Nonparametric Learning

↳ STAT 240: robust + nonparametric

→ Learning from some infinite-dimensional space.

→ parametric: linear regression w/ d-dimensional covariates

→ nonparametric: all smooth functions
- decision trees, neural networks

Still linear, but just w/ ∞ parameters

Today: Re-do basics of ordinary least squares + ridge regression, but in infinite dimensions.

Shift basis: parameter view → function view.

Ordinary least squares:

$(X_1, Y_1), \ldots, (X_n, Y_n)$

$$L(\beta) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \langle \beta, X_i \rangle)^2$$

$$\beta^* = (X^T X)^{-1} X^T Y = \left(\frac{1}{n} \sum_{i=1}^{n} X_i X_i^T \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^{n} X_i Y_i \right)$$

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

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

Ridge regression:

$$L_\lambda(\beta) = \frac{\lambda}{n}\|\beta\|_2^2 + L(\beta)$$

$$\hat{\beta}_\lambda = (\lambda n I + X^T X)^{-1} X^T Y$$

$$= (\lambda I + S)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i Y_i\right)$$

Parameter norm	Error norm
$\|\beta\|_2^2$	$(\beta-\beta^*)^T S (\beta-\beta^*)$

Function view: $f_\beta(x) = \beta^T x$ instead of β

$$f_\beta \in \mathbb{R}^n \quad (f_\beta(X_1), \ldots, f_\beta(X_n))$$

$$L(f) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - f(X_i))^2$$

$$L(f) - L(f^*) = \frac{1}{n}\|f - f^*\|_2^2$$

Question: what is $f^* = f_{\beta^*}$? what is $\|\beta\|_2^2$? $f_{\hat{\beta}_\lambda}$?

$$f_\beta = X\beta, \quad X \in \mathbb{R}^{n \times p}$$

$$f^* = f_{\beta^*} = X\beta^* = \underbrace{X(X^TX)^{-1}X^T}_{\text{projection onto linear span of } X}\, y$$

f^* : projection of y onto linear span of X

$$\|\beta\|_2^2 = \beta^T\beta$$

$$\boxed{\beta = (X^TX)^{-1}X^Tf} \rightarrow f = X\beta$$

$$(X^TX)^{-1}X^TX\beta = \beta$$

$$\beta^T\beta = f^T\boxed{X(X^TX)^{-2}X^T}f$$

$$= (XX^T)^{\dagger} \quad \leftarrow \text{pseudoinverse}$$

$$= K^{\dagger}$$

$$K = XX^T$$

$$K_{ij} = \langle x_i, x_j \rangle$$

"Kernel matrix"

$$\Rightarrow \boxed{\|\beta\|_2^2 = \begin{cases} f^TK^{\dagger}f & : \text{if } f \in \text{span}(K) \\ \infty & : \text{if } f \notin \text{span}(K) \end{cases}}$$

$$\underset{f}{\operatorname{argmin}} \ \frac{1}{n}\|y-f\|_2^2 \Rightarrow (f=y)$$

$$\underset{f:\, f\in \mathrm{Span}(K)}{\operatorname{argmin}} \ \frac{\lambda}{n} R(f) + \frac{1}{n}\|y-f\|_2^2$$

	Complexity norm	Error norm
Parameter view	$\|\beta\|_2^2$	$(\beta-\beta^*)^T S (\beta-\beta^*)$
Function view	$f^T K^\dagger f$	$\frac{1}{n}\|f-f^*\|_2^2$

Ridge regression:

$$\underset{f}{\operatorname{argmin}} \ \frac{\lambda}{n} f^T K^\dagger f + \frac{1}{n}\|f-y\|_2^2 \qquad K^\dagger = K^{-1}$$

$$\Rightarrow \lambda K^\dagger f + f - y = 0$$

$$\Rightarrow (\lambda K^\dagger + I) f = y$$

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

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

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

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

↳ y, kernel matrix K

Kernel ridge regression

$$f_\beta(x) = \beta^T x \qquad \beta = (X^T X)^{-1} X^T f$$

$$= x^T \beta \qquad K = XX^T$$

$$= x^T (X^T X)^{-1} X^T f \qquad K^\dagger = X (X^T X)^{-2} X^T$$

$$= x^T \; X^T X \, (X^T X)^{-2} X^T \; f$$

$$= x^T X^T K^\dagger f \qquad = K_x^T K^\dagger f$$

K_x^T

$X \in \mathbb{R}^{n \times p}$

$x \in \mathbb{R}^{p \times 1}$

$(Xx)_i = \langle x_i, x \rangle$

$$\hat{f} = f_\lambda$$

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

$$= K_x^T (\lambda I + K)^{-1} y \qquad \text{(if } K \text{ invertible)}$$

Reproducing Kernel Hilbert spaces

We only need to care about K $\langle x_i, x_j \rangle$

Assume we have positive definite kernel function $k(x,x')$

$T = \{x_1, \dots, x_n\}$

$K_{ij} = k(x_i, x_j)$

then K is positive semi definite, and it is strictly positive definite if the x_i are case distinct.

↳ implies $\mathcal{H}$ is infinite-dimensional otherwise $\text{rank}(K) \le p$

Def^n Given k, the corresponding reproducing Kernel Hilbert space $\mathcal{H}$ is all functions $f: \mathcal{X} \to \mathbb{R}$ s.t.

$$\sup_T \|f\|_{K[T]} < \infty.$$

↳ $\|f\|_{K[T]} = \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}^T K^{-1} \begin{bmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}$

"$\|\beta\|_2^2$"

Also define the norm

$$\|f\|_{\mathcal{H}} = \sup_T \|f\|_{K[T]}.$$

Lemma 1 (Restrictions are contractions)

For any f and sets $T \subseteq T'$, $\|f\|_{K[T]} \le \|f\|_{K[T']}$.

Lemma 2 (Isometric extension).

Let $T = \{x_1, \dots, x_n\}$ and $y_1, \dots, y_n \in \mathbb{R}$.

Then there exists an $f \in \mathcal{H}$ such that $f(x_i) = y_i$ $\forall i$

and $\|f\|_{\mathcal{H}}^2 = \|y\|_{K[T]}^2 = y^T K^{-1} y$.

Pf of Lemma 2.

Define $f(x) = k_x^T K^{-1} y$

- $f(x_i) = y_i$: $f(x_i) = k_{x_i}^T K^{-1} y$

$= K_{:i}^T K^{-1} y = e_i^T y = y_i$

- Boundary norm of f

$\|f\|_{\mathcal{H}} = \sup_{T'} \|f\|_{K[T']}$

$T' = \{x'_1, \dots, x'_n\}$

$f[T']^T K[T']^{-1} f[T']$

$f[T'] = \begin{bmatrix} f(x'_1) \\ \vdots \\ f(x'_n) \end{bmatrix}$

$f[T'] = \begin{bmatrix} K_{x'_1} \\ \vdots \\ K_{x'_n} \end{bmatrix}^T K[T]^{-1} y$

$K[T', T]$

$K[T', T]_{ij} = k(x'_i, x_j)$

$\|f\|^2_{K[T']} = y^T K[T]^{-1} K[T, T'] K[T']^{-1} K[T', T] K[T]^{-1} y$

$K_{11} = K[T]$

$K_{12} = K[T, T']$

$K_{21} = \cdots$

K_{22}

$\begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix} \succeq 0$

Kernel matrix on TUT'

$= y^T K_{11}^{-1} K_{12} K_{22}^{-1} K_{21} K_{11}^{-1} y$ ✓

WANT: $\leq y^T K_{11}^{-1} y$ (original $\|y\|_K^2$)

$\Rightarrow$ WANT: $K_{11}^{-1} K_{12} K_{22}^{-1} K_{21} K_{11}^{-1} \leq K_{11}^{-1}$ ✓

$K_{12} K_{22}^{-1} K_{21} \leq K_{11}$

↳ Holds via Schur complement lemma

$$\begin{bmatrix} K_{22} & K_{21} \\ K_{12} & K_{11} \end{bmatrix} \quad \begin{bmatrix} A & B \\ B^T & C \end{bmatrix} \succeq 0$$

$\Leftrightarrow$ $B^T A^{-1} B \leq C$

and $A \succeq 0$ ∎

Implications of Lemma 2 for kernel ridge regression.

$$\hat{f}_\lambda = \underset{f}{\operatorname{argmin}} \frac{\lambda}{n} \|f\|_{K[T]}^2 + \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

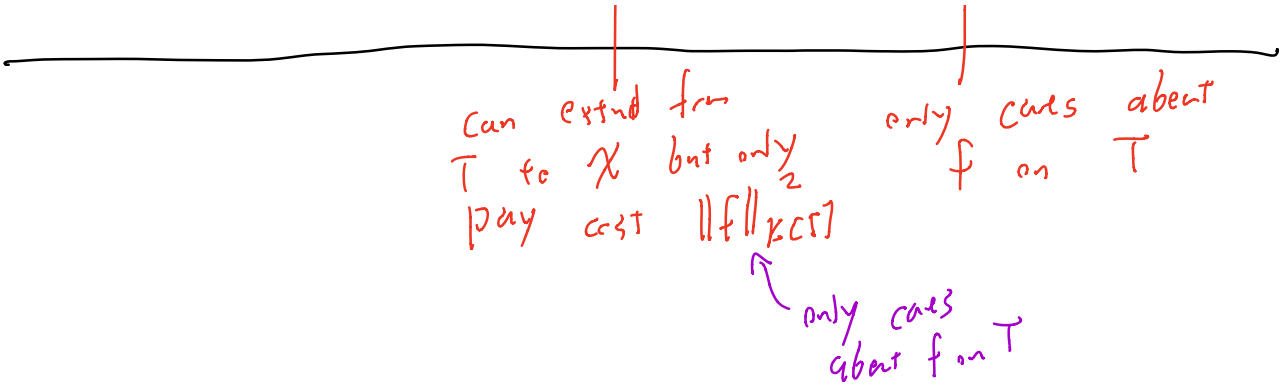
$f: T \to \mathbb{R}$

$K[T]$ depends on $T = \{x_1, \dots, x_n\}$

by Lemma 2

$$= \underset{f}{\operatorname{argmin}} \frac{\lambda}{n} \|f\|_H^2 + \frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2$$

$f: \mathcal{X} \to \mathbb{R}$



Bias-variance decomposition

$\varepsilon_i \sim N(0, \sigma^2)$

$$Y_i = f^*(X_i) + \varepsilon_i$$

$\hat{f}$: bias and variance of $\hat{f}$

$$\text{Bias}^2 \quad \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{E}_\varepsilon\left[\hat{f}(X_i)\right] - f^*(X_i) \right)^2$$

$$\text{Var} = \frac{1}{n}\sum_{i=1}^{n} \text{Var}_\varepsilon\left[\hat{f}(X_i)\right]$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_\varepsilon\left(\hat{f}(X_i) - f^*(X_i)\right)^2 = \text{Bias}^2 + \text{Var}$$

Ordinary least squares analyzed these in context of regular standard errors

$\hat{f}_\lambda$: ridge regression

$$\hat{f}_0 = K(\lambda I + K)^{-1} y \Big|_{\lambda=0} = K K^{-1} y = y$$

$$Var[\hat{f}_0(x_i)] = Var[y_i] = \sigma^2$$

Overall variance: σ^2 $\longrightarrow$ Doesn't $\to 0$ as $n \to \infty$.

Bias : 0