Consistency of probability estimation

Chanwoo Lee, July 2, 2020

Lemma 1. *Let $\mathcal{F}(k) = \{f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R} : f(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X} \rangle$ where $\|\boldsymbol{B}\| \leq k\}$. Suppose that a considered feature space is uniformly bounded such that $\|\boldsymbol{X}\| \leq G$. Then the covering number with respect to infinity norm is bounded by,*

$$\log N_\infty(\epsilon, \mathcal{F}(k)) \leq \mathcal{O}\left(d_1 d_2 \log\left(\frac{Gk}{\epsilon}\right)\right).$$

Proof. Define $\mathcal{B}_k = \{\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{B}\| \leq k\}$. By the definition of infinity norm in the function space, we have

$$\|f_{\boldsymbol{B}} - f_{\boldsymbol{B}'}\|_\infty = \|\langle \boldsymbol{B}, \cdot \rangle - \langle \boldsymbol{B}, \cdot \rangle\|_\infty = \sup_{\|\boldsymbol{X}\| \leq G} \left| \langle \boldsymbol{B} - \boldsymbol{B}', \boldsymbol{X} \rangle \right| = G\|\boldsymbol{B} - \boldsymbol{B}'\|_2$$

Therefore, the metric space $(\mathcal{F}(k), \|\cdot\|_\infty)$ is isomorphic to the metric space $(\mathcal{B}_k, G\|\cdot\|_2)$. We can check the covering number $N_2(\epsilon, \mathcal{B}_k) \leq \mathcal{O}\left(\left(\frac{k}{\epsilon}\right)^{d_1 d_2}\right)$ which proves the lemma. ☐

Remark 1. When we restrict the considered linear function class to

$$\mathcal{F}_r(M) = \{f : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R} : f(\boldsymbol{X}) = \langle \boldsymbol{B}, \boldsymbol{X} \rangle \text{ s.t. } \operatorname{rank}(\boldsymbol{B}) \leq r, \lambda_1(\boldsymbol{B}) \leq M\}.$$

rM = O(1) as d —> infty.

rM = O(g(d))

One can verify that

worst case: g(d)=d1*d2 —> required sample size n >= d1*d2 in order to achieve consistency

best case: g(d) = O(1) > required sample size n ~ O(1)

$$\log N_\infty(\epsilon, \mathcal{F}_r(M)) \leq \mathcal{O}\left(d_1 d_2 \log\left(\frac{rGM}{\epsilon}\right)\right),$$

from the inclusion $\mathcal{F}_r(M) \subset \mathcal{F}(rM)$.

Lemma 2. *Let $k > 0$ be a given constant. If $\frac{1}{Ke} > L > 0$, we have*

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} d\omega \leq \mathcal{O}\left(\sqrt{L \log\left(\frac{k}{\sqrt{L}}\right)}\right).$$

Proof.

$$\begin{aligned} \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega &= k\left[\omega \sqrt{\log\left(\frac{1}{\omega}\right)}\right]_{\mathcal{O}(L/k)}^{\mathcal{O}(\sqrt{L}/k)} \\ &= \mathcal{O}\left(\sqrt{L \log\left(\frac{k}{\sqrt{L}}\right)}\right) \end{aligned} \tag{1}$$

The first equality in (1) is from changing variable. Notice that

$$\int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \frac{1}{2\sqrt{\log\left(\frac{k}{\omega}\right)}} d\omega \geq \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \sqrt{\log\left(\frac{k}{\omega}\right)} - \mathcal{O}(1) d\omega, \tag{2}$$

from the condition on L. Combining Equation (1) and Equation (2) completes the proof. ☐

Lemma 3. $\frac{1}{\sqrt{L}}\sqrt{\log\left(\frac{k}{\sqrt{L}}\right)} \leq \sqrt{n}$ *holds if* $L \leq \frac{log(n)+2log(k)}{n}$.

Proof. Suppose $L \leq \frac{log(n)+2log(k)}{n}$. By plugging in, we have

$$\begin{aligned}\frac{1}{\sqrt{L}}\sqrt{\log\left(\frac{k}{\sqrt{L}}\right)} &\leq \sqrt{\frac{n}{log(n)+2log(k)}\left(\log(n)+2\log(k)-\log\log(nk^2)\right)}\\ &\leq \sqrt{n}.\end{aligned}$$

□

Let $\bar{f}_\pi$ be a Bayes rule. In addition, let $e_V(f,\bar{f}_\pi) = \mathbb{E}\{V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)\}$ with $V(f,\boldsymbol{X},y) = S(y)L\{yf(\boldsymbol{X})\}$.

Based on function class $\mathcal{F}_r(M)$, we have the following theorem.

Theorem 0.1. *Assume that*

1. *For some positive sequence such that* $s_n \to 0$ *as* $n \to \infty$, *there exists* $f_\pi^* \in \mathcal{F}_r(M)$ *such that* $e_V(f_\pi^*, \bar{f}_\pi) \leq s_n$.
2. *There exists* $0 \leq \alpha < \infty$ *and* $a_1 > 0$ *such that, for any sufficiently small* $\delta > 0$,

$$\sup_{\{f\in\mathcal{F}:e_V(f,\bar{f}_\pi)\leq\delta\}} \|sign(f) - sign(\bar{f}_\pi)\|_1 \leq a_1\delta^\alpha,$$

3. *Considered feature space is uniformly bounded such that there exists* $0 < G < \infty$ *satisfying* $\|\boldsymbol{X}\| \leq G$

Then, for the estimator $\hat{p}$ *obtained from our algorithm, there exists a constant* a_2 *such that*

$$\mathbb{P}\left\{\|\hat{p}-p\|_1 \geq \frac{1}{2m} + \frac{1}{2}a_1(m+1)\delta_n^{2\alpha}\right\} \leq 15\exp\{-a_2 n(\lambda J_\pi^*)\},$$

provided that $\lambda^{-1} \geq \frac{rGJ_\pi^*}{2\delta_n^2}$ *where* $J_\pi^* = \max(J(f_\pi^*),1)$ *and* $\delta_n = \max\left(\mathcal{O}\left(\frac{\log(n)+2\log(rGM)}{n}\right), s_n\right)$.

Proof. We apply Theorem 3 in [2] to our case. First, notice that trucation on the loss function V is not needed by third assumption:

$$\|yf(\boldsymbol{X})\| = \|\langle B,\boldsymbol{X}\rangle\| \leq \|B\|\|\boldsymbol{X}\| \leq rGM,$$

which implies uniformly boundness of V. Let V be bounded by T. For the second equation of Assumption 2 in [2],

$$\begin{aligned}\text{var}\{V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)\} &\leq \mathbb{E}|V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)|^2\\ &\leq T\mathbb{E}|V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)|\\ &= Te_V(f,\bar{f}_\pi).\end{aligned}$$

Therefore, β in [2] can be replaced by 1 from the following inequality.

$$\sup_{\{f\in\mathcal{F}:e_V(f,\bar{f}_\pi)\leq\delta\}} \text{var}\{V(f,\boldsymbol{X},y) - V(\bar{f}_\pi,\boldsymbol{X},y)\} \leq \sup_{\{f\in\mathcal{F}:e_V(f,\bar{f}_\pi)\leq\delta\}} Te_V(f,\bar{f}_\pi) \leq T\delta$$

Now we check Assumption 3 in [2]. Notice that

$$H_2(\epsilon, \mathcal{F}^V(k)) \leq H_2(\epsilon, \mathcal{F}(k)) \leq \log N_\infty(\epsilon, \mathcal{F}(k)), \tag{3}$$

because for functions f_ℓ and f_u, $\|V(f_\ell, \cdot) - V(f_u, \cdot)\|_2 \leq \|f_\ell - f_u\|_2$. The last inequality in (3) is from Lemma 9.22 in [1]. From Lemma 1, we have

$$\log N_\infty(\epsilon, \mathcal{F}(k)) \leq \mathcal{O}\left(d_1 d_2 \log\left(\frac{Gk}{\epsilon}\right)\right) \approx \mathcal{O}\left(\log\left(\frac{Gk}{\epsilon}\right)\right).$$

Therefore, we have the following equation from Lemma 2.

$$\phi(\epsilon, k) \approx \int_{\mathcal{O}(L)}^{\mathcal{O}(\sqrt{L})} \log\left(\frac{kG}{\omega}\right) d\omega / L \lessapprox \mathcal{O}\left(\left(\log\left(\frac{kG}{L}\right)/L\right)^{1/2}\right),$$

where $L = \min\{\epsilon^2 + \lambda(k/2-1)H_\pi^*, 1\}$. Solving Assumption 3 in [2] gives us $\epsilon_n^2 = \mathcal{O}\left(\frac{\log(n)+2\log(rGM)}{n}\right)$ by Lemma 3 when $\epsilon_n^2 \geq \lambda r G J_\pi^*$. Plugging each variable into Theorem 3 proves the theorem. Notice that condition of λ is replaced because $\{\epsilon_n^2 \geq \lambda r G J_\pi^*\} \subset \{\epsilon_n^2 \geq 2\lambda J_\pi^*\}$ when $rG \geq 2$. □

References

[1] Michael R Kosorok. *Introduction to empirical processes and semiparametric inference*. Springer Science & Business Media, 2007.

[2] Junhui Wang, Xiaotong Shen, and Yufeng Liu. Probability estimation for large-margin classifiers. *Biometrika*, 95(1):149–167, March 2008.