x) - \gamma|)^{-\beta} dw \\ &= \frac{C}{\beta - 1} |h(x) - \gamma|^{-(\beta-1)}. \end{aligned}$$

Therefore

$$w_\gamma(x) \leq \frac{2C}{\beta - 1} |h(x) - \gamma|^{-(\beta-1)} + |h(x) - \gamma|.$$

Setting $\Delta = |h(x) - \gamma|$ and $c' = 2C/(\beta - 1)$ for brevity, we have

$$H_{L,\gamma}(x) \geq \frac{1}{(2c)^s} \frac{\Delta^s}{(\Delta + c'\Delta^{-(\beta-1)})^{s-1}}.$$

Using similar reasoning, the same lower bound can be established in the case where $h_{1,\gamma}(x) > h_{-1,\gamma}(x)$. Let us now find a simpler lower bound. Notice that $\Delta = c'\Delta^{-(\beta-1)}$ when $\Delta = \Delta_0 = (c')^{1/\beta}$. When $\Delta \leq \Delta_0$,

$$\begin{aligned} H_{L,\gamma}(x) &\geq \frac{1}{(2c)^s} \frac{\Delta^s}{(2c'\Delta^{-(\beta-1)})^{s-1}} \\ &= c_1 \Delta^{s + (\beta-1)(s-1)} \end{aligned}$$

where $c_1 = (2c)^{-s}(2c')^{-(s-1)}$. When $\Delta \geq \Delta_0$,

$$H_{L,\gamma}(x) \geq \frac{1}{(2c)^s} \frac{\Delta^s}{(2\Delta)^{s-1}} = c_2 \Delta$$

where $c_2 = 2^{-(2s-1)} c^{-s}$. Putting these two cases together,

$$\begin{aligned} H_{L,\gamma}(x) &\geq \min(c_1 \Delta^{s + (\beta-1)(s-1)}, c_1 \Delta) \\ &\geq \min(c_1 \epsilon^{s + (\beta-1)(s-1)}, c_2 \epsilon) \end{aligned}$$

since $\Delta = |h(x) - \gamma| \geq \epsilon$. Now shift the function $c_2\epsilon$ to the right by some ϵ_0 so that it is tangent to $c_1 \epsilon^{s + (\beta-1)(s-1)}$. The resulting piecewise function is a closed and convex lower bound on $\mu_{L,\gamma}$. Since $\psi_{L,\gamma}$ is the largest such lower bound, the proof is complete in this case.

Now suppose (A) and (C) hold. The proof is the same up to the point where we invoke (B). At that point, we now obtain

$$\begin{aligned} & h_{1,\gamma}(x) \\ &\leq \int_0^\infty P_{Z|X=x}(|Z - h(x)| \geq w + |h(x) - \gamma|) dw \\ &\leq \int_0^\infty C e^{-C'(w + |h(x) - \gamma|)^2} dw \\ &= C\sqrt{2\pi\sigma^2} \int_0^\infty \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(w + |h(x) - \gamma|)^2 / 2\sigma^2} dw \\ &\qquad [\sigma^2 = 1/2C'] \\ &= C\sqrt{2\pi\sigma^2} P(W' \geq |h(x) - \gamma|) \\ &\qquad [\text{where } W' \sim \mathcal{N}(0, \sigma^2)] \\ &\leq C\sqrt{2\pi\sigma^2} e^{-|h(x) - \gamma|^2 / 2\sigma^2} \\ &= C\sqrt{\frac{\pi}{C'}} e^{-C'|h(x) - \gamma|^2}. \end{aligned}$$

The final inequality is a standard tail inequality for the Gaussian distribution (Ross, 2002). Now

$$w_\gamma(x) \leq \Delta + C'' e^{-C'\Delta^2}$$

where $\Delta = |h(x) - \gamma|$ and $C'' = 2C\sqrt{\pi/C'}$, and therefore

$$H_{L,\gamma}(x) \geq \frac{1}{(2c)^s} \frac{\Delta^s}{(\Delta + C'' e^{-C'\Delta^2})^{s-1}}.$$

The remainder of the proof is now analogous to the case when (B) was assumed to hold. □

Acknowledgements

Supported in part by NSF grants CCF-0830490 and CCF-0953135.

References

Bartlett, P., Jordan, M., and McAuliffe, J. Convexity, classification, and risk bounds. *J. Amer. Statist. Assoc.*, 101(473):138–156, 2006.

Blanchard, G., Lugosi, G., and Vayatis, N. On the rate of convergence of regularized boosting classifiers. *J. Machine Learning Research*, 4:861–894, 2003.

Brefeld, U., Geibel, P., and Wysotzki, F. Support vector machines with example-dependent costs. In *Proc. Euro. Conf. Machine Learning*, 2003.

Cavalier, L. Nonparametric estimation of regression level sets. *Statistics*, 29:131–160, 1997.

Devroye, L., Györfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition*. Springer, New York, 1996.

Elkan, C. The foundations of cost-sensitive learning. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence*, pp. 973–978, Seattle, Washington, USA, 2001.

Lugosi, G. and Vayatis, N. On the Bayes risk consistency of regularized boosting methods. *The Annals of statistics*, 32(1):30–55, 2004.

Mammen, E. and Tsybakov, A. B. Smooth discrimination analysis. *Ann. Stat.*, 27:1808–1829, 1999.

Mannor, S., Meir, R., and Zhang, T. Greedy algorithms for classification–consistency, convergence rates, and adaptivity. *J. Machine Learning Research*, 4:713–742, 2003.

Polonik, W. Measuring mass concentrations and estimating density contour clusters–an excess mass approach. *Ann. Stat.*, 23(3):855–881, 1995.

Polonik, W. and Wang, Z. Estimation of regression contour clusters–an application of the excess mass approach to regression. *J. Multivariate Analysis*, 94:227–249, 2005.

Reid, M. and Williamson, R. Information, divergence and risk for binary experiments. *J. Machine Learning Research*, 12:731–817, 2011.

Rockafellar, R. T. *Convex Analysis*. Princeton University Press, Princeton, NJ., 1970.

Ross, S. *A First Course in Probability*. Prentice Hall, 2002.

Scott, C. Calibrated surrogate losses for classification with label-dependent costs. Technical report, arXiv:1009.2718v1, September 2010.

Scott, C. Surrogate losses and regret bounds for classification with example-dependent costs. Technical Report CSPL-400, University of Michigan, 2011. URL http://www.eecs.umich.edu/~cscott.

Scott, C. and Davenport, M. Regression level set estimation via cost-sensitive classification. *IEEE Trans. Signal Proc.*, 55(6):2752–2757, 2007.

Steinwart, I. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Inform. Theory*, 51(1):128–142, 2005.

Steinwart, I. How to compare different loss functions and their risks. *Constructive Approximation*, 26(2): 225–287, 2007.

Tewari, A. and Bartlett, P. On the consistency of multiclass classification methods. *J. Machine Learning Research*, 8:1007–1025, 2007.

Tsybakov, A. B. Optimal aggregation of classifiers in statistical learning. *Ann. Stat.*, 32(1):135–166, 2004.

Willett, R. and Nowak, R. Minimax optimal level set estimation. *IEEE Transactions on Image Processing*, 16(12):2965–2979, 2007.

Zadrozny, B. and Elkan, C. Learning and making decisons when costs and probabilities are both unknown. In *Proc. 7th Int. Conf. on Knowledge Discovery and Data Mining*, pp. 204–213. ACM Press, 2001.

Zadrozny, B., Langford, J., and Abe, N. Cost sensitive learning by cost-proportionate example weighting. In *Proceedings of the 3rd International Conference on Data Mining*, Melbourne, FA, USA, 2003. IEEE Computer Society Press.

Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–85, 2004a.

Zhang, T. Statistical analysis of some multi-category large margin classification methods. *J. Machine Learning Research*, 5:1225–1251, 2004b.