Non-parametric II

Reminder: Pset 4 due next Tues.

Re-cap last time.

	Parameter view	Function view
Norm	$\|\beta\|_2^2$	$f^T K^{-1} f$
Error	$(\beta - \beta^*)^T \Sigma (\beta - \beta^*)$	$\frac{1}{n}\|f - f^*\|_2^2$

Kernel ridge regression:

$$f = K(\lambda I + K)^{-1} y$$

$$f(x) = K_x^T (\lambda I + K)^{-1} y$$

Define RKHS:

$$\mathcal{H} = \left\{ f \;\middle|\; \sup_T \underbrace{f[T]^T K[T]^{-1} f[T]}_{\|f\|_{\mathcal{H}}^2} < \infty \right\}$$

Two key results:

1. $\|f\|_{K[T]}$ monotonic in T
2. Isotropic extensions can make $\|f\|_{\mathcal{H}} = \|f\|_{K[T]}$.

This time: Generalization

1. Bias-variance on train set (deal w/ observation noise)

2. Generalize from train to test set

→ Generalization determined by eigenvalues of K (or of S)

↳ If $\mu_j = n \cdot j^{-\alpha}$ (for K) or $j^{-\alpha}$ (for S)

"power law decay"

then error is also a power law.

$\mu_j(S) = \frac{1}{n} \cdot \mu_j(K)$

$\frac{1}{n} X^T X$ ↓ XX^T

$x_1, \dots, x_n \sim P$

$y_i = f^*(x_i) + \varepsilon_i$ for $\varepsilon_i \sim N(0, \sigma^2)$

Kernel ridge regression:

$$\hat{f}_\lambda = K(\lambda I + K)^{-1} y$$ ← depends on ε

Error: $\mathbb{E}_\varepsilon \left[\frac{1}{n} \| \hat{f}_\lambda - f^* \|_2^2 \right]$

Bias: $\frac{1}{n} \| \mathbb{E}_\varepsilon[\hat{f}_\lambda] - f^* \|_2^2$

Var: $\frac{1}{n} \sum_{i=1}^{n} \mathrm{Var}_\varepsilon[f(x_i)]$

Bias. $\mathbb{E}_\varepsilon[\hat{f}_\lambda] = K(\lambda I + K)^{-1}\mathbb{E}_\varepsilon[y]$

$(\lambda I + K)(\lambda I + K)^{-1} f^* = K(\lambda I + K)^{-1} f^*$

$\frac{1}{n}\| f^* - K(\lambda I + K)^{-1} f^* \|_2^2$

$= \frac{1}{n}\| \lambda (\lambda I + K)^{-1} f^* \|_2^2$

$= \frac{\lambda^2}{n}\| (\lambda I + K)^{-1} f^* \|_2^2$

Further bound.

$f^T \underbrace{(\lambda I + K)^{-1}}_{\leq \frac{1}{\lambda}} \underbrace{(\lambda I + K)^{-1}}_{\leq K^{-1}} f$

$\leq \frac{1}{\lambda} f^T K^{-1} f = \frac{1}{\lambda}\|f\|_{\mathcal{H}}^2$

$\frac{\lambda}{n}\|f^*\|_{\mathcal{H}}^2$ ← often use in parametric case, but we'll see that it's too loose in nonparametric case.

Variance.

$\hat{f}_\lambda = K(\lambda I + K)^{-1} y$ ← $f^* + \varepsilon$

$\frac{1}{n}\mathbb{E}_\varepsilon\left[\| K(\lambda I + K)^{-1}\varepsilon \|_2^2\right]$

$\frac{1}{n}\sum_{i=1}^n \mathrm{Var}_\varepsilon[\hat{f}_\lambda(x_i)]$

$$\hookrightarrow = \frac{1}{n}\mathbb{E}_\varepsilon\left[\varepsilon^T (\lambda I + K)^{-1} K^2 (\lambda I + K)^{-1} \varepsilon\right]$$

$$= \frac{\sigma^2}{n} \cdot \mathrm{tr}\left((\lambda I + K)^{-1} K^2 (\lambda I + K)^{-1}\right)$$

Further bound:

$$(\lambda I + K)^{-1} \le \frac{1}{\lambda}$$

$$(\lambda I + K)^{-1} \le K^{-1}$$

$$\le \frac{\sigma^2}{n\lambda}\mathrm{tr}(K) = \boxed{\frac{\sigma^2}{\lambda}\mathrm{tr}\left(\frac{K}{n}\right)}$$

$\frac{\sigma^2}{\lambda}\mathrm{tr}(S) \simeq \frac{\sigma^2 d}{\lambda}$

often used in parametric case, but too loose here

$y = f^* + \varepsilon$

Naive bound.

$$\text{Error} \le \frac{\lambda}{n}\|f^*\|_H^2 + \frac{\sigma^2}{\lambda}\mathrm{tr}\left(\frac{K}{n}\right)$$

$\lambda \asymp \sqrt{n}$

$$\lambda = \frac{\sigma}{\|f^*\|_H}\cdot\sqrt{\mathrm{tr}(K)}$$

observable

$$\text{Error} \lesssim \underbrace{\sigma\|f^*\|_H}_{\sqrt{n}} \cdot \sqrt{\mathrm{tr}(K/n)} \simeq 1/\sqrt{n}$$

Interpretation. $\text{Error} = \frac{1}{n}\|\hat{f}_\lambda - f^*\|_2^2$

Parametric setting: $\frac{1}{n}$

Non-parametric setting: $1/\sqrt{n}$ ← doesn't depend on geometry of $\mathcal{H}$

↑ Something fishy here

Issue: $\|f^*\|_{\mathcal{H}} < \infty$ is a bad assumption

Ridge regression: Bayesian interpretation

optimal Bayesian inference under Gaussian prior on β.

$$\beta \sim N(0, \rho^2 \cdot I)$$

$\Rightarrow$ ridge regression is optimal

$$\mathbb{E}\left[\|\beta\|_2^2\right] = \rho^2 \cdot p \quad \leftarrow \text{pretty bad if } p = \infty$$

Kernel ridge regression ↙ Gaussian process prior

$$f \sim N(0, \rho^2 \cdot K)$$

$$\|f\|_{\mathcal{H}}^2 = f^T K^{-1} f \qquad \mathbb{E}[f^T K^{-1} f] = \rho^2 \cdot \mathrm{tr}(K \cdot K^{-1})$$

$$= \rho^2 \cdot n \to \infty.$$

K = Gaussian kernel

↳ smooth functions ($\hat{f}_\lambda$ always smooth)

Analysis Strategy.

Assume $f \sim N(0, \rho^2 \cdot K)$ and analyze expected error under this prior on f.

$$\text{Bias}^2 = \mathbb{E}_{f^* \sim N(0, \rho^2 \cdot K)} \left[\frac{\lambda^2}{n} \| (\lambda I + K)^{-1} f^* \|_2^2 \right]$$

$$= \frac{\lambda^2}{n} \text{tr}\left(\mathbb{E}_{f^*} \left[(\lambda I + K)^{-1} f^* (f^*)^T (\lambda I + K)^{-1} \right] \right)$$

$$= \frac{\lambda^2 \rho^2}{n} \text{tr}\left((\lambda I + K)^{-1} K (\lambda I + K^{-1}) \right)$$

$$\text{Var} = \frac{\sigma^2}{n} \cdot \text{tr}\left((\lambda I + K)^{-1} K^2 (\lambda I + K)^{-1} \right)$$

Let $\mu_j = j^{\text{th}}$ eigenvalue of K

$$\text{Error} = \frac{\lambda \rho^2}{n} \sum_j \frac{\lambda \mu_j}{(\lambda + \mu_j)^2} + \frac{\sigma^2}{n} \sum_j \frac{\mu_j^2}{(\lambda + \mu_j)^2}$$

$\frac{\lambda \mu_j}{(\lambda + \mu_j)^2}$:

- $\simeq 1$ if $\lambda \simeq \mu_j$
- $\simeq \frac{\mu_j}{\lambda}$ if $\mu_j << \lambda$
- $\simeq \frac{\lambda}{\mu_j}$ if $\mu_j >> \lambda$

$\frac{\mu_j^2}{(\lambda + \mu_j)^2}$:

- $\simeq 1$ if $\mu_j > \lambda$
- $\simeq \frac{\mu_j^2}{\lambda^2}$ if $\mu_j < \lambda$

≤ 1

$\simeq \min\left(1, \frac{\mu_j^2}{\lambda^2}\right)$

$\leq \min\left(1, \frac{\mu_j}{\lambda}\right)$

$$\asymp \min\left(1, \frac{\mu_j}{\lambda}\right)$$

$$\text{Error} \le \left(\frac{\lambda\rho^2+\sigma^2}{n}\right) \sum_j \min\left(1, \frac{\mu_j}{\lambda}\right)$$

Define $J = \max\{j \mid \mu_j > \lambda\}$

$$= \left(\frac{\lambda\rho^2+\sigma^2}{n}\right)\left(J + \frac{1}{\lambda}\sum_{j>J} \mu_j\right)$$

Example.

$$\mu_j = j^{-\alpha} \cdot n, \quad \alpha > 1$$

$J^{-\alpha} \cdot n = \lambda$

$J = \left(\frac{\lambda}{n}\right)^{-\frac{1}{\alpha}}$

$$J = \left(\frac{\lambda}{n}\right)^{-\frac{1}{\alpha}}$$

$$\frac{1}{\lambda}\sum_{j>J} \mu_j \asymp \frac{n}{\lambda}\int_J^\infty x^{-\alpha}\, dx$$

$$= \frac{n}{\lambda}\,\frac{J^{1-\alpha}}{\alpha-1} = \frac{n}{\lambda}\,\frac{\left(\frac{\lambda}{n}\right)^{-\frac{1}{\alpha}(1-\alpha)}}{\alpha-1}$$

$$= \frac{1}{\alpha-1}\left(\frac{\lambda}{n}\right)^{-\frac{1}{\alpha}}$$

$$\text{Error} \le \left(\frac{\lambda\rho^2+\sigma^2}{n}\right) \cdot \left(\frac{\alpha}{\alpha-1}\left(\frac{\lambda}{n}\right)^{-\frac{1}{\alpha}}\right)$$

$$\lambda^* = \frac{\sigma^2}{\rho^2}$$

$$\text{Error}(\lambda^*) = \frac{\sigma^2}{n} \cdot \frac{\alpha}{\alpha - 1} \left(\frac{\sigma^2}{n\rho^2} \right)^{-\alpha}$$

$$= \frac{\alpha}{\alpha - 1} n^{\frac{1}{\alpha} - 1} (\sigma^2)^{1 - \frac{1}{\alpha}} (\rho^2)^{\frac{1}{\alpha}}$$

$n^{\frac{1}{\alpha} - 1}$ → rate of decay

- $\alpha \to 1$: very poor error
- $\alpha = 2$: $1/\sqrt{n}$
- $\alpha \to \infty$: approach $1/n$ error from parametric setting

---

(2) Generalizing from train test

Train: $\frac{1}{n} \sum_{i=1}^{n} \left(f(x_i) - \hat{f}^{\lambda}(x_i) \right)^2$

Test: $\mathbb{E}_{x \sim p} \left[\left(f(x) - \hat{f}^{\lambda}(x) \right)^2 \right]$

Return to parameter view

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

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

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

Ridge regression: $\hat{\beta}_\lambda = \left(\frac{\lambda}{n} I + S \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} x_i y_i \right)$

$$= \left(\lambda_n I + S \right)^{-1} \left(S\beta^* + \sum_{i=1}^{n} x_i \varepsilon_i \right)$$

Error: $(\hat{\beta}_\lambda - \beta^*)^T \hat{\Sigma} (\hat{\beta}_\lambda - \beta^*)$

$$\mathbb{E}[\hat{\beta}_\lambda] = (\lambda_n I + \Sigma)^{-1} \Sigma \beta^*$$

Bias^2: $\delta = \mathbb{E}[\hat{\beta}_\lambda] - \beta^*$, $\delta = \lambda_n (\lambda_n I + \Sigma)^{-1} \beta^*$

$$\delta^T \Sigma^* \delta = \lambda_n^2 (\beta^*)^T (\lambda_n I + \Sigma)^{-1} \Sigma^* (\lambda_n I + \Sigma)^{-1} \beta^*$$

Under Gaussian proof:

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

Key observation: Almost the same as bias if we ran ridge regression on test distr^n:

$$\text{trace}\left((\lambda_n I + \Sigma^*)^{-1} \Sigma^* (\lambda_n I + \Sigma^*)^{-1} \right)$$

In particular, as long as

$$(\lambda_n I + \Sigma^*)^{-1} \preceq 2 (\lambda_n I + \Sigma)^{-1},$$

test error $\le C \cdot$ train error

Proposition: Assume $\Sigma^* \preceq 2\Sigma + \lambda_n I$.

Then $\mathbb{E}[\text{Error}] \le 4 \left(\lambda_n \rho^2 + \frac{\sigma^2}{n} \right) \sum_j \min\left(1, \frac{\mu_j^*}{\lambda_n}\right)$

$\frac{\mu_j}{\lambda}$

whose μ_j^* is j^{th} largest eigenvalue of S^*.

n times smaller b/c S vs. K

Take-away: As long as we pick λ_n s.t.

$$S^* \leq 2S + \lambda_n I$$ holds,

we got good generalization.

TODO: Show when this is the case.