〈통계적 기계학습〉

1. Supervised Learning.

1.1. Basic Model

— input : $x \in \mathbb{R}^d \Rightarrow x = (x_1 \dots x_d)$

output : $y \in \mathcal{Y}$ ⟨ classification / regression

model : $y \approx f(x)$, $f \in \mathcal{M}$.

$y = \phi(f(x), \varepsilon)$. $\xrightarrow{\text{additive}}$ $y = f(x) + \varepsilon$

— loss function : $l(y, a)$ ⟨ $l(y,a) = (y-a)^2$ / $l(y,a) = \mathbb{I}(y \neq a)$.

— goal : $f^0 = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(Y,X) \sim P}[l(Y, f(X))]$. $(\mathcal{F} \subseteq \mathcal{M})$.

— prediction : f^0 is unknown. → estimate f^0 by $\hat{f}$.

$$\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{\hat{P}_n}[l(Y, f(X))] \quad \hat{P}_n : \text{empirical distribution } \frac{1}{n}\sum_{i=1}^n \delta_{(Y_i, X_i)}$$
$$= \arg\min_{f} \frac{1}{n} \sum l(Y_i, f(X_i))$$

1.2. Linear Regression and Least Neighbors.

— Linear Reg. Model : $\mathcal{M} = \mathcal{F} = \left\{ \beta_0 + \sum_{j=1}^d \beta_j x_j : \beta_j \in \mathbb{R} \right\}$.

→ use least squares : empirical loss → RSS.

$$RSS(\beta) = \sum_{i=1}^n (y_i - f(x_i))^2$$
$$= \sum_{i=1}^n (y_i - \beta_0 - \sum x_{ij}\beta_j)^2$$

— k-Nearest Neighbor.

→ $\hat{f}(x) = \frac{1}{k} \sum_{x_i \in N_k(x)} y_i$ $\quad N_k(x)$: neighborhood of x.

1.3. Statistical Decision Theory.

— Risk / Expected Prediction Error (EPE)

$$R(f) = \mathbb{E}_{(Y,X)\sim P}[l(Y, f(X))] = \mathbb{E}[(Y - f(X))^2].$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}[(Y-f(X))^2 | X]].$$

where $\mathbb{E}_{Y|X}[(Y-f(x))^2|X] = \mathbb{E}_{Y|X}[(Y - \mathbb{E}[Y|X])^2|X] + (\mathbb{E}[Y|X] - f(x))^2$

⇒ $R(f)$ is minimized when $f^0(x) = \mathbb{E}[Y|X=x]$. "regression function" (usually unknown.)

— k-NN) $\hat{f}(x) = \text{Ave}(y_i | x_i \in N_k(x))$.

→ two approximations ⟨ expectation → average. / conditioning at a point → conditioning on neighbor of x.

→ under mild regularity conditions on $P(X,Y)$, $\hat{f}(x) \to \mathbb{E}[Y|X=x]$

as $n, k \to \infty$ s.t. $k/n \to 0$. model complexity grows slower than sample size. (asymptotic sample efficiency).

— linear reg.) assume $f^*(x)$ linear. $\mathbb{E}[Y|X=x] = x^T\beta$.

→ model-based approach) let $\beta^* = (\mathbb{E}(X^TX))^{-1}\mathbb{E}[XY]$.

$$R(f) = \mathbb{E}[(Y - X^T\beta)^2]$$
$$= \mathbb{E}[(Y - X^T\beta^*)^2] + \mathbb{E}[(X^T\beta^* - X^T\beta)^2]. \Rightarrow \beta^* \text{ minimizing the risk.}$$

→ $\hat{\beta} = (X^TX)^{-1}X^TY$: average over empirical dist.

→ $\hat{\beta} \xrightarrow{p} \beta^*$. ($\because$ L.L.N.)

$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \Phi)$.

— classification ($y = \{0, \dots, K-1\}$.)

$$R(f) = \mathbb{E}_{(Y,X)\sim P}[L(Y, f(X))] = \mathbb{E}[I(Y \neq f(X))].$$
$$= \mathbb{E}_X[\mathbb{E}_{Y|X}[L(y, f(x))|X]].$$
$$= \mathbb{E}_X\left[\sum_{j=0}^{K-1} L(j, f(x))\, P(Y=j|X)\right]. \quad I(j \neq k).$$

→ minimized when $f^*(x) = \underset{k=0,\dots,K-1}{\operatorname{argmin}} \sum_{j=0}^{K-1} \boxed{L(j,k)}\, P(Y=j|X=x)$.

$$= \underset{j=0,\dots,K-1}{\operatorname{argmax}} P(Y=j|X=x).$$

↳ usually unknown.

→ optimal classifier = Bayes classifier.

— K-NN) $P(Y=j|X=x) \approx \frac{1}{K}\sum_{i \in N_K(x)} I(y_i = j)$.

2.4. Curse of Dimensionality.

— $X = (X_1 \dots X_d) \sim \text{Uniform}([0,1]^d)$.

hypercubical neighbor → side length $e_d(r) = r^{1/d}$. (capture fraction r of the obs.).

but $e_{10}(0.01) = 0.63$, $e_{10}(0.1) = 0.80$. → no longer "local"

— $X = (X_1 \dots X_d) \sim \text{Uniform}(B_{R^d}(0,1))$

where $B_{R^d}(0,1) = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$.

$X_{i\cdot} = (X_{i1} \dots X_{id})$, $R_i = \sqrt{\sum_{j=1}^d X_{ij}^2}$ → $\text{Med}(R_{(1)}) = \left(1 - \left(\frac{1}{2}\right)^{\frac{1}{n}}\right)^{\frac{1}{d}}$.

min of R_i.

if $n=5000$ & $d=10$, $\text{Med}(R_{(1)}) \simeq 0.852$

— (radius of inner sphere) = (length of half diagonal) − (radius of spheres).

↓ grows as $\frac{\sqrt{d}}{2}$. ↓ (constant) $\frac{1}{2}$.

if $d > 4$, inner sphere starts poking outside. ($\frac{\sqrt{d}-1}{2} > \frac{1}{2}$).

1.5. Overfitting & the Bias–Variance Tradeoff.

— estimate risk $\mathbb{E}_{(x,y)\sim P}[\ell(Y, f(X))]$ by training error (TE). $\frac{1}{n}\sum \ell(Y_i, f(X_i))$.
$= \mathbb{E}_{\hat{P}_n}[\ell(Y, f(X))]$

— TE always underestimate prediction error!

Better estimation : test error.

— Bias – Variance Tradeoff:

suppose $Y = f(X) + \varepsilon$, $\mathbb{E}(\varepsilon) = 0$, $\mathrm{Var}(\varepsilon) = \sigma^2$.

$\hat{f}(X, \mathcal{T})$: predicted function trained on $\mathcal{T}$.

$$R(f) = \mathbb{E}[(Y - \hat{f}(X,\mathcal{T}))^2]$$
$$= \mathbb{E}[(Y - f(X))^2] + \mathbb{E}[(f(X) - \hat{f}(X,\mathcal{T}))^2]$$
$$= \sigma^2 + \mathbb{E}_X[(f(X) - \mathbb{E}_{\mathcal{T}}[\hat{f}(X,\mathcal{T}) \mid X])^2] + \mathbb{E}_X[\mathbb{E}_{\mathcal{T}}[(\hat{f}(X,\mathcal{T}) - \mathbb{E}_{\mathcal{T}}[\hat{f}(X,\mathcal{T}) \mid X])^2 \mid X]]$$
$$= \sigma^2 + \mathbb{E}_X[\mathrm{Bias}_{\mathcal{T}}(X)^2 + \mathrm{Variance}_{\mathcal{T}}(X)].$$

$\mathbb{E}[(Y - \hat{f}(X))(\hat{f}(X) - \hat{f}(X,\mathcal{T}))]$
$= \mathbb{E}_X[g(X)(\mathbb{E}[Y \mid X] - \hat{f}(X))]$
$= \mathbb{E}_X[g(X) \cdot \mathbb{E}_{Y|X}[Y - \hat{f}(X) \mid X]]$
$= 0$.

$\mathbb{E}[(f(X) - \hat{f}(X,\mathcal{T}))^2]$
$= \mathbb{E}_X[\mathbb{E}_{\mathcal{T}}[(f(X) - \hat{f}(X,\mathcal{T}))^2 \mid X]]$.
$\mathbb{E}_{\mathcal{T}}[(f(X) - \mathbb{E}_{\mathcal{T}}[\hat{f}(X,\mathcal{T}) \mid X])^2 \mid X]$ → variance
$+ \mathbb{E}_{\mathcal{T}}[(\mathbb{E}_{\mathcal{T}}[\hat{f}(X,\mathcal{T}) \mid X] - \hat{f}(X,\mathcal{T}))^2 \mid X]$.

— (k-NN) $\hat{f}(x) = \hat{f}(x, \mathcal{T}) = \frac{1}{k}\sum_{\ell=1}^{k}(f(x_{(\ell)}) + \varepsilon_{(\ell)})$. (ℓ) : ℓ^{th}-N.N.

$\Rightarrow$ $\mathrm{Bias}_{\mathcal{T}}(x) = f(x) - \frac{1}{k}\sum_{\ell=1}^{k} f(x_{(\ell)})$.

$\mathrm{Variance}_{\mathcal{T}}(x) = \frac{1}{k^2}\mathrm{Var}(\sum \varepsilon_{(\ell)}) = \frac{\sigma^2}{k}$.

$\Rightarrow$ if $k \uparrow$, bias $\uparrow$, variance $\downarrow$.

2. Linear Classifiers.

2.1. Introduction

— decision boundary (between class i and j)

$\{x \in \mathcal{X} : \text{for all } r > 0,\ B(x,r) \cap \hat{f}^{-1}\{i\} \neq \emptyset,\ B(x,r) \cap \hat{f}^{-1}\{j\} \neq \emptyset\}$

$\Rightarrow$ total decision boundary : Union over all (i, j).

— linear classifier : decision boundary btw i and j given as subset of affine hyperspace.

$\{x : H(x) = 0\}$, where $H(x) = \beta_0 + x^T\beta$.

— { model joint probability dist.
model conditional density function $P(\text{class} \mid X)$.

2.2. Bayes Classifiers

$\mathbb{E}[\sum_{j=0}^{k} \ell(j, \hat{f}(x))\, \mathbb{1}(Y = j) \mid X = x]$
$= \sum \ell(j, \hat{f}(x))\, \mathbb{E}[\mathbb{1}(Y = j) \mid X = x]$
$= \sum \ell(j, \hat{f}(x))\, P(Y = j \mid X = x)$.

Thm $R(f) = \mathbb{E}_{(Y,X)\sim P}[\ell(Y, \hat{f}(X))] = \mathbb{E}_X[\mathbb{E}_{Y|X}[\ell(Y, \hat{f}(X)) \mid X]]$.

$\rightarrow f^*(x) = \underset{k}{\mathrm{argmax}}\ P(Y = k \mid X = x)$. ∴ Bayes classifier.

— Binary) $f^*(x) = \mathbb{1}(P(Y = 1 \mid X = x) > \frac{1}{2})$.

- 베이즈 정리) $P(Y=j \mid X=x) = \frac{P(Y=j, X=x)}{P(X=x)}$.

$= \frac{P(Y=j, X=x)}{\sum P(Y=l, X=x)} = \frac{\pi_j p_j(x)}{\sum \pi_l p_l(x)}$.

$\Rightarrow P(Y=k \mid X=x) > P(Y=j \mid X=x) \Leftrightarrow \frac{p_k(x)}{p_j(x)} > \frac{\pi_j}{\pi_k}$.

Bayes rule: $f^*(x)=k$ when $\frac{p_k(x)}{p_j(x)} > \frac{\pi_j}{\pi_k}$ for all $j \neq k$.

$\left(f^*(x) = \underset{j}{\operatorname{argmax}}\ \pi_j p_j(x). \right) \rightarrow R(f^*) =$ Bayes risk.

- $f^* \in H$ that minimizes $R(f) \Rightarrow$ oracle classifier : $R(f^*) = \inf_{f \in H} R(f)$.

2.3. Logistic Classification.

- $\hat{f}(x)=k$ if $\hat{m}_k(x) > \hat{m}_j(x)$ for all $j \neq k$, $m_j(x) = P(Y=j \mid X=x)$ $= \mathbb{E}[I(Y=j) \mid X=x]$.
- $p(y; \pi_i) = \pi_i^y (1-\pi_i)^{1-y}$ for $y=0,1$.

$\rightarrow L(\pi_i) = \prod_{i=1} p(Y_i; \pi_i) = \prod \pi_i^{Y_i} (1-\pi_i)^{1-Y_i}$.

- Assumption: $P(Y=k \mid x) = \frac{\exp(\beta_{k0} + x^T \beta_k)}{1 + \sum_{j=1}^{K-1} \exp(\beta_{j0} + x^T \beta_j)}$.

- Bayes classifier : $f^*(x) = \underset{j}{\operatorname{argmax}}\ p(y=j \mid x)$.
- linear boundary : where $P(Y=i \mid X=x) = P(Y=j \mid X=x)$.

$\Rightarrow \{ x \in \mathbb{R}^d : (\beta_{i0} - \beta_{j0}) + x^T (\beta_i - \beta_j) = 0 \}$.

- motivation ①)

$P(Y=1 \mid x) = F(\beta_0 + x^T \beta) \rightarrow$ F : cont, increasing, $F(-\infty)=0$, $F(\infty)=1$.

logistic : $F(x) = \exp(x) / (1+\exp(x))$.

motivation ②) $\Rightarrow P(Y=2 \mid X=x) = P(Y=1 \mid X=x)$.

$\{ x : Pr(Y=1 \mid X=x) = 0.5 \} \Leftrightarrow \{ x : \log \left(P(Y=1 \mid X=x) / P(Y=2 \mid X=x) \right) = 0 \}$. $\leftarrow \beta_0 + x^T \beta$.

Suppose log-odds is linear. $\rightarrow P(Y=1 \mid X=x) = \frac{\exp(\beta_0 + x^T \beta)}{1+\exp(\beta_0 + x^T \beta)}$.

- estimation) MLE : $L(\{\beta_{j0}\}, \{\beta_j\}) = \prod Pr(Y=Y_i \mid X=X_i)$.

of binary, $= \prod \pi_i(x_i, \beta_0, \beta)^{Y_i} (1-\pi_i(x_i, \beta_0, \beta))^{1-Y_i}$.

$\ell(\beta_0, \beta) = \sum \{ Y_i \log \pi_i - (1-Y_i) \log(1-\pi_i) \}$

$= \sum \{ Y_i (\beta_0 + X_i^T \beta) - \log(1+\exp(\beta_0 + X_i^T \beta)) \}$

$\hookrightarrow$ Newton's method.

$\frac{\exp(\beta_0 + X_i^T \beta)}{1+\exp(\beta_0 + X_i^T \beta)}$.

2.4. LDA.

— Let $p_j(x) = p(x|y=j)$. $j = 0, \dots, K-1$.

$\pi_j = p_r(y=j)$ . . $\sum_{j=0}^{K-1} \pi_j = 1$.

(* Bayes classifier: $f^*(x) = \text{argmax}\ p_j(x)\,\pi_j$.)

— Model $p_j(x) = p(x|Y=k)$ as multivariate Gaussian. : $X|Y=j \sim N(\mu_j, \Sigma_j)$.

Thm If $X|Y=j \sim N(\mu_j, \Sigma_j)$, then ($\pi_j p_j(x) > \pi_\ell p_\ell(x)$)

Bayes rule: $f^*(x) = j$ if $\delta_j(x) > \delta_\ell(x)$ for $\forall \ell \neq j$.

where $\delta_j(x) = -\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(x-\mu_j)^T\Sigma_j^{-1}(x-\mu_j) + \log\pi_j$.

↳ $\log[\pi_j p_j(x)]$

$\Leftrightarrow$ $f^*(x) = \text{argmax}_j\ \delta_j(x)$.

decision boundary : $\{x \in \mathcal{X} : \delta_\ell(x) = \delta_j(x)\}$. → quadratic.

— In practice) use sample quantities.

$\hat{\pi}_j = \frac{n_j}{n}$, $\hat{\mu}_j = \frac{1}{n_j}\sum x_i\, I(Y_i = j)$, $\hat{\Sigma}_j = \frac{1}{n_j - 1}\sum (X_i - \hat{\mu}_0)(X_i - \hat{\mu}_0)^T I(Y_i = j)$.

— If we assume $\Sigma_j = \Sigma$ for all j,

$f^*(x) = \text{argmax}_k\ \delta_k(x)$ where $\delta_j(x) = x^T\Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T\Sigma^{-1}\mu_j + \log\pi_j$.

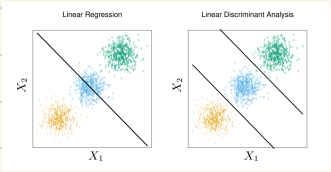
decision boundary : $\{x \in \mathcal{X} : \delta_\ell(x) = \delta_j(x)\}$ → linear.

— $\hat{\Sigma} = \frac{1}{n-K}\sum_{j=1}^{K}(n_j - 1)\hat{\Sigma}_j$.

— LDA: $d + \frac{d(d-1)}{2}$ parameters vs. QDA: $d + d(d+1)$.

2.5. LDA or Logistic classification.

— LDA
- need more assumptions.
- has troubles with categorical inputs.
- useful when some output are missing.
- works better for multi-class problems.



3. Shrinkage Methods. — Multicollinearity, Shrinkage.

3.1. Regularization.

typically convex.

— $\min_\beta \|y - X\beta\|_2^2$ s.t. $\beta \in C$ (constrained form)

$\min_\beta \|y - X\beta\|_2^2 + P(\beta)$ (penalized form)

— Three norms : $\begin{cases} \|\beta\|_0 = \sum_{j=1}^d I(\beta_j \neq 0) \\ \|\beta\|_1 = \sum_{j=1}^d |\beta_j| \\ \|\beta\|_2^2 = \sum_{j=1}^d \beta_j^2 \end{cases}$

— $\min_\beta \|y-X\beta\|_2^2$ s.t. $\|\beta\|_0 \le k$ (1) $\Leftarrow \not\Rightarrow$ $\min_\beta \frac{1}{2}\|y-X\beta\|_2^2 + \lambda\|\beta\|_0$ (4)

// s.t. $\|\beta_1\| \le s$ (2) $\Leftrightarrow$ // $+ \lambda\|\beta\|_1$ (5)

// s.t. $\|\beta_2\| \le S$ (3) $\Leftrightarrow$ // $+ \lambda\|\beta\|_2^2$ (6)

$\lambda \ge 0$ 존재.

— Toy example) $Y = (Y_1, \dots, Y_d)^T \in \mathbb{R}^d \sim N_d(\mu, \sigma^2 I)$.

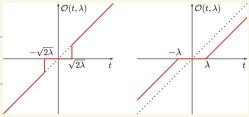
— MLE of μ = Y.

— when $d \le 2$, Y is the best estimator.

$d \ge 3$, best = $\hat{\mu}^{JS} = \left(1 - \frac{(d-2)\sigma^2}{\sum Y_i^2}\right) Y$.

— $\begin{cases} E[\|Y-\mu\|_2^2] = d\sigma^2. \\ E[\|\hat{\mu}^{JS}-\mu\|_2^2] = d\sigma^2 - (d-2)^2\sigma^4 E\left[\frac{1}{\|Y\|_2^2}\right]. \end{cases}$

— solutions of (*) $\begin{cases} \hat{\mu}_i = H(Y_i; \sqrt{2\lambda}) \to \text{hard thresholding.} & \text{dis conti} \\ \hat{\mu}_i = S(Y_i; \lambda) \to \text{soft thresholding.} & \text{conti.} \\ \hat{\mu}_i = \frac{Y_i}{1+2\lambda}. \to \text{never sparse.} \end{cases}$



— sparsity : $(p \le 1)$.

(2) : $s \downarrow$

(4), (5) $\lambda \uparrow$ $\Rightarrow$ higher degree of sparsity.

— convexity : (2), (3). $(p \ge 1)$

3.2. Variable Selection.

— $\hat{\beta}_A = (X_A^T X_A)^{-1} X_A^T y$. $\hat{\beta}_{-A} = 0$. $A = supp(\beta)$.

$\hat{\beta}_S^{oracle} = (X_S^T X_S)^{-1} X_S^T y$, $\hat{\beta}_{-S} = 0$ $S = supp(\beta_0)$. $s_0 = |S|$.

$\Rightarrow$ In-sample risk : $\frac{1}{n}\|X\beta^{oracle} - X\beta_0\|_2^2 = \sigma^2 \frac{s_0}{n}$.

┌ risk of oracle risk 비교. → $\frac{1}{n}\mathbb{E}\|X\hat\beta - X\beta_0\|_2^2 \le \sigma^2\frac{S_0}{n}\left(4\log d + 2 + o(1)\right)$. $\asymp \log d$.

- If $S \asymp \sigma^2 \log d$, best subset selection estimator $\hat\beta$ satisfies

$$\begin{cases} \dfrac{\mathbb{E}\|X\hat\beta - X\beta_0\|_2^2/n}{\sigma^2 S/n} \le 4\log d + 2 + o(1). & \text{as } n,d \to \infty. \\ \inf_{\hat\beta}\sup_{X\beta_0} \dfrac{\mathbb{E}\|X\hat\beta - X\beta_0\|_2^2/n}{\sigma^2 S/n} \ge 2\log d - o(\log d). \end{cases}$$

3.3. Ridge Regression.

Def $\hat\beta^{ridge} = \operatorname{argmin} \sum_{i=1}^n \left(y_i - \beta_0 - \sum_{k=1}^d x_{ik}\beta_k\right)^2$ subject to $\sum_{k=1}^d \beta_k^2 \le S$.

$= \operatorname{argmin}$ // $+ \lambda\sum\beta_k^2$

↳ $\|Y - X\beta\|_2^2 + \lambda\|\beta\|_2^2$, $\hat\beta = (X^TX + \lambda I)^{-1}X^TY$.

3.4. LASSO

Def $\hat\beta^{lasso} = \operatorname{argmin} \sum_{i=1}^n \left(y_i - \beta_0 - \sum x_{ik}\beta_k\right)^2$ subject to $\sum|\beta_k| \le S$.

$= \operatorname{argmin}$ // $+ \lambda\sum|\beta_k|$ ↳ non-differentiable around 0.

- optimization w/ subgradient optimality (KKT conditions).

∴ $\hat\beta$ must satisfy $X^T(y - X\hat\beta) = \lambda s$. (i)

where $s \in \partial\|\hat\beta\|_1 = \begin{cases} +1 & \hat\beta_j > 0 \\ -1 & \hat\beta_j < 0 \\ [-1,1] & \hat\beta_j = 0. \end{cases}$ $j = 1,\dots,d$. (ii)

→ (i)의 의미: $\hat\beta$가 유일하게 결정되지 않아도, s는 $X\hat\beta$의 함수이므로 유일.

→ (ii)로부터, s의 유일성 ⇒ overlapped support에서는 같은 부호.

($\hat\beta_j > 0$ but $\tilde\beta_j < 0$ for some j 불가).

→ X의 general position ($\simeq$ 스무스) 이면 $\hat\beta^{LASSO}$는 유일?

pf) Let $A = \operatorname{supp}(\hat\beta)$. $s_A = \operatorname{sign}(\hat\beta_A)$.

(i), (ii) ⇒ $X_A^T(y - X_A\hat\beta_A) = \lambda s_A$.

↔ $\hat\beta_A = (X_A^TX_A)^{-1}(X_A^Ty - \lambda s_A)$. $\hat\beta_{-A} = 0$. (invertible)

$= (X_A^TX_A)^{-1}X_A^Ty - \lambda(X_A^TX_A)^{-1}s_A$

shrinkage. (cf. best subset : no shrinkage).

↳ orthogonal 이면 $X_A^TX_A = I$.

$\hat\beta_j^{LASSO} = \hat\beta_j^{OLS} - \lambda s_j$. 어떻게? $\operatorname{sign}(\hat\beta_j)$

교재 표현	의미
Prediction error	$\|X\hat\beta - X\beta_0\|^2$ (확률변수)
In-sample risk	$\frac{1}{n}\mathbb{E}\|X\hat\beta - X\beta_0\|^2$
Predictive risk	$\mathbb{E}(x_0^\top\hat\beta - x_0^\top\beta)^2$ (out-of-sample)

3.5. Theoretical Analysis of LASSO.

3.5.1. Slow Rates.

- 핵심 결론 : X에 대한 가정 없이도 prediction error 가 $\sqrt{\log d/n}$ 속도로 줄어듦.
- 가정) (2) 에서 $s = \|\beta_0\|_1$.
  - $\|y - X\hat\beta\|_2^2 \le \|y - X\beta_0\|_2^2 = \|\varepsilon\|_2^2$ (* ∵ $s = \|\beta_0\|_1$ 이므로. β_0도 조건 만족).
  - $\|y - X\hat\beta\|_2^2 = \|X\beta_0 + \varepsilon - X\hat\beta\|_2^2$

$$= \|X(\hat\beta - \beta_0)\|_2^2 + \|\varepsilon\|_2^2 - 2\langle \varepsilon, X(\hat\beta - \beta_0)\rangle. \text{ 대입.}$$

$$\Rightarrow \|X\hat\beta - X\beta_0\|_2^2 \le 2\langle \varepsilon, X\hat\beta - X\beta_0\rangle = 2\langle X^\top\varepsilon, \hat\beta - \beta_0\rangle$$

$$\le 4\|\beta_0\|_1 \|X^\top\varepsilon\|_\infty \quad (\because \text{Hölder's Inequality}).$$

  - By maximal ineq. for Gaussians, for $\forall \delta > 0$,

$$\max_{j=1,\dots,d} |X_j^\top\varepsilon| \le \sigma\sqrt{2n\log\left(\frac{2d}{\delta}\right)}. \text{ w.p. at least } 1-\delta.$$

$$\Rightarrow \frac{1}{n}\|X\hat\beta - X\beta_0\|_2^2 \le 4\sigma\|\beta_0\|_1\sqrt{2\log\left(\frac{2d}{\delta}\right)/n} \quad \text{w.p. } \text{〃} \qquad (\Rightarrow \text{Thm}).$$

$$\Leftrightarrow P\left(\frac{1}{n}\|X\hat\beta - X\beta_0\|_2^2 \ge \bigcirc\right) \le \delta.$$

$$\Longrightarrow E\left[\frac{1}{n}\|X\hat\beta - X\beta_0\|_2^2\right] \lesssim \|\beta_0\|_1\sqrt{\frac{\log d}{n}}.$$

※ 유의: lasso-slow-rate-proof.

$S_0\sqrt{\log d/n}$. ↳ slower. (cf. subset : $S_0 \log d/n$.)

Lemma. *Suppose we have random variables $X_1, \dots, X_d$ with $X_j \sim N(0, \sigma^2)$, but X_j's are not necessarily independent. Then*

$$P\left(\max_{j=1,\dots,d}|X_j| \ge t\right) \le 2d\exp\left(-\frac{t^2}{2\sigma^2}\right),$$

or equivalently, with probability $1-\delta$,

$$\max_{j=1,\dots,d}|X_j| \le \sigma\sqrt{2\log(2d/\delta)}$$

Theorem. *Suppose the linear model $Y = X\beta_0 + \epsilon$ with ϵ_i i.i.d. from $N(0,\sigma^2)$. If we choose $s = \|\beta_0\|_1$ as the tuning parameter, then*

$$\frac{1}{n}\|X\hat\beta - X\beta_0\|_2^2 \lesssim \frac{\|\beta_0\|_1\|X^T\epsilon\|_\infty}{n}.$$

Suppose $\|X_j\|_2 = \sqrt{n}$. Then with high probability,

$$\frac{1}{n}\|X\hat\beta - X\beta_0\|_2^2 \lesssim \|\beta_0\|_1\sqrt{\frac{\log d}{n}}.$$

- Out-of-sample risk : $E(x_0^\top\hat\beta - x_0^\top\beta)^2 \lesssim \|\beta_0\|_1^2\sqrt{\frac{\log d}{n}}$.
  ↳ generally higher than In-sample risk.
  ↳ $\|\beta_0\|_1^2$ 단이 큼.
- Oracle Inequality : (X assume linearity of the mean ; $y = \mu(X) + \varepsilon$).
  - It holds that $\langle X^\top(y - X\hat\beta), \tilde\beta - \hat\beta\rangle \le 0$. ($\tilde\beta$: other feasible lasso).

$$\Leftrightarrow \langle \mu(X) - X\hat\beta, X\tilde\beta - X\hat\beta\rangle \le \langle X^\top\varepsilon, \hat\beta - \tilde\beta\rangle.$$

  - Using polarization identity,

$$\|X\hat\beta - \mu(X)\|_2^2 + \|X\hat\beta - X\tilde\beta\|_2^2 \le \|X\tilde\beta - \mu(X)\|_2^2 + 2\langle X^\top\varepsilon, \hat\beta - \tilde\beta\rangle$$

$$\Leftrightarrow \frac{1}{n}\|X\hat\beta - \mu(X)\|_2^2 + \underbrace{\frac{1}{n}\|X\hat\beta - X\tilde\beta\|_2^2}_{A} \le \underbrace{\frac{1}{n}\|X\tilde\beta - \mu(X)\|_2^2}_{B} + 4\sigma t\sqrt{\frac{2\log(2d/\delta)}{n}}$$

w.p. at least $1-\delta$.

$$\Rightarrow \frac{1}{n}\|X\hat\beta - \mu(X)\|_2^2 \le \left\{\inf_{\|\tilde\beta\|_1 \le s} \frac{1}{n}\|X\tilde\beta - \mu(X)\|_2^2\right\} + 4\sigma t\sqrt{\frac{2\log(2d/\delta)}{n}}$$

A, B 차이 inf ⇒ $\mu(X)$가 linear, best 이면 0. ($\mu(X) = X\tilde\beta^{best}$).

3.7.2. Fast Rates.

— Under very strong assumptions → faster rates.

· X satisfies the restricted eigenvalue condition with const. $\phi_0 > 0$,

[if $\frac{1}{n}\|Xv\|_2^2 \geq \phi_0^2 \|v\|_1^2$ for all subsets $J \subseteq \{1,\dots,d\}$ s.t. $|J| = S_0$.
and all $v \in \mathbb{R}^d$ s.t. $\|v_{J^c}\|_1 \leq 3\|v_J\|_1$.]

Theorem. *Suppose $Y = X\beta_0 + \epsilon$ with X fixed, ϵ_i i.i.d. from $N(0,\sigma^2)$, with $\|\beta_0\|_0 = s_0$. Suppose X satisfies the restricted eigenvalue condition in (16). If $\lambda \geq 2\|X^\top \epsilon\|_\infty$, then*

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \frac{s_0\lambda^2}{n^2\phi_0^2},$$

and

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{s_0\lambda^2}{n^2\phi_0^4}.$$

Suppose $\|X_j\|_2 = \sqrt{n}$, and λ is chosen as $\lambda = 2\sigma\sqrt{2n\log(2d/\delta)}$. Then with high probability,

$$\frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \lesssim \frac{\sigma^2 s_0 \log(d)}{n\phi_0^2},$$

and

$$\|\hat{\beta} - \beta_0\|_2^2 \lesssim \frac{\sigma^2 s_0 \log d}{n\phi_0^4}, \tag{17}$$

with probability tending to 1.

pf) $\|y - X\hat{\beta}\|_2^2 + 2\lambda\|\hat{\beta}\|_1 \leq \|y - X\beta_0\|_2^2 + 2\lambda\|\beta_0\|_1$. ($\because \hat{\beta}$: minimizer).

$\Leftrightarrow \|X\hat{\beta} - X\beta_0\|_2^2 \leq 2\langle \epsilon, X(\hat{\beta} - \beta_0)\rangle + 2\lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1)$.

$\leq 2\|X^\top\epsilon\|_\infty \|\hat{\beta} - \beta_0\|_1 + 2\lambda(\;\;)$.

$\leq \lambda\|\hat{\beta} - \beta_0\|_1 + 2\lambda(\;\;)$. ($\because$ cond.)

Let $\hat{\Delta} = \hat{\beta} - \beta_0$ for convenience. Then since $\beta_{0,-S} = 0$,

$$\begin{aligned}\|\beta_0\|_1 - \|\hat{\beta}\|_1 &= \|\beta_{0,S}\|_1 - \|\beta_{0,S} + \hat{\Delta}_S\|_1 - \|\hat{\Delta}_{-S}\|_1 \\ &\leq \|\hat{\Delta}_S\|_1 - \|\hat{\Delta}_{-S}\|_1,\end{aligned}$$

$\leq \lambda(\|\hat{\Delta}_S\|_1 + \|\hat{\Delta}_{-S}\|_1) + 2\lambda(\;\;)$.

$= 3\lambda(\;\;) \geq 0$.

$\Rightarrow \|\hat{\Delta}_{-S}\|_1 \leq 3\|\hat{\Delta}_S\|_1 \Rightarrow$ R.E.C. $\frac{\theta}{2}$ $\hat{\Delta} = \hat{\beta} - \beta_0$에 적용.

$\|\hat{\Delta}_S\|_1 \leq \sqrt{S_0}\,\|\hat{\Delta}_S\|_2 \leq \sqrt{S_0}\,\|\hat{\Delta}\|_2$ 이므로,

(C-S.)

$\|X\hat{\beta} - X\beta_0\|_2^2 \leq 3\lambda\|\hat{\Delta}_S\|_1$

$\leq 3\lambda\sqrt{S_0}\,\|\hat{\beta} - \beta_0\|_2$.

$\leq 3\lambda\sqrt{\frac{S_0}{n\phi_0^2}}\,\|X\hat{\beta} - X\beta_0\|_2$. ← R.E.C.

$\Leftrightarrow \frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{9S_0\lambda^2}{n^2\phi_0^2}$ → $\lambda \geq \|X^\top\epsilon\|_\infty$.

Since $\|X_j\|_2 = \sqrt{n}$, w.p. $1-\delta$, $\|X^\top\epsilon\|_\infty \leq \sigma\sqrt{2n\log(2d/\delta)}$.

Choose $\lambda = 2\sigma\sqrt{2n\log(2d/\delta)}$

$\Rightarrow \frac{1}{n}\|X\hat{\beta} - X\beta_0\|_2^2 \leq \frac{72\sigma^2 S_0 \log(2d/\delta)}{n\phi_0^2}$ (order : $S_0 \log\frac{d}{n}$).

Coefficients) $\|\hat\beta - \beta_0\|_2^2 \le \frac{1}{n\phi_0^2}\|X\hat\beta - X\beta_0\|_2^2$

$\le \frac{9 s_0 \lambda^2}{n^2 \phi_0^4} = \frac{n2\sigma^2 s_0 \log(2d/\delta)}{n\phi_0^4}$

↳ ϕ_0 : eigenvalue 의 LB.
작으면 작을 UB ↑. (≈ singular)

3.4.3. Support Recovery.

– Additional Assumptions:

(mutual coherence) $\|(X_S^T X_S)^{-1} X_S^T X_j\|_1 \le 1-\gamma$. for $j \notin S$

(minimum eigenvalue) for some $C>0$, $\Lambda_{min}\left(\frac{1}{n}X_S^T X_S\right) \ge C$.

(minimum signal) $\beta_{0,min} = \min|\beta_{0,j}| \ge \lambda\|(X_S^T X_S)^{-1}\|_\infty + \frac{4\sigma\lambda}{\sqrt{C}}$.

$\Rightarrow P(\text{support}(\hat\beta) = \text{support}(\beta)) \to 1$.

4. Basis Expansion & Kernel Methods.

4.1. Introduction.

– Basis expansions
- reg : $f(x) = \sum_{j=1}^{q} \beta_j h_j(x)$. $h_j : \mathbb{R}^d \to \mathbb{R}$.
- polynomial reg / regression spline / smoothing spline.

– Kernel Smoothing
- Nadaaya-Watson estimator
- local linear reg.

↳ 모든 estimator는 linear smoother : $\hat f(x) = \sum_i w_i(x) Y_i$ (weights $w_i(x)$는 Y_i와 무관).

$\Rightarrow \hat\mu = (\hat f(x_1), \dots, \hat f(x_n)) = SY$. (for some $S \in \mathbb{R}^{n\times n}$).

↳ h, λ에 따라 달라짐. Y_i와는 무관

$\Rightarrow$ effective df $(\hat\mu) = tr(S)$.

4.2. Basis Expansion

– $\mathcal{F} = \{h(x)^T\beta : \beta \in \mathbb{R}^q\}$. $h : \mathbb{R}^d \to \mathbb{R}^q$. (typically $q \gg d$).
(basis)

4.2.1. Splines.

– set of basis functions.

– k-th order spline : piecewise polynomial ftn of degree k,
has continuous derivatives of order $1, \dots, k-1$ at knots $t_1 \dots t_p$.

– parameterize splines with $t_1 \dots t_p$

→ truncated power basis $g_1 \dots g_{p+k+1}$. $\{ g_1(x) = 1,\ g_2(x) = x,\ \dots\ g_{k+1}(x) = x^k$, $g_{k+1+j}(x) = (x - t_j)_+^k$. $(j = 1, \dots, p)$.

$(p+1)$개 × k차 poly − $(p\times k)$개 조건
$= (p+1)(k+1) - pk = p+k+1$.

4.2.2. Regression Splines

- $f(x) = \sum_{j=1}^{p+k+1} \beta_j g_j(x)$
  - → define basis matrix G by $G_{ij} = g_j(x_i)$.
  - → $\hat{\beta} = \text{argmin}_{\beta \in \mathbb{R}^{p+k+1}} \|y - G\beta\|_2^2 = (G^TG)^{-1}G^Ty$
    $\hat{f}(x) = \sum \hat{\beta}_j g_j(x)$
    $\hat{y} = (\hat{f}(x_1), \dots, \hat{f}(x_n)) = G(G^TG)^{-1}G^Ty$. (linear smoother).

4.2.3. Natural Splines.

- f: poly- degree of k on $[t_1, t_2], \dots, [t_{p-1}, t_p]$
  f: // $\frac{k-1}{2}$ on $(-\infty, t_1]$ and $[t_p, \infty)$
  f is cont & has cont- derivatives of orders $1, \dots, k-1$ at $t_1, \dots, t_p$.
- $\dim$(span of k^{th} order N.S.) $= (k+1)(p-1) + \left(\frac{(k+1)}{2}+1\right) \cdot 2 - pk = p$.

4.2.4. Smoothing Splines

- regularized reg. over natural spline basis. (knots at all inputs $x_1, \dots, x_n$).

motivation) $x_1 \dots x_n \in [0,1]$. smoothing spline estimate of f

→ $\hat{f} = \text{argmin}_f \sum (y - f(x_i))^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx$ $\left(m = \frac{k+1}{2}\right)$.

↳ 곡률 제한 (곡률) penalty. 즉 $f''(x)$

⇒ unique, natural k^{th} order spline w/ knots at $x_1, \dots, x_n$.

pf) basis : n-dimensional. → $z_i = \hat{f}(x_i)$. $i = 1, \dots, n$.

We can always find natural spline f w/ knots at $x_1 \dots x_n$ that satisfies $f(x_i) = z_i$.

Define $h(x) = \tilde{f}(x) - f(x)$.

- $\int_0^1 f''(x) h''(x)\, dx = f''(x)h'(x)\Big|_0^1 - \int_0^1 f'''(x) h'(x)\, dx$ ($f''(0) = f''(1) = 0$. $f'''(x) = 0$ for $x < x_1$ & $x > x_n$.)
  $= -\int_{x_1}^{x_n} f'''(x) h'(x)\, dx$
  $= -\sum_{j=1}^{n-1} f'''(x) h(x)\Big|_{x_j}^{x_{j+1}} + \int_{x_1}^{x_n} f^{(4)}(x) h(x)\, dx$ ($f^{(4)}(x) = 0$.)
  $= -\sum_{j=1}^{n-1} f'''(x_j)\left(h(x_{j+1}) - h(x_j)\right)$ ($h(x_j) = 0$ for all $j = 1, \dots, n$.)
  $= 0$
- $\int_0^1 \tilde{f}''(x)^2\, dx = \int_0^1 (f''(x) + h''(x))^2\, dx$
  $= \int_0^1 f''(x)^2\, dx + \int_0^1 h''(x)^2\, dx + 2\int_0^1 f''(x) h''(x)\, dx$.
  $= \int_0^1 f''(x)^2\, dx + \int_0^1 h''(x)^2\, dx$.
  $\geq \int_0^1 f''(x)^2\, dx$. ($= \Leftrightarrow h''(x) = 0$ for all $x \in [0,1]$.)
  $\Leftrightarrow h$ is linear. $\because h = 0$.

4.2.a. Finite Dimensional form

— Reparameterization) choosing basis $\eta_1, \dots, \eta_n$

$\Rightarrow \hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{n} \beta_j \eta_j(x_i)\right)^2 + \lambda \int_0^1 \left(\sum_{j=1}^{n} \beta_j \eta_j^{(m)}(x)\right)^2 dx.$

(→ finite-dimensional form).

— N (basis matrix), Ω (penalty matrix) $\in \mathbb{R}^{n \times n}$,

$$\begin{cases} N_{ij} = \eta_j(d_i) \\ \Omega_{ij} = \int_0^1 \eta_i^{(m)}(x)\, \eta_j^{(m)}(x)\, dx \end{cases}$$

$\Rightarrow \hat{\beta} = \underset{\beta \in \mathbb{R}^n}{\operatorname{argmin}} \|y - N\beta\|_2^2 + \lambda \beta^T \Omega \beta.$

$= (N^T N + \lambda \Omega)^{-1} N^T y.$ / $\hat{\mu} = N(N^T N + \lambda\Omega)^{-1} N^T y$: linear (in y).

4.2.b. Reinsch Form

— $(*) = N(N^T N + \lambda\Omega)^{-1} N^T y$

$= N\left(N^T \left(I + \lambda \underbrace{(N^T)^{-1} \Omega N^{-1}}_{Q\ (\text{p.s.d.})}\right) N\right)^{-1} N^T y$

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

— S의 eigen decomposition : $S = \sum_{j=1}^{n} \frac{1}{1+\lambda d_j} u_j u_j^T.$ $\left(Q = UDU^T,\ D = \operatorname{diag}(d_1 \dots d_n)\right)$

→ 해석) orthonormal basis $u_1 \dots u_n \in \mathbb{R}^n$ 로 표현가능.

shrinkage : 큰 eigenvalue에 해당하는 eigenvector shrink.

4.3. Kernel Smoothing

— Kernel fun $K : \mathbb{R}^d \to \mathbb{R}$.

$$\begin{cases} \int K(t)\, dt = 1 \\ \int t K(t)\, dt = 0 \\ 0 < \int t^2 K(t)\, dt < \infty. \end{cases}$$

— Nadaraya-Watson estimate (각 점 마다 상수 θ로 locally fit).

: $\hat{f}(x) = \dfrac{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) Y_i}{\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)} = \sum W_i(x) Y_i.$ $= \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)(Y_i - \theta)^2$

4.3.1. Error Analysis

— Kernel Smoothing estimator : universally consistent ($\mathbb{E}\|\hat{f} - f_0\|_2^2 \to 0$ as $n \to \infty$).

+) 추가 내용

Theorem. Suppose that $d = 1$ and that f_0'' is bounded. Also suppose that X has a non-zero, differentiable density p and that the support is unbounded. Then, the risk is

$$R_n = \frac{h_n^4}{4}\left(\int x^2 K(x)dx\right)^2 \int \left(f_0''(x) + 2f_0'(x)\frac{p'(x)}{p(x)}\right)^2 dx + \frac{\sigma^2 \int K^2(x)dx}{nh_n} \int \frac{dx}{p(x)} + o\left(\frac{1}{nh_n}\right) + o(h_n^4)$$

where p is the density of P_X.

const. $= O\left(h_n^4 + \frac{1}{nh_n}\right)$.

→ h_n이 커지면 $h_n^4 \uparrow \Leftrightarrow \frac{1}{nh_n} \downarrow$. $\Rightarrow h_n^4 = \frac{1}{nh_n}$, $h_n = n^{-\frac{1}{5}}$. → $n^{-\frac{2\alpha}{2\alpha+2}}$. α높아지면 local bias/high-order kernel로..

$\Rightarrow R_n = O(n^{-\frac{4}{5}})$. [$d$-차원 : $O(n^{-\frac{4}{2\alpha+d}})$]

→ support의 boundary 근처면 bias order $\begin{cases} O(h) & \text{near boundary} \\ O(h^2) & \text{interior} \end{cases}$ $\Rightarrow R_n = O(h^3)$.

$\Leftrightarrow$ minimax rate over $H_d(\alpha, L)$

: $\inf_{\hat{f}} \sup_{f_0 \in H_d(\alpha,L)} \mathbb{E}_p[(\hat{f}(x) - f_0(x))^2] \gtrsim n^{-\frac{2\alpha}{2\alpha+d}}$.

↳ 어떤 방법 가져와도 kernel reg. 보다 좋지x.

4.7.2. Local Linear Regression.

— local constant fit → local linear fit.

at each input x_i, define $\hat{f}(x) = \hat{\beta}_x$. $\leftarrow \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)(Y_i - \beta)^2$

→ consider $\hat{f}(x) = \hat{\alpha}_x + \hat{\beta}_x^T x \leftarrow \underset{\alpha,\beta}{\operatorname{argmin}} \sum_{i=1}^n K\left(\frac{x-X_i}{h}\right)(Y_i - \alpha - \beta^T X_i)^2$.

$\Leftrightarrow (Y - B\beta)^T \Omega (Y - B\beta)$

$\Rightarrow \hat{f}(x) = \underbrace{b(x)^T}_{w(x)^T} (B^T \Omega B)^{-1} B^T \Omega Y$ where $\begin{cases} b(x) = (1, x) \in \mathbb{R}^{d+1}. \\ B \in \mathbb{R}^{n \times (d+1)} \text{ with } i^{th} \text{ row } b(X_i). \\ \Omega: \text{diag}, \ \Omega_{ii} = K\left(\frac{x-X_i}{h}\right). \end{cases}$

$w(x)^T = \Omega B (B^T \Omega B)^{-1} b(x)$

$\Rightarrow$ linear.

$\hat{m} = (w(x_1)^T Y, \ldots, w(x_n)^T Y) = B(B^T \Omega B)^{-1} B^T \Omega Y = SY$.

— Local fit reduces bias?

$\mathbb{E}[\hat{f}(x)] = \sum_{i=1}^n w_i(x) f_0(X_i)$. ② Taylor expansion of f_0 about x

$= f_0(x) \sum_{i=1}^n w_i(x) + \nabla f_0(x)^T \sum_{i=1}^n (X_i - x) w_i(x) + R$.

$\overset{(*)}{=} f_0(x) + R$.

$\Rightarrow \hat{f}$ is unbiased to first-order. (bias ← 2nd order term에 의존).

— Bias: universal 하게 h^2의 order. (bias 작음).

4.7.3. Higher-Order Smoothness.

— $\hat{f}(x) = \hat{\beta}_{x,0} + \sum_{j=1}^k \hat{\beta}_{x,j} x^j \leftarrow \underset{\beta_0 \ldots \beta_k}{\operatorname{argmin}} \sum K\left(\frac{x-X_i}{h}\right)(Y_i - \beta_0 - \sum \beta_j X_i^j)^2$

$= b(x)(B^T \Omega B)^{-1} B^T \Omega y$.

↳ $(1, x, \ldots, x^k)$.

— Taylor expansion 통해 $\mathbb{E}[(\hat{f}(x) - f_0(x))^2] \lesssim n^{-\frac{2\alpha}{2\alpha+1}}$ 얻을 수 있다.

— higher-order kernel: $\int K(t)dt = 1$, $\int t^j K(t)dt = 0$. $0 < \int t^k K(t)dt < \infty$.

$j = 1, \ldots, k-1$.

↳ 대체로: kernel reg + higher-order kernel.

Smoothing Splines & Local Polynomial.

[2024 mid #3] f_0 : cubic smoothing spline

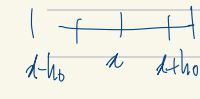
(a) argmin → f_0, f_{CSS} 증명.

(b) $E[\hat{f}^{lp}(x)] = f_0(x)$. : local polynomial은 x_0에서 bias = 0.

pf) $E[\hat{f}^{lp}(x)] = E[\sum w_i(x) y_i]$

$= \sum w_i(x) f_0(x_i)$. × $I(|x-x_i| \le h_0)$.

$[x-h_0, x+h_0] \cap \{t_1, \dots, t_m\} = \emptyset$ 인 $h_0 > 0$ 존재.



f_0 테일러전개 : $f_0(x_i) = \sum_{j=0}^{3} \frac{(x_i-x)^j}{j!} f_0^{(j)}(x)$. $(\because f_0^{(4)} \equiv 0)$.

$E[\hat{f}^{lp}(x)] = \sum w_i(x) \cdot \| \cdot I(|x-x_i| \le h)$.

$= \sum_{j=0}^{3} \frac{f_0^{(j)}(x)}{j!} \sum w_i(x) (x_i - x)^j$

$\left(f_0(x) + f_0'(x) + \frac{1}{2} f_0''(x) + \frac{1}{6} f_0'''(x) \right)$.

$\| \quad 0$.

$= f_0(x)$.

[2024 mid #3] $\hat{f}^{CSS}$ $\forall x, y \in [0,1]$.

(a) $\hat{f}^{CSS} \in H_1(3, L)$ 임을 보이기. $\Leftrightarrow |f^{CSS\prime\prime}(x) - f^{CSS\prime\prime}(y)| \le L(y-x)$.

pf) 각 $[t_i, t_{i+1}]$ 에서, $\hat{f}^{CSS}$는 3차. $\Rightarrow f^{CSS\prime\prime\prime}$ 존재.

$\Rightarrow |f^{CSS\prime\prime\prime}(x)| \le L_2$ for some $x \in (t_i, t_{i+1})$.

$\Rightarrow |f^{CSS\prime\prime}(x)| \le L$ $(L := \max L_2)$.

$\Rightarrow |f''(x) - f''(y)| = \int_y^x f'''(t)\, dt \le L|x-y|$.

→ 모든 $x \le y$ 에 대해.

$|f''(x) - f''(y)| \le | \quad | + \sum | \quad | + | \quad |$.

$\le L(y-x)$

(b) $\mathrm{Bias}\left(\hat{f}^{lp}(x)\right) = \sum W_i(x) \underbrace{\hat{f}^{css}(x_i)}_{f_0} - \hat{f}^{css}(x)$

(1차 Taylor 전개)

$= \sum W_i(x) \left(\sum_{j=0}^{1} \frac{(x_i-x)^j}{j!} \hat{f}^{css(j)}(x) + \frac{(x_i-x)^2}{2!} \hat{f}^{css''}(\xi_{i2}) \right) - f.$

$= \sum W_i(x) \frac{(x_i-x)^2}{2} \left((\hat{f}^{css''})(\xi_{i2}) - (\hat{f}^{css})''(x) \right).$

$\leq L|\xi_{i2}-x| \leq L|x_i-x|.$

$\leq \frac{L}{2!} |W_i(x)| |x_i-x|^3 \; 1(|x_i-x| \leq h).$

$\leq \frac{L}{2!} h^3 |W_i(x)| = C.$

(c) 위의 (b)와 같음.

(d) $\mathrm{Bias} = 0$ on $[t_i+h, t_{i+1}-h]$.

$\int_0^1 \mathrm{Bias}^2 dx \leq \int_{\text{knot}: \cup[t_i-h, t_i+h] \cap [0,1]} (Ch^3)^2 dx$

$\leq \sum_{i=1}^{m} \int_{t_i-h}^{t_i+h} C^2 h^6 dx$

$2h \cdot C^2 h^6 \cdot m.$

[HW2 #4] $\hat{f}: lp. = \sum W_i(x) y_i.$

(a) $\sum W_i(x) = 1, \quad \sum W_i(x)(x_i-x)^j = 0$

(b) $f_0 \in H_1(k+1, L)$ → 국소적인 bias?

$H_1(k+1, L)$ 조건 ⇒ k-테일러 + R.E.

Pf) $E[\hat{f}(x)] = \sum W_i(x) f_0(x_i).$

$\mathrm{Bias} = \sum_{i=1}^{n} W_i(x) \left(\sum_{j=0}^{k+1} \frac{(x_i-x)^j}{j!} f^{(j)}(x) + \frac{(x_i-x)^k}{k!} f^{(k)}(\xi_{i2}) \right) - f_0(x)$

$\left(f(x) + f'(x)(x_i-x) + \cdots \right)$ (*)

$= \cancel{f(x)} + (\;) \cdot 0 + \left(\sum W_i(x) \frac{(x_i-x)^k}{k!} \left(f^{(k)}(\xi_{i2}) - f_0^{(k)}(x) \right) \right) - \cancel{f_0(x)}$

$\leq L|\xi_{i2}-x|.$

$|x_i-x| > h$ 면 0이므로

$$\le \frac{1}{k!} \sum |w_2(x)| \, h^k \cdot L h.$$

$$= \frac{L}{k!} h^{k+1} \sum |w_2(x)| .$$

$$\le \frac{CL}{k!} h^{k+1}$$ 1.

<LASSO>.

$\frac{1}{n}\|Xv\|_2^2 \ge \phi_0^2 \|v\|_2^2$ ($\|v_{J^c}\|_1 \le 3\|v_J\|_1$). ← R. eigenvalue.