Lecture 26. Nonparametric Regression

Recap so far:

- Reproducing Kernel Hilbert space perspective
- Bias-variance decomposition to analyze error of Kernel ridge regression
- Power law generalization: $\mu_j = j^{-\alpha}$ $\Rightarrow$ generalization error of $n^{\frac{1}{\alpha}-1}$ $(\alpha > 1)$

"noise removal", not really generalization

This time:

- Look at generalization to test distⁿ
- Mercer's theorem: important analytic tool
  - ↳ Useful for computing μ_j's
  - ↳ Helps draw connection b/t random features and Kernel regression

Notation:

$x_1, \dots, x_n \sim \rho$

$y_i = \langle \beta^*, x_i \rangle + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2)$

$\hookrightarrow f^*(x_i) + \xi_i$

$S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$

$S^* = \mathbb{E}_{x\sim p}[xx^T]$

$L^*(\beta)$: excess risk of population level

$\Rightarrow L^*(\beta) = (\beta-\beta^*)^T S^* (\beta-\beta^*)$

$L(\beta)$: excess risk on training samples

$= \frac{1}{n}\sum_{i=1}^{n} \langle \beta-\beta^*, x_i\rangle^2 = (\beta-\beta^*)^T S (\beta-\beta^*)$

Ridge regression estimator:

$$\hat{\beta}_\lambda = \frac{1}{n}\left(\frac{\lambda}{n} I + S\right)^{-1} X^T (X\beta^* + \xi)$$

$$= \left(\lambda I + X^T X\right)^{-1} X^T y$$

$$\mathbb{E}_\xi[\hat{\beta}_\lambda] = \left(\frac{\lambda}{n} I + S\right)^{-1} S\beta^*$$

$$\delta = \frac{\lambda}{n}\left(\frac{\lambda}{n} I + S\right)^{-1}\beta^*$$

Bias: $\delta^T S^* \delta$, $\delta = \mathbb{E}_\xi[\hat{\beta}_\lambda] - \beta^*$

$$\text{Bias}^2 = \left(\frac{\lambda}{n}\right)^2 (\beta^*)^T \left(\frac{1}{n} X^T X + S\right)^{-1} S^* \left(\frac{\lambda}{n} X^T X + S\right)^{-1} \beta^*$$

Contrast to before's
Bias^2 was w.r.t. train set $\Rightarrow S$ instead of S^*

old formula:

$$\left(\frac{\lambda}{n}\right)^2 (\beta^*)^T \left(\frac{1}{n} X^T X + S\right)^{-1} S \left(\frac{\lambda}{n} X^T X + S\right)^{-1} \beta^*$$

$\lambda_n = \frac{\lambda}{n}$

$\mathbb{E}[\text{Bias}^2]$ under Gaussian prior $\mathbb{E}[\beta^* (\beta^*)^T] = \rho^2 I$

$$\mathbb{E}[\text{Bias}^2] = \lambda_n^2 \rho^2 \operatorname{trace}\left(\left(\lambda_n I + S\right)^{-1} S^* \left(\lambda_n I + S\right)^{-1} \right)$$

$$\text{Variance} = \frac{\sigma^2}{n} \operatorname{trace}\left((S^*)^{1/2} \left(\lambda_n I + S\right)^{-1} S \left(\lambda_n I + S\right)^{-1} (S^*)^{1/2} \right)$$

old formula: $\frac{\sigma^2}{n} \operatorname{trace}\left(\left(\lambda_n I + S\right)^{-1} S^2 \left(\lambda_n I + S\right)^{-1} \right)$

Good generalization want Bias^2 + Variance to be small

Issue: S instead of S^* ← generally assume is nice

For bias $(\lambda_n I + S)^{-1} \approx (\lambda_n I + S^*)^{-1}$

$$(\lambda_n I + S)^{-1} \preceq 2(\lambda_n I + S^*)^{-1}$$ ← Want this to be true.

$\Rightarrow$ can call S w/ S^* losing only a factor of 2

For variance:

$$(\lambda_n I + S)^{-1} S (\lambda_n I + S)^{-1} \preceq (\lambda_n I + S)^{-1}$$

since $S(\lambda_n I + S)^{-1} \preceq I$

$$\text{Variance} \le \frac{\sigma^2}{n} \operatorname{trace}\left((S^*)^{1/2} (\lambda_n I + S)^{-1} (S^*)^{1/2} \right)$$

Once again, as long as $(\lambda_n I + S)^{-1} \preceq 2(\lambda_n I + S^*)^{-1}$, lose at most a factor of 2 moving to S^*.

$$2(\lambda_n I + S) \succeq \lambda_n I + S^*$$

$$\Rightarrow S^* \preceq \lambda_n I + 2S$$

$\lambda_n I$: takes care of small eigenvalues

$2S$: Need concentration in large eigenvalues

Proposition. Suppose $S^* \le \lambda_n I + 2S$.

Let $\mu_j^* = j^{th}$ largest eigenvalue of S^*.

Then, generalization error satisfies

$$\mathbb{E}_{\beta^*}\left[\mathbb{E}_{\xi}\left[L^*(\beta_\lambda)\right] - L^*(\beta^*)\right]$$

$$\le \boxed{4\lambda_n \rho^2 \sum_j \frac{\lambda_n \mu_j^*}{(\lambda_n + \mu_j^*)^2} + \frac{2\sigma^2}{n} \sum_j \frac{\mu_j^*}{\lambda_n + \mu_j^*}}$$

$$\le \left(4\lambda_n \rho^2 + \frac{2\sigma^2}{n}\right) \sum_j \min\left(1, \frac{\mu_j^*}{\lambda_n}\right).$$

$$\frac{\lambda_n}{\lambda_n + \mu_j^*} \le 1 \Rightarrow \frac{\lambda_n \mu_j^*}{(\lambda_n + \mu_j^*)^2} \le \frac{\mu_j^*}{\lambda_n + \mu_j^*} \le \min\left(1, \frac{\mu_j^*}{\lambda_n}\right)$$

$$\mathbb{E}[\text{Bias}^2] = \lambda_n^2 \rho^2 \operatorname{trace}\left((\lambda_n I + S)^{-1} S^* (\lambda_n I + S)^{-1}\right)$$

$$\text{Variance} = \frac{\sigma^2}{n} \operatorname{trace}\left((S^*)^{1/2} (\lambda_n I + S)^{-1} S (\lambda_n I + S)^{-1} (S^*)^{1/2}\right)$$

$$\hookrightarrow \quad \mathbb{E}[\text{Bias}^2] \le 4\lambda_n^2 \rho^2 \, \text{trace}\left((\lambda_n I + S^*)^{-1} S^* (\lambda_n I + S^*)^{-1} \right)$$

$$= \sum_j \frac{\mu_j^*}{(\lambda_n + \mu_j^*)^2}$$

$$\mu_j = j^{-\alpha}$$

Corollary: If we plug in $\mu_j^* \asymp j^{-\alpha}$, (S) get same $n^{\frac{1}{\alpha}-1}$ rate as before.

Upshot: Really do just need to care about when $S^* \preceq \lambda_n I + 2S$.

$$S \succeq \frac{1}{2}(S^* + \lambda_n I) - \lambda_n I$$

$$(S^* + \lambda_n I)^{-1/2} S (S^* + \lambda_n I)^{-1/2} \succeq \frac{1}{2} I - \lambda_n (S^* + \lambda_n I)^{-1}$$

simple

looks like a covariance matrix Σ

find something like Σ^*

S: sample covariance matrix of X;

Σ: sample covariance matrix of $(S^2+\lambda_n I)^{-1/2} X$;

$$\Sigma^*: (S^2+\lambda_n I)^{-1/2} S^2 (S^2+\lambda_n I)^{-1/2}$$

$$= S^2 (S^2+\lambda_n I)^{-1}$$

$$\frac{1}{2} I = I - \frac{1}{2} I$$

$$\downarrow$$

$$I = (S^2+\lambda_n I)(S^2+\lambda_n I)^{-1}$$

$$(S^2+\lambda_n I)(S^2+\lambda_n I)^{-1} - \lambda_n I (S^2+\lambda_n I)^{-1} - \frac{1}{2} I$$

$$= S^2 (S^2+\lambda_n I)^{-1} - \frac{1}{2} I = \Sigma^* - \frac{1}{2} I$$

Upshot Condition is equivalent to saying that

$$\hat{\Sigma} \succeq \Sigma^* - \frac{1}{2} I$$

Suffices to show that $\|\Sigma^* - \hat{\Sigma}\|_{op} \leq \frac{1}{2}$.

Theorem (Koltchinskii & Lounici)

If X_i are Gaussian w/ covariance matrix Σ^*, then

$$\|\hat{\Sigma} - \Sigma^*\| \leq \Theta\left(\max\left(\frac{\mathrm{tr}(\Sigma^*)}{n}, \sqrt{\frac{\mathrm{tr}(\Sigma^*)\|\Sigma^*\|}{n}} \right) \right)$$

$\mathrm{tr}(\Sigma^*) \leq p \cdot \|\Sigma^*\|$ $\quad p = \infty$

$\Rightarrow \sqrt{\frac{p}{n}\|\Sigma^*\|}$

Recall $\Sigma^* = S^*(\lambda_n I + S^*)^{-1}$

$\Rightarrow \|\Sigma^*\| \leq 1$

$\text{trace}(\Sigma^*) = \sum_j \frac{\mu_j^*}{\lambda_n + \mu_j^*}$

$\Rightarrow$ As long as $\boxed{\sum_j \frac{\mu_j^*}{\lambda_n + \mu_j^*} \ll n,}$

$\|\Sigma^* - \hat{\Sigma}\|_{op} \leq \frac{1}{2}$

$\Rightarrow S^* \leq 2S + \lambda_n I$

If $\mu_j^* = j^{-\alpha}$, condition works out to $\lambda_n \Rightarrow \frac{\alpha}{\alpha - 1} n^{-\alpha}$.

Before, optimal $\lambda_n = \frac{\sigma^2}{n\rho^2}$ $\frac{1}{n^\alpha}$

$\frac{1}{n}$

Upshot: Generally fine in terms of generalization, UNLESS σ^2 is really small.

If we kind propagate lower bound

on λ_n, result by bound on generalization error is

$$\mathbb{E}[\text{Error}] \leq \mathcal{O}\left(\frac{\alpha}{\alpha - 1} \cdot \max\left(n^{\frac{1}{\alpha}-1}(\sigma^2)^{\frac{\alpha-1}{\alpha}}(\rho^2)^{\frac{1}{\alpha}}, \; n^{1-\alpha}\rho^2\right)\right)$$

what we had before

new term that accounts for generalization

Mercer's theorem.

Motivation: Eigenvalues of S, K seem important for understanding generalization.

$$K = \begin{bmatrix} k(x_1, x_1) & \cdots & k(x_1, x_n) \\ & \ddots & \\ & \cdots & \end{bmatrix}$$

What about eigenvalues of K?

Define operator T_K

arbitrary finite, positive measure ν

$$T_K f(x) = \int_X K(x,s) f(s)\, d\nu(s)$$

"$(T_K f)_x = \sum_s K_{xs} f_s$"

Theorem (Mercer).

Suppose $K: X \times X \to \mathbb{R}$ is a positive definite kernel, K is continuous, and X is compact w/ finite positive Borel measure ν supported on all of X.

Then, there is an orthonormal basis

$($. $)$

(wrt $\langle f, g \rangle_\nu$) $\int f(x) g(x) \, d\nu(x)$

of eigenfunctions $e_1, e_2, \ldots$ of T_K, w/ corresponding eigenvalues $\mu_1, \mu_2, \ldots$, s.t.

$$K(x,y) = \sum_{\ell=1}^{\infty} \mu_\ell e_\ell(x) e_\ell(y).$$

observations: 3 inner products

$\langle \cdot, \cdot \rangle_{\mathcal{H}}$, $\langle \cdot, \cdot \rangle_{\ell_2}$ (function error)

$\langle \cdot, \cdot \rangle_{L^2(\nu)}$ ← ν arbitrary useful for computation

---

Consequences:

- Any f can be represented as

$$f(x) = \sum_{\ell} c_\ell e_\ell(x)$$

$$c_\ell = \int_X f(x) e_\ell(x) \, d\nu(x)$$

$$= \langle f, e_\ell \rangle_{L^2(\nu)}$$

- $\|f\|_H^2 = \sum_{\ell=1}^{\infty} \frac{c_\ell^2}{\mu_\ell}$ $\left(\text{vs. } \|f\|_{L^2(\nu)}^2 = \sum_{\ell=1}^{\infty} c_\ell^2 \right)$

INDEPENDENT OF ν

- If we take $\nu = p$, $\langle \cdot, \cdot \rangle_{\ell_2} = \langle \cdot, \cdot \rangle_{L^2(p)}$