Spring 2018: STA 6448
Advanced Probability and Inference II
Lecture 25

Yun Yang

- Non-parametric least squares

Example: Kernel ridge regression, continued

- Recall the KRR estimator

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

- Let $\mathcal{K} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ denote the reproducing kernel associated with the RKHS $\mathbb{H}$.

Theorem (Representer theorem)

Any solution $\widehat{f}$ of the KRR optimization problem takes the form

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i \, \mathcal{K}(\cdot, x_i).$$

Example: Kernel ridge regression, continued

- Define the empirical kernel matrix $K \in \mathbb{R}^{n \times n}$, with $K_{ij} = n^{-1}\,\mathcal{K}(x_i, x_j)$, and recall

$$\widehat{f}(\cdot) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \widehat{\alpha}_i\,\mathcal{K}(\cdot, x_i).$$

- Then, we can write

$$(\widehat{f}(x_1), \ldots, \widehat{f}(x_n))^T = \sqrt{n}\,K\,\widehat{\alpha},$$

where $\widehat{\alpha} = (\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n)^T$.

- Solving the KRR optimization problem is inequivalent to solving the following quadratic programming

$$\widehat{\alpha} \in \operatorname*{argmin}_{\alpha \in \mathbb{R}^n} \left\{ \frac{1}{2n} \|y - \sqrt{n}\,K\,\alpha\|_2^2 + \lambda_n \underbrace{\alpha^T K \alpha}_{\|f\|_{\mathbb{H}}^2} \right\}.$$

Example: Convex regression

- Now suppose $f^* : \mathcal{C} \to \mathbb{R}$ is known to be a convex function over its domain $\mathcal{C}$, where $\mathcal{C}$ is some convex and open subset of $\mathbb{R}^d$.
- It is natural to consider the least-squares estimator with a convexity constraint,

$$\widehat{f} \in \underset{f \text{ is convex}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\}.$$

- Although this optimization problem is infinite-dimensional, we can convert it to an equivalent finite-dimensional problem.
- The convexity constraint implies there exist sub-gradient vectors $\{\widetilde{z}_i\}_{i=1}^n$, such that for all $i = 1, \ldots, n$,

$$f(x) \geq f(x_i) + \langle \widetilde{z}_i, x - x_i \rangle \quad \text{for all } x \in \mathcal{C}.$$

Example: Convex regression

- Since the cost function depends only on the values $\widetilde{y}_i = f(x_i)$, the optimum does not depend on the function behavior elsewhere.
- It suffices to solve the optimization problem

$$\min_{\{(\widetilde{y}_i, \widetilde{z}_i)\}_{i=1}^n} \frac{1}{2n} \sum_{i=1}^{n} (y_i - \widetilde{y}_i)^2$$

such that $\quad \widetilde{y}_j \geq \widetilde{y}_i + \langle \widetilde{z}_i, x_j - x_i \rangle \quad$ for all $i, j = 1, \ldots, n$.

- An optimal solution $\{(\widehat{y}_i, \widehat{z}_i)\}_{i=1}^n$ can be used to define an estimate $\widehat{f} : \mathcal{C} \to \mathbb{R}$ via

$$\widehat{f}(x) = \max_{i=1,\ldots,n} \left\{ \widehat{y}_i + \langle \widehat{z}_i, x - x_i \rangle \right\}.$$

- $\widehat{f}$ is convex, and by the feasibility of the solution $\{(\widehat{y}_i, \widehat{z}_i)\}_{i=1}^n$, we are guaranteed that $\widehat{f}(x_i) = \widehat{y}_i$.

Statistical error bounds

$$\widehat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \right\},$$

- Question: how well the non-parametric least squares estimate $\widehat{f}$ approximates the true regression function f^*.
- We will bound the error $\|\widehat{f} - f^*\|_n$ as measured in the $L^2(\mathbb{P}_n)$-norm.
- The difficulty of estimating the function f^* should depend on the complexity of the function class $\mathcal{F}$.
- Define the f^*-shifted version of the function class $\mathcal{F}$,

$$\mathcal{F}^* := \left\{ f - f^* \,\middle|\, f \in \mathcal{F} \right\}.$$

Localized form of Gaussian complexity

- We define a complexity measure of $\mathcal{F}$, locally in a neighborhood around the true regression function f^*.

Definition

Local Gaussian complexity For a given radius $\delta > 0$, the local Gaussian complexity around f^* at scale δ is given by

$$\mathcal{G}_n(\delta;\, \mathcal{F}^*) := \mathbb{E}_w \Bigg[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \leq \delta}} \left| \frac{1}{n} \sum_{i=1}^{n} w_i\, g(x_i) \right| \Bigg],$$

where $\{w_i\}_{i=1}^n$ are i.i.d. $\mathcal{N}(0,\, 1)$ variables.

- A central object in our analysis is the set of positive δ that satisfy the critical inequality

$$\frac{\mathcal{G}_n(\delta;\, \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma}.$$

Critical radius

We call a set $\mathcal{H}$ star-shaped if for any $h \in \mathcal{H}$ and $\alpha \in [0, 1]$, the rescaled function$\alpha\, h$ also belongs to $\mathcal{H}$.

Lemma

If $\mathcal{F}^* := \{f - f^* \mid f \in \mathcal{F}\}$ is star-shaped, then $\frac{\mathcal{G}_n(\delta;\, \mathcal{F}^*)}{\delta}$ is a non-increasing function of $\delta > 0$.

Definition

When $\mathcal{F}^* := \{f - f^* \mid f \in \mathcal{F}\}$ is star-shaped, we define the critical radius $\delta_n > 0$ as the smallest solution of

$$\frac{\mathcal{G}_n(\delta;\, \mathcal{F}^*)}{\delta} \leq \frac{\delta}{2\sigma}.$$

Heuristic illustration

By the optimality of $\widehat{\theta}$ and the feasibility of θ^*, we have

$$\frac{1}{2n}\sum_{i=1}^{n}(y_i-\widehat{f}(x_i))^2 \leq \frac{1}{2n}\sum_{i=1}^{n}(y_i-f^*(x_i))^2.$$

Recalling $y_i = f^*(x_i) + \sigma w_i$, some simple algebra leads to

$$\text{(basic inequality):}\quad \frac{1}{2}\|\widehat{f}-f^*\|_n^2 \leq \frac{\sigma}{n}\sum_{i=1}^{n} w_i\left(\widehat{f}(x_i)-f^*(x_i)\right).$$

Let $\delta^2 = \|\widehat{f}-f^*\|_n^2$, then we have (this step is heuristic!)

$$\frac{\delta^2}{2} \leq \sigma\,\mathcal{G}_n(\delta;\,\mathcal{F}^*), \quad \text{or} \quad \frac{\mathcal{G}_n(\delta;\,\mathcal{F}^*)}{\delta} \geq \frac{\delta}{2\sigma}.$$

By definition of δ_n, this implies $\delta \leq \delta_n$.

Formal statement

$$\mathcal{G}_n(\delta;\, \mathcal{F}^*) := \mathbb{E}_w\Bigg[\sup_{\substack{g \in \mathcal{F}^* \\ \|g\|_n \le \delta}} \Big| \frac{1}{n} \sum_{i=1}^{n} w_i\, g(x_i) \Big| \Bigg],$$

$$\text{(critical radius)} \qquad \frac{\mathcal{G}_n(\delta_n;\, \mathcal{F}^*)}{\delta_n} \le \frac{\delta_n}{2\sigma}.$$

Theorem (Non-parametric least squares error bound)

Suppose that the shifted function class $\mathcal{F}^*$ is star-shaped. Then there are universal positive constants (c_0, c_1, c_2) such that for any $t \ge \delta_n$, the non-parametric least squares estimate $\widehat{f}_n$ satisfies

$$\mathbb{P}\big[\|\widehat{f}_n - f^*\|_n^2 \ge c_0\, t\, \delta_n\big] \le c_1\, e^{-\frac{c_2\, n\, t\, \delta_n}{\sigma^2}}.$$

This theorem implies

$$\|\widehat{f}_n - f^*\|_n^2 = \mathcal{O}_p(\delta_n^2).$$

Bounds via metric entropy

Recall that the localized Gaussian complexity corresponds to expected absolute maximum of a Gaussian process.

Define $\mathbb{B}_n(\delta; \mathcal{F}^*) = \{f \in \mathcal{F}^* : \|f\|_n \leq \delta\}$, and $\mathcal{N}_n(t, \mathbb{B}_n(\delta; \mathcal{F}^*))$ denote the t-covering number of $\mathbb{B}_n(\delta; \mathcal{F}^*)$ in the $\|\cdot\|_n$ norm.

Corollary

The critical radius δ_n is upper bounded by any $\delta \in (0, \sigma]$ such that

$$\frac{32}{\sqrt{n}} \int_{\frac{\delta^2}{2\sigma}}^{\delta} \sqrt{\log \mathcal{N}_n(t, \mathbb{B}_n(\delta; \mathcal{F}^*))}\, dt \leq \frac{\delta^2}{\sigma}.$$

Proof: Apply Dudley's entropy integral bound.

Example: Bound for linear regression

- Consider the standard linear model $y_i = \langle x_i, \theta \rangle + w_i$, where $\theta \in \mathbb{R}^d$.
- The usual least-squares estimate corresponds to optimizing over the function class

$$\mathcal{F}_{\mathrm{lin}} = \{ f_\theta(\cdot) = \langle \cdot, \theta \rangle : \theta \in \mathbb{R}^d \}.$$

- Let $X \in \mathbb{R}^{n \times d}$ denote the design matrix. Then $\| f_\theta \|_n = \frac{\|X\theta\|_2}{\sqrt{n}}$.
- Therefore, we have ($r = \mathrm{rank}(X)$),

$$\log \mathcal{N}_n(t, \mathbb{B}_n(\delta; \mathcal{F}^*)) \leq r \log \left(1 + \frac{c\,\delta}{t} \right),$$

$$\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log \mathcal{N}_n(t, \mathbb{B}_n(\delta; \mathcal{F}^*))}\, dt \leq c' \,\delta \sqrt{\frac{r}{n}}.$$

- Finally, we reach that

$$\| \widehat{f}_n - f^* \|_n^2 \leq C\, \sigma^2 \frac{\mathrm{rank}(X)}{n} \qquad \text{w.h.p.}$$

Example: Bounds for Lipschitz functions

- Consider the function class

$$\mathcal{F}_{\text{Lip}}(L) = \{f : [0,1] \to \mathbb{R} : f(0) = 0,\, f \text{ is } L\text{-Lipschitz.}\}.$$

- In our previous lecture, we showed

$$\log \mathcal{N}(t,\, \mathcal{F}_{\text{Lip}}(L),\, \|\cdot\|_\infty) \le \frac{2L}{t}.$$

- Therefore, we have

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log \mathcal{N}_n(t,\, \mathbb{B}_n(\delta;\, \mathcal{F}^*))}\, dt \\
&\le \frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log \mathcal{N}(t,\, \mathcal{F}_{\text{Lip}}(L),\, \|\cdot\|_\infty)}\, dt \le \frac{c}{\sqrt{n}} \sqrt{L\delta}.
\end{aligned}$$

- Finally, this implies that

$$\|\widehat{f}_n - f^*\|_n^2 \le C \left(\frac{L\sigma^2}{n}\right)^{2/3} \qquad \text{w.h.p.}$$

Example: Bounds for convex regression

- Consider the class of convex 1-Lipschitz functions,

  $\mathcal{F}_{\text{conv}} = \{f : [0,1] \to \mathbb{R} : f(0) = 0,\ f$ is L-Lipschitz and convex.$\}$.

- It can be shown that

$$\log \mathcal{N}(t,\, \mathcal{F}_{\text{conv}},\, \|\cdot\|_\infty) \leq \frac{C}{\sqrt{t}}.$$

- Therefore, we have

$$\frac{1}{\sqrt{n}} \int_0^\delta \sqrt{\log \mathcal{N}_n(t,\, \mathbb{B}_n(\delta;\, \mathcal{F}^*))}\, dt$$
$$\leq \frac{1}{\sqrt{n}} \int_0^\delta \frac{\sqrt{C}}{t^{1/4}}\, dt \leq \frac{c}{\sqrt{n}}\, \delta^{3/4}.$$

- Finally, this implies that

$$\|\widehat{f}_n - f^*\|_n^2 \leq C \left(\frac{\sigma^2}{n}\right)^{4/5} \qquad \text{w.h.p.}$$