기타

소볼레프 공간 : $f \in H^s(\mathbb{R}^d) \iff \int_{\mathbb{R}^d} \overbrace{\underbrace{|\hat{f}(\xi)|^2}_{\text{f 안에 ξ 성분이 얼마나 많은가?}} \underbrace{(1+|\xi|^2)^s}_{\text{ξ에 대한 penalty}}}^{\text{거칠기 점수}} \, d\xi < \infty$

"H^s에 가입하려면,

고주파 성분을 많이 가져서는 안된다"

- $B^s(\Omega) = \{ f : \Omega \to \mathbb{R} \ : \ \|f\|_{B^s(\Omega)} < \infty \}$

- $\|f\|_{B^s(\Omega)} := \inf_{\underbrace{f_e|\Omega = f}_{}} \left[\int_{\mathbb{R}^d} \overbrace{(1+|\xi|)^s}^{\text{거칠기 점수}} |\hat{f_e}| \, d\xi \right]$

f를 Ω 밖으로

확장한 f_e (푸리에 변환)

머신러닝의 목표 : 일반화 (generalization)

- 학습 데이터 $\{x_i, y_i\}_{i=1}^{n}$
- 손실 함수 $\ell(y, f(x))$

$$\begin{cases} \text{일반화 오류} : E_{(Y,X)\sim P}\left[\ell(Y, f(X))\right] \\ \text{경험적 오류} : \frac{1}{n}\sum_{i=1}^{n} \ell(Y_i, f(X_i)) \end{cases}$$

generalization error

Estimation Error　　Approximation Error

$$\text{Total error} \approx O\left(\frac{1}{\sqrt{n}}\right) + O\left(\frac{1}{m}\right)$$

데이터 n개　　뉴런 m개

<Estimation Error>

- Rademacher Complexity

- 라데마커 확률변수 $P(\sigma_i = 1) = P(\sigma_i = -1) = 50\%$ (노이즈)

- 라데마커 복잡도 $$Rad_n(\mathcal{F}) = E\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\underbrace{\sigma_i f(z_i)}_{\text{노이즈를 따라가는 정도.}}\right)\right] \propto \text{과적합 위험.}$$

---

let $$F_{m,\sigma} = \left\{ f_\theta(x) = \sum_{j=1}^{m} \underbrace{\beta_j\, \sigma(w_j^T x)}_{NN} \right\}$$

then $$Rad(F_{m,\sigma,Q}) \lesssim \frac{Q}{\sqrt{n}}$$

Q ← 파라미터 크기 제한

$\sqrt{n}$ ← 데이터 샘플 갯수

UAT의 의의 : NN은 $C(I_n)$에서 dense 하다.

∴ 한계 : NN의 width는 얼마여야 하는가?

Curse of dimensionality!

· L^p norm : $\|f\|_{L^p} = \left(\int_X |f(x)|^p d\mu\right)^{1/p}$

$= \left(\int_\Omega |f(x)|^p dx\right)^{1/p}$ where $\Omega \subseteq \mathbb{R}^d$

오픈닫 .

· 푸리에 변환 : $\hat{f}(\xi) = \int_{\mathbb{R}^d} f(x) \cdot e^{-2\pi i \langle \xi, x \rangle} dx$

· 역푸리에 변환 : $f(x) = \int_{\mathbb{R}^d} \hat{f}(\xi)\, e^{2\pi i \langle \xi, x \rangle} d\xi$

· 바런 클래스 : $\Gamma_{C_1} = \{ f \in L^1(\mathbb{R}^d) : \|\hat{f}\|_{L^1(\mathbb{R}^d)} < \infty$

and $\int_{\mathbb{R}^d} |2\pi\xi| |\hat{f}(\xi)| d\xi < C_1 \}$

$L^1(\mathbb{R}^d)$

: 함수의 절대값을 $\mathbb{R}^d$ 에 대해 적분값들 때 유한한 값을 가진다.

· 2계층 신경망 : $\Phi(x) = \sum_{i=1}^{N} a_i\, \sigma(w_i^T x_i + b_i)$

정리 Sigmoid σ,

$f \in \Gamma_{C_1}$, $c > 4C_1^2$, $N \in \mathbb{N}$

B_1^d : 반지름 1짜리 구 $\subseteq \mathbb{R}^d$

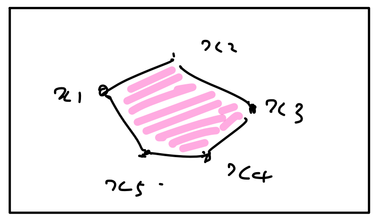
then 有 Φ_{NN} s.t. $\frac{1}{|B_1^d|} \int_{B_1^d} |f(x) - \Phi(x)|^2 \, dx \le \frac{c}{N}$ ⟵ d와는 무관하다!

보조정리 ①

- H : Hilbert Space
- $G \subseteq H$: 어떤 $B > 0$에 대해 모든 $g \in G$가 $\|g\|_H \le B$인 집합
- $f \in \overline{co}(G)$ (closure of the convex hull of G)
- $c > B^2$
- then 有 $(g_i)_{i=1\ldots}^N \subseteq G$ s.t. $\left\| f - \frac{1}{N}\sum_{i=1}^N g_i \right\|_H^2 \le \frac{c}{N}$

$$co(G) = \left\{ \sum_{j=1}^n \alpha_j x_j : n \in \mathbb{N},\ x_j \in G,\ \alpha_j > 0,\ \sum_{j=1}^n \alpha_j = 1 \right\}$$

$\|g\|_H \le B$



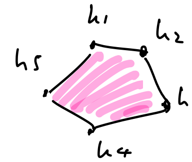
<보조정리 1 증명>

step 1

- since $f \in \overline{co(G)}$
- if f is limit point, $\forall \varepsilon > 0$

  有 $f^* \in co(G)$ s.t $\|f - f^*\|_H < \varepsilon$

$$f^* = \sum_{j=1}^{m} \alpha_j \cdot h_j$$



$\alpha_j > 0 \quad \sum \alpha_j = 1, \quad h_i \in G$

---

step 2 claim

$\exists \{\theta_1, \ldots, \theta_N\} \in \{h_1, \ldots, h_m\}$ s.t $\left\| f^* - \frac{1}{N}\sum_{j=1}^{N} \theta_j \right\|_H^2 \le \frac{B^2}{N}$

---

step 3 let $P(X_i = h_j) = \alpha_j$

then
$$E[X_i] = \sum_{j=1}^{m} P(X_i = h_j) \cdot h_j$$
$$= \sum_{j=1}^{m} \alpha_j h_j = f^*$$

---

step 4

- $E\left[\| f^* - \frac{1}{N}\sum_{i=1}^{N} X_i \|^2\right]$

$$= \frac{1}{N^2} E\left[\| \sum_{i=1}^{N} (f^* - X_i) \|^2\right]$$

$$= \frac{1}{N^2} E\left[\sum_{i=1}^{N} \| f^* - X_i \|^2 + \sum_{i \ne j} \langle \cancel{f^* - X_i}, \cancel{f^* - X_j} \rangle \right]$$

$\begin{cases} X_i, X_j\text{는 독립} \\ E[X_i] = f^* \end{cases}$

$$= \frac{1}{N^2} E\left[\sum_{i=1}^{N} \| f^* - X_i \|^2 \right]$$

$$= \frac{1}{N^2} \cdot N \cdot E\left[\cancel{\|f^*\|^2} - 2 \cdot \langle \cancel{f^*, X_i} \rangle + \|X_i\|^2 \right] = \frac{1}{N}\left(E[\|X_i\|^2] - \|f^*\|^2 \right)$$

$$\le \frac{B^2}{N}$$

- $E[\|X_i\|^2] = \sum_{j=1}^{m} P(X_i = h_j) \cdot \|h_j\|^2$

$$= \sum_{j=1}^{m} \alpha_j \|h_j\|^2$$

$$\le \sum_{j=1}^{m} \alpha_j B^2 \le B^2$$

<정리 증명>

$$G_{C_1} = \left\{ x \mapsto \mathbb{1}_{\mathbb{R}_+}(\langle a, x\rangle + b) \;:\; a \in \mathbb{R}^d,\ b \in \mathbb{R},\ \|a\| \le 2C_1 \right\}$$

$\begin{cases} +1 & : \text{양수} \\ 0 & : \text{음수} \end{cases}$

Lemma 2. $f \in \Gamma_{C_1}$ 이면, $f(x) - f(0) \in \overline{co(G_{C_1})}$ 이다.

$$f(x) - f(0) = \int_{\mathbb{R}^d} \hat{f}(\xi)\left(e^{2\pi i \langle \xi, x\rangle} - 1\right) d\xi$$

$$= \int_{\mathbb{R}^d} |\hat{f}(\xi)| \left\{ \cos\left[2\pi\langle x, \xi\rangle + k(\xi)\right] - \cos k(\xi) \right\} d\xi \qquad (\text{실수부})$$

"무한히 많은 코사인의 가중합" . . . .

- $f \in \Gamma_{C_1}$ 이면 $f(x) - f(0) \in \overline{co}(G_{C_1})$ 이다.
- $\begin{cases} H & : L^2(B_1^d) \\ G & : G_{C_1} \\ \text{유계} & : \|g\|_{L^2} \le 2C_1 \quad (g \in G_{C_1}) \end{cases}$

보조정리 ①

$\Downarrow$

- $$\frac{1}{|B_1^d|} \int_{B_1^d} \Big| (f(x) - f(0)) - \underbrace{\boxed{\sum_{i=1}^{N} w_i \mathbb{1}_{\mathbb{R}^+}(\cancel{\langle a_i, x\rangle} + b_i) \quad w_i\,\sigma(\langle \tilde{a}_i, x\rangle + \tilde{b}_i)}}_{} \Big|^2 dx \le \frac{(2C_1)^2}{N} + \delta$$

  $\Phi(x)$

- $\lim_{r \to \infty} \sigma(r x) = \mathbb{1}_{\mathbb{R}^+}(x)$ (아주 가파른 시그모이드)

<Estimation Error>

- Rademacher Complexity

  - 라데마커 확률변수 $P(\xi_i = 1) = P(\xi_i = -1) = 50\%$ (노이즈)

  - 라데마커 복잡도 $\mathrm{Rad}_n(\mathcal{F}) = E\left[\sup_{f\in\mathcal{F}} \left(\frac{1}{n}\sum_{i=1}^{n} \underbrace{\xi_i f(z_i)}_{\text{노이즈를 따라하는 정도.}}\right)\right] \propto$ 과적합 위험.

---

let $F_{m,\sigma} = \left\{ f_\theta(x) = \sum_{j=1}^{m} \underbrace{\beta_j \sigma(w_j^T x)}_{NN} \right\}$

with (prob) $\geq 1-\delta$

then $\sup_{f\in F_{m,\sigma}} \left| \underbrace{\frac{1}{n}\sum_{i=1}^{n} f(x_i)}_{\text{empirical risk}} - \underbrace{E[f]}_{\text{population risk.}} \right| \leq 2\,\mathrm{Rad}(F_{m,\sigma}) + \sqrt{\frac{1}{2n}\cdot \log\frac{2}{\delta}}$

also, $\mathrm{Rad}(F_{m,\sigma,Q}) \leq \frac{Q}{\sqrt{n}}$ $\quad F_{m,\sigma,Q} = \{ f\in F_{m,\sigma} : \|f\| \leq Q \}$

(유계 II)

$B > 0$ 이고, 데이터가 $\|x_i\|_2 \le C$ 이며

σ가 ReLU 이면 $\Rightarrow$ $Rad(F_{m,\sigma,B}) \le \dfrac{2BC}{\sqrt{n}}$

· 2계층 신경망 $f_\theta^{(m)} = \sum_{j=1}^{m} \beta_j \sigma(w_i^T x)$

· 복잡도 척도 $C(\theta) = \sum_{j=1}^{m} |\beta_j| \|w_j\|_2$

· $F_{m,\sigma,B} = \{ f_\theta \in F_{m,\sigma} : C(\theta) \le B \}$

Positive Homogeneity

· ReLU의 핵심 성질 $\alpha \sigma(x) = \sigma(\alpha x) \quad \forall \alpha > 0$

· 파라미터 재조정 $\theta = \{(\beta_j, w_j)\}_{j \le m}$

$\theta' = \{(\lambda_j \beta_j, w_j / \lambda_j)\}_{j \le m} \quad (\lambda_i > 0)$

Note. 탈라그랑의 축약 원리

$Rad(\phi \circ F) \le K \cdot Rad(F)$

when ϕ: 립시츠 함수

K: 립시츠 상수

ReLU는 립시츠 함수이다.

· 가중치 방향의 분리 let $\lambda_j = \|w_j\|_2$

방향 단위벡터

$\sigma(w_j^T x) = \|w_j\|_2 \cdot \sigma(\bar{w}_j^T x)$

· 라데마허 복잡도 축약

$F_1 = \{ x \mapsto w^T x : w \in \mathbb{R}^d, \|w\|_2 \le 1 \}$

$f_\theta(x) = \sum \beta_j \sigma(w_j^T x) = \sum \beta_j \|w_j\|_2 \sigma(\bar{w}_j^T x)$

· $Rad(F_{m,\sigma,B}) \le 2B \cdot Rad(F_1) = 2B \cdot \left(\frac{C}{\sqrt{n}}\right)$