Penalized empirical risk minimization in the estimation of thresholds

Leila Mohammadi

EURANDOM

October 5, 2005

Introduction

* (X, Y), $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d.

$$\mathbf{X} = (X_1, \ldots, X_n)$$

* $X \in \mathcal{X}$ *instance*

* $Y \in \{-1, 1\}$ *label*

Definitions

- Base classifier $h : \mathcal{X} \to \{-1, 1\}$.
- Collection of base classifiers $\mathcal{H}$.
- Misclassification if $Yh(X) < 0$
- Loss function $l : \mathbb{R} \to [0, \infty)$

- Empirical risk

$$L_n(h) := E_n(l(-Yh(X))) = \frac{1}{n}\sum_{i=1}^{n} l(-Y_i h(X_i)),$$

- Theoretical risk

$$L(h) := E(l(-Yh(X))).$$

Penalized empirical risk minimizer

- The class $\mathcal{H}$ is too complex
- Consider a penalty p over $\mathcal{H}$

$$\hat{h}_n := \arg\min_{h \in \mathcal{H}} (L_n(h) + p^2(h))$$

$$h_0 := \arg\min_{h \in \mathcal{H}} L(h)$$

$$\tau^2(h|\tilde{h}) := L(h) - L(\tilde{h}) + p^2(h)$$

$$\mathcal{H}_0(\delta) := \{h \in \mathcal{H} : \tau^2(h|h_0) \leq \delta^2\}.$$

$$U_i(h) := l(-Y_i h(X_i)) - E(l(-Y_i h(X_i))|\mathbf{X}).$$

Assume that $d_1, \ldots, d_n$ are some metrics on $\mathcal{H}$,

$$|U_i(h) - U_i(\tilde{h})| \leq |W_i| d_i(h(X_i), \tilde{h}(X_i)),$$
$$i = 1, \ldots, n, \ h, \tilde{h} \in \mathcal{H},$$

where $W_1, \ldots, W_n$ are uniformly sub-Gaussian, so that for an M and σ_0^2,

$$\max_{i=1,\ldots,n} M^2(E(\exp[|W_i|^2/M^2]) - 1) \leq \sigma_0^2.$$

$$d^2(h, \tilde{h}) := \frac{1}{n} \sum_{i=1}^{n} d_i^2(h, \tilde{h}).$$

Definition Let T be a (subset of a) metric space endowed with a metric m. The u-covering number $N(u, T; m)$ is defined as the number of balls with radius u necessary to cover T with respect to the metric m. The u-entropy is defined as $H(u, T; m) := \log N(u, T; m)$.

$$v_n(h) := L_n(h) - L(h).$$

Lemma 1. *Suppose all the above assumptions hold and let* $\sup_{h\in\mathcal{H}} d(h, h_0) \leq R$. *Then, for some* c_1 *depending only on* M *and* σ_0, *and for all* $\sigma > 0$ *and* $\delta > 0$ *satisfying*

$$\sqrt{n}\delta \geq c_1 \left(\int_{\delta/(8\sigma)}^{R} H^{1/2}(u, \mathcal{H}; d) du \vee R \right),$$

we have

$$\mathbf{P}\left(\sup_{h\in\mathcal{H}} |v_n(h_0) - v_n(h)| \geq \delta \wedge \frac{1}{n}\sum_{i=1}^{n} W_i^2 \leq \sigma^2 \right)$$

$$\leq c_1 \exp\left[-\frac{n\delta^2}{c_1^2 R^2} \right].$$

Assumption (A).

There are $\eta > 0$ and $k > 1$, such that

$$L(h) - L(h_0) \geq \eta d^k(h, \tilde{h}), \quad \forall h \in \mathcal{H}.$$

Theorem 1. *Suppose Assumption (A). Let*

$$\Psi(\delta) \geq \int_0^\delta H^{1/2}(u, \mathcal{H}_0(\delta); d)du \vee \delta,$$

and assume that $\Psi(\delta^{2/k})/\delta^2$ *is a non-increasing function of* δ, $\delta > 0$. *There exists a constant* c_2 *such that for*

$$\sqrt{n}\delta_n^2 \geq c_2 \Psi(\delta_n^{2/k}),$$

and for all $\delta \geq \delta_n$,

$$\mathbf{P}(\tau^2(\hat{h}_n | h_0) \geq 2(p^2(h_0) + \delta^2))$$
$$\leq c_2 \exp\left[-\frac{n\delta^{4(1-1/k)}}{c_2^2}\right].$$

Sketch of a proof

We use

$$L_n(\hat{h}_n) + p^2(\hat{h}_n) \le L_n(h_0) + p^2(h_0)$$

or

$$L(\hat{h}_n) - L(h_0) + p^2(\hat{h}_n)$$
$$\le [L_n(h_0) - L(h_0) - (L_n(\hat{h}_n) - L(\hat{h}_n))] + p^2(h_0)$$

or

$$\tau^2(\hat{h}_n|h_0) - p^2(h_0) \le v_n(h_0) - v_n(\hat{h}_n).$$

We obtain

$$\mathbf{P}(\tau^2(\hat{h}_n|h_0) \ge 2(p^2(h_0) + \delta^2))$$
$$\le \sum_{s=1}^{\infty} \mathbf{P}\left(\sup_{h \in \mathcal{H}_0(2^s\delta)} |v_n(h_0) - v_n(h)| \ge \frac{1}{12} 2^{2s}\delta^2 \right).$$

If $h \in \mathcal{H}_0(2^s\delta)$, then, $d(h, h_0) \leq \frac{2^{2s/k}\delta^{2/k}}{\eta^{1/k}}$, by Assumption (A).

We use Lemma 1, for $\sqrt{n}r \geq c_3\Psi(\frac{2^{2s/k}\delta^{2/k}}{\eta^{1/k}})$, one has

$$\mathbf{P}\left(\sup_{h\in\mathcal{H}_0(2^s\delta)} |v_n(h_0) - v_n(h)| \geq r\right)$$

$$\leq c_3 \exp\left[-\frac{nr^2\eta^{2/k}}{c_3^2 2^{4s/k}\delta^{4/k}}\right].$$

□

Here is a simple consequence of Theorem 1.

Theorem 2. *Under the conditions of Theorem 1, for $k = 2$, we arrive at the inequality*

$$E(\tau^2(\hat{h}_n|h_0)) \leq 2(p^2(h_0) + \delta_n^2) + \frac{c_4}{n}$$

where c_4 is a constant depending only on M.

An application: Threshold estimation

Let $\mathcal{X} = [0, 1]$,

$$h_a(x) = \sum_{k=1}^{K+1} b_k \mathbb{1}\{a_{k-1} \leq x < a_k\},$$

$$\mathcal{H} = \{h_a(x) : a \in U\},\ b_1, \ldots, b_{K+1} \in \{-1, 1\},$$

where

$$U = U_K = \{a = (a_1, \ldots, a_K) \in (0, 1)^K : a_1 < \ldots < a_K\}.$$

$a_0 = 0$ and $a_{K+1} = 1$.

- Prediction error (theoretical error)

$$L(f) = P(Yf(X) < 0).$$

- Percentage misclassified (empirical error)

$$L_n(f) = \frac{\#\{Y_i f(X_i) < 0,\ 1 \leq i \leq n\}}{n}.$$

- Regression

$$F_0(x) = P(Y = 1 | X = x).$$

$$l(t) := \mathbb{1}(t > 0)$$

$$d^2(h, \tilde{h}) := P(h(X) \neq \tilde{h}(X))$$

$$a_0 := \arg\min_a L(h_a)$$

Assumption (B). There is an $\eta > 0$, such that

$|2F_0(x)-1| > \eta, \ \ \forall x \in (0,1), \ \ x \neq a_{0,i}, \ \ i = 1, \ldots, K.$

$$V := \{h_a \neq h_{a_0}\}$$

$$L(h_a) - L(h_{a_0}) \geq \eta d^2(h_a, h_{a_0}).$$

$$h = h_a \in \cup_{K=1}^{\infty} \mathcal{H}_1(K), \quad K = K_a$$

$$p^2(h_a) := \lambda^2 \frac{K_a}{n}.$$

Lemma 2. *Consider the L_1 metric on $\mathbb{R}^m$,*

$$\tilde{d}(a, b) := \sum_{i=1}^{m} |a_i - b_i|,$$

$$a = (a_1, \ldots, a_m), \ b = (b_1, \ldots, b_m).$$

Then a ball $\tilde{B}_m(R)$ (with respect to the above metric) in $\mathbb{R}^m$ can be covered by $N \leq (\lfloor \frac{R}{u} \rfloor + 1)^m$ balls with radius u.

Theorem 3. *Suppose Assumption (B). Take $\lambda \geq c_5 \sqrt{\log n}$, where c_5 is a large constant depending only on M. Then*

$$E(\tau^2(\hat{h}_n | h_0)) \leq 2p^2(h_0) + \frac{c_5}{n}.$$