〈통계적 기계학습〉

7. RKHS

7.1. Review.

- Goal: find f that minimizes expected prediction error.

$$f^* = \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; \mathbb{E}_{(Y,X) \sim P}\left[\ell(Y, f(X)) \right].$$

- f^* is unknown. → estimated by $\hat{f}$

$$\hat{f} = \underset{f}{\operatorname{argmin}} \; \mathbb{E}_{\mathcal{D}_n}\left[\ell(Y, f(X)) \right] = \underset{f}{\operatorname{argmin}} \; \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)).$$

7.1.1. Hölder & Sobolev Spaces.

- Hölder class $H(\beta, L)$ on $T \subset \mathbb{R}^d$.
  - ↳ $g : T \to \mathbb{R}$ 중 $\ell = \beta - 1$ times differentiable and satisfies $|g^{(\ell)}(y) - g^{(\ell)}(x)| \le L \cdot |y - x|$. $(\alpha = 1)$.
    (bounded derivatives).
- Sobolev class : $\mathcal{S}(\beta, L)$.
  - ↳ $g : T \to \mathbb{R}$ s.t. $\int_x \left(g^{(\beta)}(x) \right)^2 dx \le L^2$

7.2. Hilbert Space.

- norm : mapping $\|\cdot\| : V \to (0, \infty)$
  1. $\|x\| = 0$ iff $x = 0$.
  2. $\|ax\| = |a| \, \|x\|$ $\forall a \in \mathbb{R}$
  3. $\|x + y\| \le \|x\| + \|y\|$
- normed space : $\mathbb{R}^d \to \|x\|_p = \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}$ 중 하나.

- inner product : mapping $\langle \cdot \rangle : V \times V \to \mathbb{R}$
  1. $\langle x, x \rangle \ge 0$ and $\langle x, x \rangle = 0$ iff $x = 0$.
  2. $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
  3. $\langle x, y \rangle = \langle y, x \rangle$.
- inner product defines a norm : $\|v\| = \sqrt{\langle v, v \rangle}$.

- complete : every Cauchy seq. converges to a limit.

complete, normed Banach. $\supset$ complete, inner product Hilbert → ($f_n \to f$ ⊢ 여기 RKHS 이야기 나옴. $\Rightarrow \|f_n - f\| \to 0$ as $n \to \infty$).

- If V is Hilbert space & L is a closed subspace,
  for $\forall v \in V$, $\exists$ unique $y \in L$. (projection of v onto L).
  → y minimizes $\|v - z\|$ over all $z \in L$.
  → v can be written as $v = w + z$. ($w \in L^{\perp}$, z = projection of v).

- orthogonal sum: $L \oplus M = \{l + m : l \in L,\ m \in M\}$.
- every Hilbert space has an orthonormal basis. $\mathcal{E}$.
- Hilbert space is separable if $\exists$ countable orthonormal basis

Thm Let V: separable Hilbert space. → $\{e_1, e_2, \dots\}$ orthonormal basis
then for any $x \in V$, $x = \sum_{j=1}^{\infty} \theta_j e_j$ where $\theta_j = \langle x, e_j \rangle$.
and $\|x\|^2 = \sum_{j=1}^{\infty} \theta_j^2$.

Thm (Riesz Representation theorem).
In a Hilbert space X, for every continuous linear functional $L: X \to \mathbb{R}$,
$\exists$ unique $y \in X$ s.t. $Lx = \langle x, y \rangle$.

11.4. L_p spaces

- f: functions $[a,b] \to \mathbb{R}$.
- L_p norm: $\|f\|_p = \left(\int_a^b |f(x)|^p \, dx \right)^{1/p}$ → $\langle f, g \rangle = \int_a f(x) g(x) \, dx$.

  - C-S: $\left(\int f(x) g(x) \, dx \right)^2 \le \int f^2(x) dx \int g^2(x) dx$.
  - Minkowski: $\|f+g\|_p \le \|f\|_p + \|g\|_p$.
  - Hölder: $\|fg\|_1 \le \|f\|_p \|g\|_q$ $\left(\frac{1}{p} + \frac{1}{q} = 1\right)$.

- Special properties of L_2 → Hilbert space. $\langle f, g \rangle = \int f(x) g(x) dx$.
  ⇒ orthonormal basis $\phi_1, \phi_2, \dots$

  → $f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$
  $\theta_j = \int_a^b f(x) \phi_j(x) dx$.
  $\int_a^b f^2(x) \, dx = \sum_{j=1}^{\infty} \theta_j^2$

  → span of $\{\phi_1 \dots \phi_n\}$ = $\{ \sum a_j \phi_j(x) : a_1 \dots a_n \in \mathbb{R} \}$.

→ projection of $f(x) = \sum_{j=1}^{\infty} \theta_j \phi_j(x)$ onto the span of $\{\phi_1 \dots \phi_n\}$

: $f_n = \sum_{j=1}^{n} \theta_j \phi_j(x)$. (n-term linear approximation of f).

7.5. Sobolev Space.

- weakly differentiable if $\exists f'$ s.t. $\int_x^y f'(s)\,ds = f(y) - f(x)$.
- Sobolev space of order m

: $S_{m,p} = \{ f \in L_p(0,1) : \| D^m f \| \in L_p(0,1) \}$. $(p=2)$

Thm S_m is a Hilbert space under the inner product

$\langle f, g \rangle = \sum_{k=0}^{m-1} f^{(k)}(0)\, g^{(k)}(0) + \int_0^1 f^{(m)}(x)\, g^{(m)}(x)\,dx$.

→ Define $k(x,y) = \sim$.

Then for each $f \in S_m$, $f(y) = \langle f, k(\cdot, y) \rangle$ and $k(x,y) = \langle k(\cdot,x), k(\cdot,y) \rangle$.

⇒ S_m is a reproducing kernel H. S.

7.6. Mercer Kernels and RKHS.

- choose f to minimize $\sum_i (Y_i - f(X_i))^2 + \lambda J(f)$.

$\Leftrightarrow$ $f \in \mathcal{H}$ to minimize $\sum (Y_i - f(X_i))^2$ where $\mathcal{H} = \{f : J(f) \le L\}$.

7.6.2. Evaluational functional.

Def $\mathcal{L}_x : H \to \mathbb{R}$ is defined as $\mathcal{L}_x f = f(x)$.

↳ Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$

Def① H is RKHS if $\mathcal{L}_x$ is continuous for any $x \in \mathcal{X}$.

7.6.3. Reproducing Kernel.

Def $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel of H

if $\begin{cases} \forall x \in \mathcal{X} & K_x = K(\cdot, x) \in \mathcal{H}. \\ \forall x \in \mathcal{X} \text{ and } \forall f \in H, & \langle f, K_x \rangle = f(x). \end{cases}$

→ K is the representer of the evaluational functional.

↳ Riesz Thm : H.S. 에서 bounded linear functional 은 unique 하게 representer 존재.

Def② H is RKHS if it has a reproducing kernel.

7.6.4. Positive Semidefinite function.

Def Symmetric function $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called positive semidefinite if $\forall n \geq 1$, $(a_1 \dots a_n) \in \mathbb{R}^n$, $(x_1 \dots x_n) \in \mathcal{X}^n$, $\sum_{i,j=1}^{n} a_i a_j K(x_i, x_j) = a^T K a \geq 0$.

- every inner product / reproducing kernel $\Rightarrow$ p.s.d.

- given p.s.d. function K, $K_x(y) = K(y, x)$ 고정.
  - $\Rightarrow$ can create fns by linear comb. of the kernel.
  - $\Rightarrow$ $f(x) = \sum \alpha_j K_{x_j}(x)$. $g(x) = \sum \beta_j K_{y_j}(x)$
  - $\Rightarrow$ define inner product $\langle f, g \rangle = \langle f, g \rangle_K = \sum\sum \alpha_i \beta_j K(x_i, y_j)$. (*)
    and norm $\|f\|_K = \sqrt{\langle f, f \rangle_K} = \sqrt{\alpha^T K \alpha}$.

Def ② H_K : completion of H_0 with respect to $\|\cdot\|$. (RKHS generated by K)
↳ verify: $\langle f, f \rangle = 0 \Rightarrow f = 0$.

7.6.5. Mercer Kernels.

Def Mercer kernel $K(x, y)$: symmetric & p.s.d. ($K(x,y) = K(y,x)$; $\iint K(x,y) f(x) f(y) dx dy \geq 0$, $\forall f$)
↳ e.g. Sobolev space의 kernel.

Thm (Mercer's Thm) $\mathcal{X}$: compact, μ: measure on $\mathcal{X}$. $\text{supp}(\mu) = \mathcal{X}$.
Suppose $K: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ cont., symmetric, $\sup K(x,y) < \infty$.
Define $T_K f(x) = \int_{\mathcal{X}} K(x,y) f(y) d\mu(y)$
Suppose T_K is p.s.d. ($\Leftrightarrow \langle f, T_K f \rangle \geq 0$) $\int_{\mathcal{X}}\int_{\mathcal{X}} K(x,y) f(x) f(y) d\mu(x) d\mu(y) \geq 0$.
$\Rightarrow$ $\exists$ countable eigenvalues & eigenfunctions λ_i, Φ_i, i.e.,
$\int_{\mathcal{X}} K(x,y) \Phi_i(y) d\mu(y) = \lambda_i \Phi_i(x)$.

- $K(x,y) = \sum_{i=1}^{\infty} \lambda_i \Phi_i(x) \Phi_i(y)$ $\Rightarrow$ (Mercer thm 이용. 유한차원 spectral thm.)
  $= \langle \{\sqrt{\lambda_i}\Phi_i(x)\}, \{\sqrt{\lambda_i}\Phi_i(y)\} \rangle_{\ell^2(\mathbb{N})}$

7.6.6. Spectral Representation

— Define feature map Φ

: $\Phi(x) = \left(\sqrt{\lambda_1}\,\phi_1(x), \sqrt{\lambda_2}\,\phi_2(x) \dots\right)$.

$\Rightarrow k(x,y) = \langle \Phi(x), \Phi(y) \rangle$

— Expand f in terms of basis $\phi_1, \phi_2, \dots$

: $f(x) = \sum_{j=1}^{\infty} \beta_j \phi_j(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$

— If $f(x) = \sum \alpha_j \phi_j(x)$ and $g(x) = \sum \beta_j \phi_j(x)$, $\langle f, g \rangle_k = \sum_{j=1}^{\infty} \frac{\alpha_j \beta_j}{\lambda_j}$.

↳ since $k_x(\cdot) = k(\cdot, x) = \sum \lambda_i \phi_i(x) \phi_i(\cdot)$

$\langle \phi_i, \phi_j \rangle = 0$, $C: \langle f, k_x \rangle = f(x)$.

$\phi_j(x) = \langle \phi_j, k_x \rangle_k = \langle \phi_j, \sum \lambda_i \phi_i(x) \phi_i(\cdot) \rangle$

$= \lambda_j \phi_j(x) \langle \phi_j, \phi_j \rangle = \lambda_j$.

7.6.8. Representer Theorem

— Let $\hat{f}$ minimize $\ell + g(\|f\|_K^2)$

then $\hat{f}(x) = \sum_{i=1}^{n} \alpha_i k(x_i, x)$.

7.6.9. RKHS Regression.

— Define $\hat{f}$ to minimize $R = \sum (y_i - f(x_i))^2 + \lambda \|f\|_K^2$

↳ By representer thm, $\hat{f} = \sum \alpha_i k(x_i, x)$

$\rightarrow$ $R = \|Y - K\alpha\|^2 + \lambda \alpha^T K \alpha = \langle Y - K\alpha, Y - K\alpha \rangle + \lambda \alpha^T K \alpha$.

$\rightarrow$ minimizer: $\hat{\alpha} = (K + \lambda I)^{-1} Y$.

$\hat{Y} = K\hat{\alpha} = K(K + \lambda I)^{-1} Y = LY$. $\rightarrow$ linear smoother.

7.6.12. Kernel Trick.

— replace $\langle x_i, x_j \rangle$ with $k(x_i, x_j)$

$\Leftrightarrow$ replacing $\langle x_i, x_j \rangle$ with $\langle \Phi(x_i), \Phi(x_j) \rangle = k(x_i, x_j)$.

8. Concentration of Measure.

8.2. Notation

- $Pf = P(f) = \int f(z) dP(z) = E[f(z)]$.
- $P_n f = P_n(f) = \int f(z) dP_n(z) = \frac{1}{n} \sum f(Z_i)$

8.4. Basic Inequalities

① Hoeffding.

(Markov) $P(Z \geq \varepsilon) \leq \frac{E(Z)}{\varepsilon}$

(Chebyshev) $P(|Z-\mu| \geq \varepsilon) = P(|Z-\mu|^2 \geq \varepsilon^2) \leq \frac{E(Z-\mu)^2}{\varepsilon^2} = \frac{\sigma^2}{\varepsilon^2}$.

$\hookrightarrow P(|\bar{Z}_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$.

$\rightarrow$ Chernoff's method: $P(Z \geq \varepsilon) = P(e^Z \geq e^\varepsilon) = P(e^{tZ} \geq e^{t\varepsilon}) \leq e^{-t\varepsilon} E(e^{tZ})$

$P(Z \geq \varepsilon) \leq \inf_{t>0} e^{-t\varepsilon} E(e^{tZ})$.

Lemma $E(e^{tZ}) \leq e^{t^2(b-a)^2/8}$ $(a \leq Z \leq b)$.

(Hoeffding) $P(|\bar{Z}_n - \mu| \geq \varepsilon) \leq 2\exp\left(-\frac{2n\varepsilon^2}{c}\right)$. $(c = n^{-1}\sum (b_i - a_i)^2)$

Corollary w.p. at least $1-\delta$, $|\bar{Z}_n - \mu| \leq \sqrt{\frac{c}{2n} \log\left(\frac{2}{\delta}\right)} = O_p\left(\frac{1}{\sqrt{n}}\right)$.

② McDiarmid

(McDiarmid) Suppose $\sup |f(z_1 \cdots z_i \cdots z_n) - f(z_1 \cdots z_i' \cdots z_n)| \leq c_i$

then, $P(|f(z_1 \cdots z_n) - E(f(z_1 \cdots z_n))| \geq \varepsilon) \leq 2\exp\left(-\frac{2\varepsilon^2}{\sum c_i^2}\right)$.

$\hookrightarrow$ Thm $P\left(\left\|\frac{1}{n}\sum \phi(Z_i) - E[\phi(Z)]\right\| \geq \frac{B}{\sqrt{n}}\left[2 + \sqrt{2\log\left(\frac{1}{\delta}\right)}\right]\right) \leq \delta$.

$\hookrightarrow$ Ex Let $F_n(z) = \frac{1}{n}\sum I(Z_i \leq x)$: empirical cdf.

$f(z_1 \cdots z_n) = \sup_x |F_n(z) - F(z)| = \frac{1}{n}$.

$\Rightarrow P\left(\sup_x |F_n(z) - F(z)| - E\left(\sup_x |F_n(z) - F(z)|\right) \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}$.

③ Gaussian Tail Inequality

Let $X \sim N(0,1)$ $\rightarrow \phi(x) = (2\pi)^{-1/2} e^{-x^2/2}$

$P(X \geq \varepsilon) = \int_\varepsilon^\infty \phi(s) ds \leq \frac{1}{\varepsilon} e^{-\varepsilon^2/2}$

$\rightarrow \quad P(|\bar{X}_n - \mu| \geq \epsilon) \leq e^{-n\epsilon^2/2\sigma^2}$

$\Rightarrow$ X_i 에 대한 제한 (var. bounded) 없음.

④ Bernstein

Lemma $\quad \mathbb{E}(e^{tX}) \leq \exp\left\{ t^2\sigma^2 \left(\frac{e^{tc} - 1 - tc}{(tc)^2} \right) \right\} \quad \left(|X| \leq c \text{ and } E(X) = 0 \right).$

$\rightarrow$ (Bernstein) If $P(|X_i| \leq c) = 1$ and $\mathbb{E}(X_i) = 0$,

$$P(|\bar{X}_n| \geq \epsilon) \leq 2\exp\left\{ -\frac{n\epsilon^2}{2\sigma^2 + 2c\epsilon/3} \right\}.$$

Corollary w.p. at least $1-\delta$, $|\bar{X}_n - \mu| < \sqrt{\frac{2\sigma^2 \log(1/\delta)}{n}} + \frac{2c\log(\frac{1}{\delta})}{3n}$

8.4. Measures of Complexity $\rightarrow$ uniform bnds $P(\sup|f - \mu(f)| \geq \epsilon) < \delta$.

— F 가 finite $\rightarrow$ size of the set.

Lemma $\quad P(\max_i Z_i \geq t) \leq \sum P(Z_i \geq t)$. (maximal ineq).

$$\Rightarrow P(\sup|f(Z_1 \dots Z_n) - \mu(f)| \geq \epsilon) = P(\max |f(Z_1 \dots Z_n) - \mu(f)| \geq \epsilon)$$
$$\leq \sum P(|f_i(Z_1 \dots Z_n) - \mu(f_i)| \geq \epsilon).$$

① Rademacher Complexity

$(\sigma_1 \dots \sigma_n)$, $(f(Z_1) \dots f(Z_n))$ 가 independent 하게 움직임.

— $\sigma_1 \dots \sigma_n$: Rademacher uniform variables. $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.

Def $\quad Rad_n(F, z^n) = \mathbb{E}\left(\sup\left(\frac{1}{n} \sum \sigma_i f(Z_i) \right) \mid Z_1 \dots Z_n \right)$.

$$Rad_n(F) = \mathbb{E}(Rad_n(F, z^n)) = \mathbb{E}\left(\sup\left(\frac{1}{n} \sum \sigma_i f(Z_i) \right) \right).$$

Lemma $\quad$ Let g : Lipschitz coeff. $\Rightarrow Rad_n(g \circ F, z^n) \leq L \, Rad_n(F, z^n)$.

② Shattering Numbers

Def $\quad F$: class of binary fns. $f \in F : Z \rightarrow \{0,1\}$.

$F_{z_1 \dots z_n} = \{ (f(z_1), \dots f(z_n)) : f \in F \}$. : projection of F onto $z_1 \dots z_n$.

Def $\quad s(F, n) = \sup |F_{z_1 \dots z_n}|$

$\Leftrightarrow \quad s(A, n)$: where $F = \{ I_A : A \in A \}$

Thm $\quad Rad_n(\mathcal{F}, Z^n) \le \frac{2\log S(\mathcal{F}, n)}{n}$

② **VC Dimension**

Def $\quad VC(\mathcal{A}^{\mathcal{F}}) = \sup\{n : S(\mathcal{A}^{\mathcal{F}}, n) = 2^n\}$.

↳ $s(\mathcal{F}, n) = 2^n$ for $n < d$ → poly.

Thm $\quad$ suppose $\mathcal{F}$ has finite VC dimension d.

then, $\quad S(\mathcal{F}, n) \le \sum_{i=1}^{d} \binom{n}{i}$.

and for $n \ge d$, $\quad S(\mathcal{F}, n) \le \left(\frac{en}{d}\right)^d$.

+) $\exists C$ s.t. $\quad Rad_n(\mathcal{F}) \le C\sqrt{d/n}$.

8.6. Uniform Bounds

thm $\quad$ for any $t > \sqrt{\varepsilon/n}$,

$$P\left(\sup_{f\in\mathcal{F}} |(P_n - P)f| > t\right) \le 4\, S(\mathcal{F}, 2n)\, e^{-nt^2/8}$$

$\Leftrightarrow$ w.p. at least $1-\delta$, $\quad \sup |P_n(f) - P(f)| \le \sqrt{\frac{8}{n}\log\left(\frac{4\, S(\mathcal{F}, 2n)}{\delta}\right)}$.

Lemma (Symmetrization) $\quad P\left(\sup |(P_n - P)f| > t\right) \le 2P\left(\sup |(P_n - P_n')f| > t/2\right)$.

finite VC-dim : $S(\mathcal{F}, n) \le \left(\frac{en}{d}\right)^d$

$$\sup |P_n(f) - P(f)| \le \sqrt{\frac{8}{n}\left(\log\left(\frac{4}{\delta}\right) + d\log\left(\frac{ne}{d}\right)\right)}.$$

Lemma $\quad \mathbb{E}\left[\sup |(P_n - P)f|\right] \le 2\, Rad_n(\mathcal{F})$

Thm $\quad$ W.p. at least $1-\delta$,

$$\sup |P_n(f) - P(f)| \le 2\, Rad_n(\mathcal{F}) + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}.$$

$$\le 2\, Rad_n(\mathcal{F}, Z^n) + \sqrt{\frac{4}{n}\log\left(\frac{2}{\delta}\right)}.$$

Cor $\quad \sup |P_n(f) - P(f)| \le \sqrt{\frac{8\log S(\mathcal{F}, n)}{n}} + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$.

finite VC dim) $\quad \le 2C\sqrt{\frac{d}{n}} + \sqrt{\frac{1}{2n}\log\left(\frac{2}{\delta}\right)}$.

8.7. Ex. Classification

- h_* minimizes $R(h) = P(Y \neq h(X))$.
- $\hat{h}$ minimizes $\hat{R}_n(h) = \frac{1}{n}\sum I(Y_i \neq h(X_i))$.

↳ from VC theorem, $\sup |\hat{R}(h) - R(h)| \leq \sqrt{\frac{d \log n}{n}}$.

$$\Rightarrow R(\hat{h}) \leq \hat{R}(\hat{h}) + \sqrt{\frac{d\log n}{n}} \leq \hat{R}(h_*) + \sqrt{\frac{d\log n}{n}} \leq R(h_*) + \sqrt{\frac{d\log n}{n}}.$$

8.8. Ex: K-means clustering $C = \{c_1 \dots c_k\}$. $C_* = \{c_1^* \dots c_k^*\}$ → minimizer of $R(C)$.

Def risk $R(C) = E\|X - \pi_C[X]\|^2$ where $\pi_C[x] = \operatorname{argmin}_{c_j} \|x - c_j\|^2$.

Thm Suppose $P(\|X\|^2 \leq B) = 1$ for some B.
Then, $E(R(\hat{C}) - R(C_*)) \leq C\sqrt{\frac{k(d+1)\log n}{n}}$.

pf.

9. Minimax

9.2. Definitions

Def Expected risk : $E_P[d(\hat{\theta}, \theta(P))]$

Maximum risk : $\sup_{P \in \mathcal{P}} E_P[d(\hat{\theta}, \theta(P))]$. → est 에서.

Minimax risk : $R_n = R_n(\mathcal{P}) = \inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P[d(\hat{\theta}, \theta(P))]$

→ maximum risk 최소화하는 estimator의 risk.

↳ θ : function or scalar.

Notation $KL(P_0, P_1) = \int \log\left(\frac{p_0(x)}{p_1(x)}\right) p_0(x)\, dx$.

- Minimax risk의 bound : L_n → any estimator의 maximum risk.
  L_n → Le Cam. ...

9.4. Le Cam.

Thm ① $\inf_{\hat{\theta}} \sup_{P \in \mathcal{P}} E_P[d(\hat{\theta}, \theta(P))] \geq \frac{s}{4}\int [p_0^n(x) \wedge p_1^n(x)]\, dx$

$= \frac{s}{4}[1 - TV(P_0^n, P_1^n)]$.

② $\inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))] \geq \frac{s}{2}\exp(-n\,KL(P_0,P_1))$

$\geq \frac{s}{2}\exp(-n\,\chi^2(P_0,P_1))$

③ $\inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))] \geq \frac{s}{2}\left(1-\frac{1}{2}\int|p_0-p_1|\right)^{2n}$.

Corollary Suppose $\exists\ P_0, P_1$ s.t. $KL(P_0,P_1) \leq \log\frac{2}{n}$.

then $\inf_{\hat\theta} \sup_{P\in\mathcal{P}} \mathbb{E}_P[d(\hat\theta, \theta(P))] \geq \frac{s}{16}$.

Ex $(X_1,Y_1)\dots(X_n,Y_n) \to X_i \sim \mathrm{Unif}(0,1)$, $Y_i = f(X_i)+\varepsilon_i$, $\varepsilon_i \sim N(0,1)$

P : $p(x,y) = p(x)\,p(y|x) = \phi(y-f(x))$.

WLOG, $d=0$ $\Rightarrow$ $\theta = f(0)$. $d(\theta_0,\theta_1) = |\theta_1-\theta_0|$.

Let $f_0(x) = 0$ for all x.

$f_1(x) = \begin{cases} L(\varepsilon - x) & 0 \leq x \leq \varepsilon \\ 0 & x \geq \varepsilon \end{cases}$

$\Rightarrow$ $KL(P_0,P_1) = \int_0^1\int p_0(x,y)\log\left(\frac{P_1(x,y)}{P_0(x,y)}\right)dy\,dx$

$= \int_0^1\int \phi(y)\log\left(\frac{\phi(y)}{\phi(y-f_1(x))}\right)dy\,dx$

$= \int_0^\varepsilon KL(N(0,1), N(f_1(x),1))\,dx$.

$= \frac{L^2}{2}\int_0^\varepsilon(\varepsilon-x)^2dx = \frac{L^2\varepsilon^3}{6}$. ④

5. Ensemble

5.3. Partitions and Trees

- Partitioning range of X. → $\{A_1, \dots, A_N\}$. / $x \in A_j$.

$$\hat{h}(x) = \begin{cases} 1 & \text{if } \sum_{X_i \in A_j} Y_i \ge \sum_{X_i \in A_j} (1-Y_i). \\ 0 & \text{o.w.} \end{cases}$$

$$\hat{f}(x) = \sum_{j=1}^{N} \bar{Y}_j \, I(x \in A_j) = \sum \frac{\sum_{i=1}^{n} Y_i \, I(X_i \in A_j)}{n_j} I(x \in A_j)$$

$n_j = \sum I(X_i \in A_j)$. ($A_j$ 에서 data point 수).

- Tree의 성장 방식: impurity ↓

5.4. Bagging

- Algorithm : Data $\mathcal{L} = \{(x_i, y_i) : i = 1, \dots n\}$

$\{\mathcal{L}^{(b)} : b = 1, \dots, B\}$ → $f(x, \mathcal{L}^{(b)})$. learned.

- Explanation

$f_A(x, P) = E_{\mathcal{L}} f(x, \mathcal{L})$ → aggregated predictor.

- avg. pred error in $f(x, \mathcal{L})$ $\quad e = E_{\mathcal{L}} E_{X,Y} (Y - f(X, \mathcal{L}))^2$
- avg. 〃 in f_A $\quad e_A = E_{X,Y} (Y - f_A(X, P))^2$.

→ $e \ge e_A$.

- $e - e_A$ depends on $f_A^2(x, P) = (E_{\mathcal{L}} f(x, \mathcal{L}))^2 \le E_{\mathcal{L}} f^2(x, \mathcal{L})$. 의 차이. ($e_A$, e ↑)

↳ $f(x, \mathcal{L})$이 $\mathcal{L}$에 대해 stable ⇒ $e \simeq e_A$.

↳ f_A always improves f. (unstable 할수록. improve 多).

- f_B는 f_A의 estimator ($\because$ $\mathcal{L} \simeq P$.)

↔ Bagging은 unstable learner 에서 잘 작동.

- Analysis) 언제 $Var(f(X, \mathcal{L}))$ large?
  - $f(x, \mathcal{L}) = I(\bar{Y}_n \le x)$ 라고 하자. ($\mathcal{L} = \{Y_1, \dots, Y_n\}$).

$E[Y_i] = \mu$, $Var[Y_i] = 1$.

x is close to μ relative to n $\quad x = x_n = \mu + c/\sqrt{n}$.

$F(x, L) = I(\sqrt{n}(\bar{Y}_n - \mu) \le c) \simeq I(Z \le c)$.

→ limiting mean : $\Phi(c)$, variance : $\Phi(c)(1-\Phi(c))$.

— Bootstrap $\bar{Y}^* \sim N(\bar{Y}, \frac{1}{n})$.

$\Leftrightarrow \sqrt{n}(\bar{Y}^* - \bar{Y}) \mid Y_1 \dots Y_n \xrightarrow{d} N(0,1)$.

— $F_B(x) = \mathbb{P}^*[I(\bar{Y}^* \le x)] \simeq \mathbb{P}^*[I(\sqrt{n}(\bar{Y}^* - \bar{Y}) \le \sqrt{n}(x - \bar{Y}))]$.

$\simeq \Phi(\sqrt{n}(x - \bar{Y})) \simeq \Phi(c + Z)$.

↳ $\sqrt{n}(x - \mu) + \sqrt{n}(\mu - \bar{Y})$.

c. $N(0,1)$

⇒ $F(x, L) \simeq I(Z \le c)$, $F_B \simeq \Phi(c + Z)$.

↳ 0 or 1. ↳ smoothed ver.

$(c=0)$ → Ber, mean $\frac{1}{2}$, var $\frac{1}{4}$. → $\Phi(Z) \sim \mathrm{Unif}(0,1)$, mean $\frac{1}{2}$, var $\frac{1}{12}$.

5.5. Random Forest

5.5.1. Theories.

— Suppose each tree $T_k(x) = T(x; \theta_k)$. $\theta_k \overset{iid}{\sim} \pi(\theta)$.

— Let $mg(x, y) = \frac{1}{n}\sum I(T_k(x) = y) - \frac{1}{n}\sum I(T_k(x) = -y)$.

↳ $y \in \{-1, 1\}$.

$PE^*_n = P_{(X,Y)}(mg(X,Y) < 0)$.

$PE^*_n \longrightarrow PE^* = P_{(X,Y)}\{P_\theta(T(X;\theta) = Y) - P_\theta(T(X;\theta) = -Y) < 0\}$

SLLN.

We will derive the upper bound of PE^* as a function of the correlation and strength.

Let

$$mr(X,Y) = P_\theta(T(X:\theta) = Y) - P_\theta(T(X:\theta) = -Y)$$

Then the strength of a single tree is

$$s = \mathbb{E}_{(X,Y)}(mr(X,Y)).$$

Chebychev's inequality gives

$$PE^* \le Var(mr)/s^2.$$

Now we calculate $Var(mr)$. $mr(X,Y)$ can be expanded as

$$\begin{aligned} mr(X,Y) &= \mathbb{E}_\theta(I(T(X:\theta) = Y) - I(T(X:\theta) = -Y)) \\ &= \mathbb{E}_\theta(rmg(\theta, X, Y)), \end{aligned}$$

where $rmg(\theta, X, Y) = I(T(X:\theta) = Y) - I(T(X:\theta) = -Y)$. Note that

$$mr(X,Y)^2 = \mathbb{E}_{\theta,\theta'} rmg(\theta, X, Y) rmg(\theta', X, Y),$$

where θ and θ' are independent and $\theta, \theta' \sim \pi(\theta)$. Then

$$\begin{aligned} Var(mr) &= \mathbb{E}_{\theta,\theta'} cov(rmg(\theta, X, Y), rmg(\theta', X, Y)) \\ &= \mathbb{E}_{\theta,\theta'}\left(\rho(\theta,\theta') sd(\theta) sd(\theta')\right), \end{aligned}$$

where

$$\rho(\theta, \theta') = \mathrm{corr}_{(X,Y)}\left(rmg(\theta, X, Y), rmg(\theta', X, Y) | \theta, \theta'\right)$$

and $sd(\theta) = \mathrm{std}_{(X,Y)}(rmg(\theta, X, Y)|\theta)$. Then

$$Var(mr) = \bar{\rho}\left(\mathbb{E}_\theta(sd(\theta))\right)^2 \le \bar{\rho}\mathbb{E}_\theta(sd^2(\theta))$$

where

$$\bar{\rho} = \mathbb{E}_{\theta,\theta'}\left(\rho(\theta,\theta') sd(\theta) sd(\theta')\right) / \mathbb{E}_{\theta,\theta'}(sd(\theta) sd(\theta')).$$

Write

$$\begin{aligned} \mathbb{E}_\theta(sd^2(\theta)) &\le \mathbb{E}_\theta(\mathbb{E}_{(X,Y)}(rmg(\theta, X, Y)^2)) - \mathbb{E}_\theta(\mathbb{E}_{(X,Y)}(rmg(\theta, X, Y)))^2 \\ &\le 1 - s^2. \end{aligned}$$

To sum up, we have

$$PE^* \le \bar{\rho}(1 - s^2)/s^2.$$

It shows that the two ingredients involved in the generalization error for random forests are the strength of the individual classifiers in the forest, and the correlation between them in terms of the raw margin functions. It is the first of its kind to explain that an important ingredient in success of ensemble methods is diversity of base learners. For example, random perturbation can be understood as a tool of increasing the diversity.

One example of a simplified tree is the *centered forest* is defined as follows. Suppose the data are on $[0,1]^d$. Choose a random feature, split in the center. Repeat until there are k leaves. This defines one tree. Now we average M such trees. Breiman (2004) and Biau (2012) proved the following.

Theorem. *If each feature is selected with probability $1/d$, $k = o(n)$ and $k \to \infty$ then*

$$\mathbb{E}[\hat{f}(X) - f^0(X)|^2] \to 0$$

as $n \to \infty$.

Under stronger assumptions we can say more:

Theorem. *Suppose that f_0 is Lipschitz and that f_0 only depends on a subset S of the features and that the probability of selecting $j \in S$ is $(1/S)(1 + o(1))$. Then*

$$\mathbb{E}|\hat{f}(X) - f^0(X)|^2 = O\left(\frac{1}{n}\right)^{\frac{3}{4|S|\log 2 + 3}}.$$

This is better than the usual Lipschitz rate $n^{-2/(d+2)}$ if $|S| \le d/2$. But the condition that we select relevant variables with high probability is very strong and proving that this holds is a research problem.

A significant step forward was made by Scornet, Biau, and Vert (2015). Here is their result.

Theorem. *Suppose that $Y = \sum_j f_j(X(j)) + \epsilon$ where $X \sim \mathrm{Uniform}[0,1]^d$, $\epsilon \sim N(0, \sigma^2)$ and each f_j is continuous, with $f^0(X) = \sum_j f_j(X(j))$. Assume that the split is chosen using the maximum drop in sums of squares. Let t_n be the number of leaves on each tree and let a_n be the subsample size. If $t_n \to \infty$, $a_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$ then*

$$\mathbb{E}[\hat{f}(X) - f^0(X)|^2] \to 0$$

as $n \to \infty$.

Again, the theorem has strong assumptions but it does allow a greedy split selection. Scornet, Biau, and Vert (2015) provide another interesting result. Suppose that (i) there is a subset S of relevant features, (ii) $p = d$, (iii) m_j is not constant on any interval for $j \in S$. Then with high probability, we always split only on relevant variables.

5.5.4. Variable Importance — $\hat{f}$: random forest est. / $X(j)$의 상대 중요도?

[1] LOCO

$$\hat{\Delta}_j = \frac{1}{m} \sum_{i \in H} W_i \qquad (H:\ \text{hold-out},\ \text{size } m).$$

$$W_i = (Y_i - \hat{f}_{(-j)}(X_i))^2 - (Y_i - \hat{f}(X_i))^2$$

→ Δ_j : prediction risk inflation of estimate.

— $\mathbb{E}[\hat{\Delta}_j \mid \mathcal{T}] = \mathbb{E}[(Y - \hat{f}_{(-j)}(X))^2 - (Y - \hat{f}(X))^2 \mid \mathcal{T}] = \Delta_j.$

[2] Permutation Importance

— permute $X(j)$ for the OOB.

- O : OOB obs. for tree k.
- O^* : OOB 〃 with $X(j)$ permuted.

— $\hat{\Gamma}_j = \frac{1}{M} \sum_{t=1}^{M} W_{tj}$

where $W_{tj} = \frac{1}{m_t} \sum_{i \in O_t^*} (Y_i - \hat{f}_t(X_i))^2 - \frac{1}{m} \sum_{i \in O_t} (Y_i - \hat{f}_t(X_i))^2$

Estimating → $\Gamma_j = \mathbb{E}[(Y - \hat{f}(X_j'))^2] - \mathbb{E}[(Y - \hat{f}(X))^2].$

→ (조건 하에서 $\Gamma_j = 2\left(\frac{\tau}{1-c+dc}\right)^2$).

5.6. Boosting.

— combining weak learners.

— Let $Z_i = (X_i, Y_i)$, $Y_i \in \{-1, +1\}$

— weak learning 가정 : for some $\gamma > 0$, $\exists$ algorithm that returns $h \in \mathcal{H}$ s.t. for all P, $P(R(h) \le \frac{1}{2} - \gamma) \ge 1 - \delta$

— Algorithm :

1. Set $D_1(i) = \frac{1}{n}$. → data 비중의 weight.
2. Repeat for $t = 1, \dots, T$
   - (a) Let $h_t \leftarrow \arg\min_{h \in H} P_{D_t}(Y_i \ne h(X_i))$.
   - (b) $\varepsilon_t = P_{D_t}(Y_i \ne h_t(X_i))$. $= \sum D_t(i)\, \mathbb{1}(Y_i \ne h(X_i))$.
   - (c) $\alpha_t = \frac{1}{2} \log \frac{1-\varepsilon_t}{\varepsilon_t}$. → $\varepsilon_t \to 0$: error ↓. $\alpha_t \to \infty$. / $\varepsilon_t \to 1$: error ↑. $\alpha_t \to -\infty$.
   - (d) Let $D_{t+1}(i) = \frac{D_t(i)\, e^{-Y_i \alpha_t h_t(X_i)}}{Z_t}$ (Y_i: ±1)
3. Set $g(x) = \sum_t \alpha_t h_t(x)$.
4. Return $h(x) = \text{sign}\, g(x)$.

5.6.2. Training Error.

Lemma $\quad Z_t = 2\sqrt{\varepsilon_t(1-\varepsilon_t)} = \sum_i D_t(i)\, e^{-\alpha_t \sum_i h_t(x_i)}$.

Thm Suppose that $\gamma \le \frac{1}{2} - \varepsilon_t \quad \forall t$.
Then, $\hat{R}(h) \le e^{-2\gamma^2 T}$ $\quad$ (error $\to 0$).

5.6.3. Generalization Error.

— $\mathcal{F}_t = \{ \operatorname{sgn}(\sum_t \alpha_t h_t) : \alpha_t \in \mathbb{R},\ h_t \in \mathcal{H} \}$. $\to$ VC dimension T.

$\Rightarrow$ shattering number. $\left(\frac{en}{T}\right)^T$.

— If $\mathcal{H}$ is finite, $\left(\frac{en}{T}\right)^T \cdot |H|^T$.

infinite & VC dimension d, $\left(\frac{en}{T}\right)^T \left(\frac{en}{d}\right)^{dT} \lesssim n^{Td}$.

— By the VC thm, w.p. at least $1-\delta$,

$\mathbb{E}[I(Y \ne \hat{h}(X))] = R(\hat{h}) \lesssim \hat{R}(h) + \sqrt{\frac{T d \log n}{n}}$. $\longrightarrow$ T에 dep. 하지 않게 하면...

↓ n, d에만 의존.

Thm w.p. at least $1-\delta$, $\quad R(g) \le \hat{R}_\rho(g/\|\alpha\|_1) + \frac{1}{\rho}\sqrt{\frac{2d\log(en/d)}{n}} + \sqrt{\frac{2\log(2/\delta)}{n}}$.

Thm $\quad \hat{R}(g/\|\alpha\|_1) \le \prod_{t=1}^{T} \sqrt{4\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}} \le b^T$.

$\sqrt{4\varepsilon_t^{1-\rho}(1-\varepsilon_t)^{1+\rho}} < 1$.

(assuming $\gamma \le \frac{1}{2} - \varepsilon_t$).

6. Model Selection and Assessment.

— $\hat{f} \simeq \underset{f \in \mathcal{F}}{\operatorname{argmin}}\ \mathbb{E}[(Y - f(X))^2] = f_0^*$ $\quad (\ne f \text{ s.t. } Y = f(X))$.

↳ $\mathcal{F}_1, \mathcal{F}_2$ 中 무엇이 더 좋나?

6.2. Introduction

— expected prediction error : $\mathbb{E}[(Y - f(X))^2]$

$\to$ minimized at $f(x) = \mathbb{E}(Y \mid X = x)$. : true regression function.

— training sample 주어지면 $\hat{f}$ construct.

expected prediction (test) error : $\mathbb{E}[(Y - \hat{f}(X))^2]$.

↳ 1) assessment

2) selection.

— Goal
- Prediction : cross-validation. AIC. (C_p).
- selecting true : BIC.

6.3. Prediction and B-V. tradeoff.

— expected test error

: $\mathbb{E}[(Y-\hat{f}(x_0))^2 \mid X=x_0] = \sigma^2 + \mathbb{E}[(f(x_0)-\hat{f}(x_0))^2]$.

(training dataset randomness. $\mathbb{E}[Y|X]$. $\mathbb{E}$ → Risk $(\hat{f}(x_0))$)

$= \sigma^2 + (f(x_0)-\mathbb{E}[\hat{f}(x_0)])^2 + \mathbb{E}[(\hat{f}(x_0)-\mathbb{E}[\hat{f}(x_0)])^2]$.

($(f(x_0)-\mathbb{E}[\hat{f}(x_0)])^2$ ← $\mathrm{Bias}^2(\hat{f}(x_0))$, $\mathbb{E}[(\hat{f}(x_0)-\mathbb{E}[\hat{f}(x_0)])^2]$ ← $\mathrm{Var}(\hat{f}(x_0))$)

⇒ unbiased 여도 큰 error 발생 가능.

6.4. Training Error and Optimism.

— How to estimate E.P.E? $\mathbb{E}[(Y-\hat{f}(X))^2] = R(\hat{f})$.

↳ test data $(x_1', y_1')\cdots$ 상태로 $\frac{1}{m}\sum (y_i'-\hat{f}(x_i'))^2$ 로 추…

— $(x_1', y_1')\cdots(x_m', y_m') \overset{d}{=} (x_1, y_1), \ldots, (x_n, y_n)$ 와 같음.

($y_i = f(x_i)+\varepsilon_i$, $y_i' = f(x_i') + \varepsilon_i'$).

→ Expected Test Error $= \frac{1}{n}\mathbb{E}\|y'-\hat{f}\|_2^2$

$= \frac{1}{n}\mathbb{E}\|y'-f+f-\hat{f}\|_2^2$

$= \underbrace{\frac{1}{n}\mathbb{E}\|y-\hat{f}\|_2^2}_{\text{expected training error}} + \underbrace{\frac{2}{n}\mathrm{tr}(\mathrm{Cov}(y,\hat{f}))}_{\text{양수. (= optimism)}}$.

→ $\frac{2}{n}\mathrm{tr}(\mathrm{Cov}(y,\hat{f})) = \frac{2}{n}\sum \mathrm{Cov}(y_i, \hat{f}(x_i))$.

6.5. D.f and Unbiased Risk Estimation.

— $df(\hat{f}) = \frac{1}{\sigma^2}\sum \mathrm{Cov}(y_i, \hat{f}(x_i))$ ⇒ optimism $= \frac{2\sigma^2}{n} df(\hat{f})$.

• # of params. used by the fit $\hat{f}$. (complexity of $\hat{f}$).

e.g. $\hat{f}(x_i) = \bar{y}$, $df(\hat{f}) = 1$.

— $\hat{Err} = \frac{1}{n}\|y-\hat{f}\|_2^2 + \frac{2\sigma^2}{n} df(\hat{f})$: unbiased estimate for the expected test error.

⇝ $\hat{R} = \hat{Err} - \sigma^2$ → $\mathbb{E}[\hat{R}] = \frac{1}{n}\|f-\hat{f}\|_2^2$

— (model selection) $\hat{f}_\theta$ 로 fitting.

→ $\hat{\theta} = \mathrm{argmin}\ \frac{1}{n}\|y-\hat{f}_\theta\|_2^2 + \frac{2\sigma^2}{n} df(\hat{f}_\theta)$.

(추가 optimistic 하게 해줌?)

6.6. CV

- estimate $\mathbb{E}[(Y-\hat{f}(X))^2]$ w/ minimal assumptions. $((x_i, y_i) \sim \text{iid})$
- split training set into K folds. $(F_1 \dots F_K)$
  k^{th} fold 제외하고 fit. $\rightarrow \hat{f}^{-(k)}$
  evaluate errors on the k^{th} fold. $\rightarrow CV_k(\hat{f}^{-(k)}) = \frac{1}{n_k}\sum_{i\in F_k}(y_i - \hat{f}^{-(k)}(x_i))^2$.
  average $\rightarrow CV(\hat{f}) = \frac{1}{K}\sum CV_k(\hat{f}^{-(k)})$. $(K=n : \text{LOOCV})$.
  - K 의 선택 : bias-variance trade off.

- $\text{Var}(CV(\hat{f})) \approx \text{Var}\left(\frac{1}{K}\sum CV_k(\hat{f}^{-(k)})\right) \hat{=} \frac{1}{K}\text{Var}(CV_1(\hat{f}^{-(1)}))$.
  valid for small K.

$\rightarrow$ $\frac{1}{K}\sum(CV_k(\hat{f}^{-(k)}) - CV(\hat{f}))^2$ 로 estimate 가능 $(SE(\hat{\theta}))^2$.

6.7. AIC

- KL divergence) p: true density ; $p_j = p(\cdot;\hat{\theta}_j)$ (MLE) : estimated density.
  $\rightarrow \hat{j} = \underset{j}{\text{argmin}}\, KL(p, p_j)$
  minimizing $KL(p, p_j) \Leftrightarrow$ maximizing $R_j = \int p(x)\log p_j(x)\,dx$.
  $CV(\hat{f}) = \frac{1}{K}\sum\frac{1}{n_k}\sum_{i\in F_k}\log p(x_i;\hat{\theta}^{-(k)}_j)$
- $\hat{R}_j = \frac{1}{n}\sum\log p(x_i;\hat{\theta}_j) := \frac{1}{n}\ell_j(\hat{\theta}_j)$: log likelihood 의 평균.
  ↓
  adjust bias : $\hat{R}_j = \frac{1}{n}[\ell_j(\hat{\theta}_j) - d_j]$.
  $AIC_j = 2\ell_j(\hat{\theta}_j) - 2d_j$

6.8. BIC

- $BIC = 2\ell_j(\hat{\theta}_j) - d_j\log n$. ($\frac{d_j}{n}$ 가 true model 보다 smallest 하면서 p. $\Rightarrow$ $\hat{j} \to j$)

6.9. Linear Smoothers

- $\hat{f}(x) = \sum w_i(x)y_i = w(x)^T y$.
  $\hat{\mu} = Sy$. $\rightarrow$ $df(\hat{f}) = \frac{1}{\sigma^2}\text{tr}(Cov(y,\hat{y})) = \text{tr}(S)$.
- Cross Validation : (linear smoother 의 경우, $\hat{\mu} = Sy$)
  $CV(\hat{f}) = \frac{1}{n}\sum(y_i - \underbrace{\hat{f}^{-(i)}(x_i)}_{\|}\,)^2 = \frac{1}{n}\sum\left(\frac{y_i - \hat{f}(x_i)}{1-S_{ii}}\right)^2$.
  $\frac{1}{(1-S_{ii})}(\hat{f}(x_i) - S_{ii}y_i)$.