Nonparametric Least Square Estimator

Luo Fangzhi

Sun Yat-sen University

Examples of Nonparametric Least Square

Kernel Ridge Regression : Let H be a reproducing kernel Hilbert space, equipped with the norm $\|\cdot\|_H$.
A constraint nonparametric least square estimator is by minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \tag{1}$$

subject to $\|f\|_{\mathbb{H}} < C$. Under some condition, this is equivalent to solve

$$\widehat{f} \in \arg\min_{f\in\mathbb{H}} \left\{ \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda_n \|f\|_{\mathbb{H}}^2 \right\}.$$

for a corresponding regularization parameter $\lambda_n > 0$. As mentioned in last chapter, if the kernel of $\mathbb{H}$ is $\mathcal{K}$, and the correponding kernel function is $K(s,t)$, then the solution to f is $\widehat{f}(\cdot) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\widehat{\alpha}_i\mathcal{K}(\cdot, x_i)$, where $\widehat{\alpha} := (\mathbf{K} + \lambda_n \mathbf{I}_n)^{-1}\frac{y}{\sqrt{n}}$.

Problem Setup

- obs : (x_i, y_i)
- model : $y_i = f^*(x_i) + v_i, v_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$

We want to evaluate f^* by

$$\text{argmin}_{f \in \mathcal{F}} \sum_i (y_i - f(x_i))^2. \tag{2}$$

The first intuition is the difficulty of evaluate f^* is from the complexity of $\mathcal{F}$.

The quality of evaluation is estimated by

$$\frac{1}{n} \sum_i (\hat{f}(x_i) - f^*(x_i))^2. \tag{3}$$

Remark that this is evaluation converge to $\mathbb{E}_{X,Y}[\hat{f}(x) - f^*(x)]^2$.

The Nonasymptotic Bound

Theorem

for any c, $\exists\delta_c$ *such that,* $\forall u > \delta_c$, $\frac{\mathbb{E}\mathcal{G}(u,\mathcal{F}^*)}{u} < c$. *Then* $\forall c_1$ *subject to* $2\sigma(c + c1) > \delta_c$,

$$\mathbb{P}[\|\hat{f} - f^*\|_n > 2\sigma(c + c1)] \leq exp(-nc_1^2) \tag{4}$$

The Nonasymptotic Bound

Our goal is to establish a nonasymptotic bound of $\sum_i [\hat{f}(x_i) - f^*(x_i)]^2$. First tool is the basic inequality

$$\sum_i [y_i - f^*(x_i)]^2 \geq \sum_i [y_i - \hat{f}(x_i)]^2. \tag{5}$$

What follows can be obatined by simple transformation of the above inequality.

$$\frac{1}{2n} \sum_i [\hat{f}(x_i) - f^*(x_i)]^2 \leq \frac{\sigma}{n} \sum_i \omega_i [\hat{f}(x_i) - f^*(x_i)]. \tag{6}$$

Or equivalently,

$$\sum_i [\hat{f}(x_i) - f^*(x_i)]^2 \leq 2\sigma \sum_i \omega_i [\hat{f}(x_i) - f^*(x_i)]. \tag{7}$$

The right handside is very similar to a sub-Gaussian. But some observation is the right hand side is not L-Lipschitz continuous. So we cannot simple apply Thm 2.26.

The Nonasymptotic Bound

Moreover, if we let $\mathcal{G}(\delta, \mathcal{F}^*) = \sup_{g \in \mathcal{F}^*, \|g\|_n < \delta} \frac{1}{n} \sum_i \omega_i g(x_i)$, where $\mathcal{F}^* = \mathcal{F} - f^*$, $\|g\|_n = (\frac{1}{n} \sum_i g(x_i)^2)^{1/2}$. Then

$$\frac{1}{n} \sum_i [\hat{f}(x_i) - f^*(x_i)]^2 \leq \frac{2\sigma}{n} \sum_i \omega_i [\hat{f}(x_i) - f^*(x_i)] \leq 2\sigma lim_{\delta \to \infty} \mathcal{G}(\delta, \mathcal{F}^*). \tag{8}$$

An important observation is $\mathcal{G}(t, \mathcal{F}^*)$ is increasing linearly or no more fatser linearly with respect to a constant C(δ) for any $t > \delta$ if $\mathcal{F}^*$ is star-shaped. Specifically,

$$\frac{\mathcal{G}(\delta, \mathcal{F}^*)}{\delta} \geq \frac{\mathcal{G}(u, \mathcal{F}^*)}{u}, \quad \frac{\mathbb{E}\mathcal{G}(\delta, \mathcal{F}^*)}{\delta} \geq \frac{\mathbb{E}\mathcal{G}(u, \mathcal{F}^*)}{u}, \forall \delta < u. \tag{9}$$

The Nonasymptotic Bound

Let $Z_\delta(r) = \sup_{g \in \mathcal{F}^*, \|g\|_n < \delta} \frac{1}{n} \sum_i r_i g(x_i)$. For fixed δ, $Z_\delta(r)$ is L-Lipschitz continuous as a function of r. The textbook claims that the Lipschitz constant is at most $\frac{\delta}{\sqrt{n}}$. But i can't prove this. Consequently,

$$\mathbb{P}[\mathcal{G}(\delta, \mathcal{F}^*) - \mathbb{E}(\mathcal{G}(\delta, \mathcal{F}^*)) > t] \le \exp(-\frac{nt^2}{\delta^2}), \forall t \tag{10}$$

for any $\|\hat{f} - f^*\|_n = \delta$,

$$\delta^2 \le 2\sigma\mathcal{G}(\delta, \mathcal{F}^*) \le 2\sigma(\mathbb{E}\mathcal{G}(\delta, \mathcal{F}^*) + t) \tag{11}$$

with prob $(1 - \exp(-\frac{nt^2}{\delta^2}))$.

So with prob $(1 - \exp(-\frac{nt^2}{\delta^2}))$,

$$\delta \le 2\sigma(\frac{\mathbb{E}\mathcal{G}(\delta, \mathcal{F}^*)}{\delta} + \frac{t}{\delta}). \tag{12}$$

The Nonasymptotic Bound

Fortunately, for any c, $\exists \delta_c$, $\forall u > \delta_c$, $\frac{\mathbb{E}\mathcal{G}(u,\mathcal{F}^*)}{u} < c$, so $\forall \delta > \delta_c$,

$$\delta \leq 2\sigma(\frac{\mathbb{E}\mathcal{G}(\delta,\mathcal{F}^*)}{\delta} + \frac{t}{\delta}) \leq 2\sigma(c + \frac{t}{\delta}). \tag{13}$$

with prob $(1 - \exp(-\frac{nt^2}{\delta^2}))$. Let $t = c_1\delta$, then

$$\delta \leq 2\sigma(c + c_1). \tag{14}$$

with prob $(1 - \exp(-nc_1^2))$.

Determination of δ_c

The problem remained is the quantity of δ_c. It's related to $\mathcal{F}^*$. If δ_c is very big, we can only bound δ when δ_c is very big, then the bound is meaningless. Specifically we let $c = \frac{u}{2\sigma}$.

We want to know, for $c = \frac{u}{2\sigma}$, which δ_c is valid subject to $\forall u > \delta_c$, $\frac{\mathbb{E}\mathcal{G}(u,\mathcal{F}^*)}{u} < \frac{u}{2\sigma}$. Remark that here 2σ is not important.

Determination of δ_c

For any given $\delta \in (0, \sigma]$, let set $\mathbb{B}(\delta, \mathcal{F}^*) = \{g | g \in \mathcal{F}^*, \|g\|_n < \delta\}$. The minimal $\frac{\delta^2}{4\sigma}$-covering of $\mathbb{B}(\delta, \mathcal{F}^*)$ is finite and is assumed to consist of $\{g^1, \ldots, g^M\}$. By the property of covering set, for any $g \in \mathbb{B}(\delta, \mathcal{F}^*)$ we have

$$
\begin{aligned}
|\frac{1}{n}\sum_i \omega_i g(x_i)| &\leq |\frac{1}{n}\sum_i \omega_i g^j(x_i)| + |\frac{1}{n}\sum_i \omega_i (g(x_i) - g^j(x_i))| \\
&\leq max_j |\frac{1}{n}\sum_i \omega_i g^j(x_i)| + \sqrt{\frac{\sum_i \omega_i^2}{n}}\sqrt{\frac{\sum_i (g(x_i) - g^j(x_i))^2}{n}} \qquad (15) \\
&\leq max_j |\frac{1}{n}\sum_i \omega_i g^j(x_i)| + \sqrt{\frac{\sum_i \omega_i^2}{n}}\frac{\delta^2}{4\sigma}
\end{aligned}
$$

Determination of δ_c

Then

$$
\begin{aligned}
sup_{g\in\mathcal{F}^*}\{\mathbb{E}|\frac{1}{n}\sum_i \omega_i g(x_i)|\} &\leq \mathbb{E}\mathcal{G}(\delta,\mathcal{F}^*) \\
&\leq \mathbb{E}max_j|\frac{1}{n}\sum_i \omega_i g^j(x_i)| + \mathbb{E}\sqrt{\frac{\sum_i \omega_i^2}{n}}\frac{\delta^2}{4\sigma} \qquad (16) \\
&\leq \mathbb{E}max_j|\frac{1}{n}\sum_i \omega_i g^j(x_i)| + \frac{\delta^2}{4\sigma}
\end{aligned}
$$

So $\frac{\mathbb{E}\mathcal{G}(u,\mathcal{F}^*)}{u} < \frac{u}{2\sigma}$ exists if $\mathbb{E}max_j|\frac{1}{n}\sum_i \omega_i g^j(x_i)| < \frac{\delta^2}{4\sigma}$

Determination of δ_c

By definition, for any j, $\frac{1}{n}\sum_i \omega_i g^j(x_i)$ is sub-Gaussian with $L = \frac{\delta^2}{2\sqrt{n}\sigma}$. The therorem 5.22 can be applied.
Let $Z = max_j \frac{1}{n}\sum_i \omega_i g^j(x_i)$,

$$\mathbb{E}Z \leq \frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n\left(t; \mathbb{B}_n\left(\delta; \mathcal{F}^*\right)\right)} dt. \tag{17}$$

So $\frac{\mathbb{E}\mathcal{G}(u,\mathcal{F}^*)}{u} < \frac{u}{2\sigma}$ exists if

$$\frac{16}{\sqrt{n}} \int_{\frac{\delta^2}{4\sigma}}^{\delta} \sqrt{\log N_n\left(t; \mathbb{B}_n\left(\delta; \mathcal{F}^*\right)\right)} dt \leq \frac{\delta^2}{4\sigma} \tag{18}$$

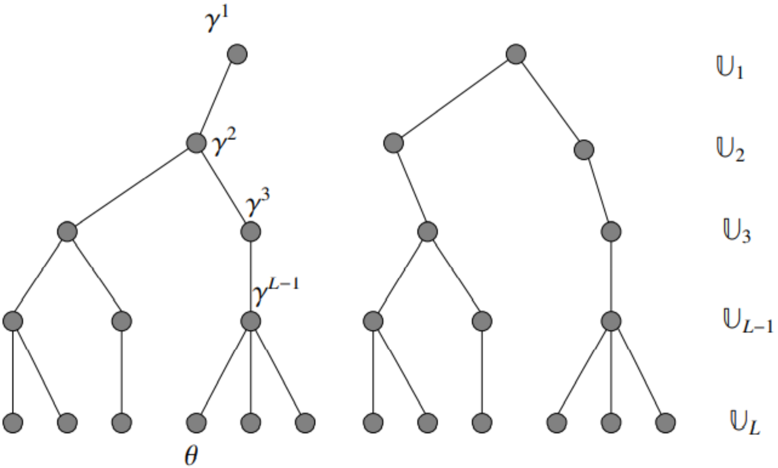
Let $\mathbb{U} = \{g^1, \ldots, g^M\}$let $\mathbb{U}_m$ be a minimal $\epsilon_m = D2^{-m}$ covering set of U in the metric ρ_X, where we allow for any element of $\mathbb{B}(\delta, \mathcal{F}^*)$ to be used. Define the mapping $\pi_m : \mathbb{U} \to \mathbb{U}_m$ via

$$\pi_m(g) = \arg\min_{\beta \in \mathbb{U}_m} \rho_X(g, \beta),$$

Using this notation, we can decompose the random variable X_θ into a sum of increments in terms of an associated sequence $\left(\gamma^1, \ldots, \gamma^L\right)$, where we define $\gamma^L = \theta$ and $\gamma^{m-1} := \pi_{m-1}\left(\gamma^m\right)$ recursively for $m = L, L-1, \ldots, 2$. By construction, we then have the chaining relation

$$X_g - X_{\gamma^1} = \sum_{m=2}^{L} \left(X_{\gamma^m} - X_{\gamma^{m-1}}\right)$$

and hence $\left|X_g - X_{\gamma^1}\right| \leq \sum_{m=2}^{L} max_{\beta \in \mathbb{U}_m} \left|X_\beta - X_{\pi_{m-1}(\beta)}\right|$.


γ^1
γ^2
γ^3
γ^{L-1}
θ
$\mathbb{U}_1$
$\mathbb{U}_2$
$\mathbb{U}_3$
$\mathbb{U}_{L-1}$
$\mathbb{U}_L$

$\left|X_g - X_{\gamma^1}\right| \leq \sum_{m=2}^{L} \max_{\beta\in\mathbb{U}_m} \left|X_\beta - X_{\pi_{m-1}(\beta)}\right|$. From exercise 2.12 we know

$$\mathbb{E}\left[\max_{\gamma,\tilde{\gamma}\in\mathbb{U}_1} |X_\gamma - X_{\tilde{\gamma}}|\right] \leq 2D\sqrt{\log N(D/2)}.$$

Similarly, for each $m = 2, 3, \ldots, L$, the set U_m has $N\left(D2^{-m}\right)$ elements, and, moreover, $\max_{\beta\in\mathbb{U}_m} \rho_X\left(\beta, \pi_{m-1}(\beta)\right) \leq D2^{-(m-1)}$, whence

$$\mathbb{E}\left[\max_{\beta\in\mathbb{U}_m} \left|X_\beta - X_{\pi_{m-1}(\beta)}\right|\right] \leq 2D2^{-(m-1)}\sqrt{\log N\left(D2^{-m}\right)}.$$

So

$$\mathbb{E}\left[\max_{g,\tilde{g}\in\mathbb{U}} |X_g - X_{\tilde{g}}|\right] \leq 4\sum_{m=1}^{L} D2^{-(m-1)}\sqrt{\log N\left(D2^{-m}\right)}.$$

Since the metric entropy $\log N(t)$ is non-increasing in t, we have

$$D2^{-(m-1)}\sqrt{\log N\left(D2^{-m}\right)} \leq 4\int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N(u)}du,$$

Hence $2\mathbb{E}\left[\max_{g,\tilde{g}\in\mathbb{U}}|X_g - X_{\tilde{g}}|\right] \leq 32\int_{\delta/4}^{D}\sqrt{\log N(u)}du$.