Section 1. Basic Decomposition of Risk

[문제상황]

data $\{(x_i, y_i)\}_{i=1,\dots,n}$ $\Rightarrow$ map $f: \mathcal{X} \to \mathcal{Y}$

loss $f: \mathcal{Y} \times \mathcal{Y} \to \mathcal{Y}$

(pointwise)

$\begin{cases} \text{i.i.d} \\ x_i \in \mathcal{X} \text{ (high dim)} \\ y_i \in \mathcal{Y} \text{ (label)} \end{cases}$

[Generalization]

(population risk) $R(f) = \mathbb{E}_{(x,y)}[\ell(f(x), y)] = \mathbb{E}[\hat{R}(f)]$

(empirical risk) $\hat{R}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i), y_i)$

[Approximation]

- Hypothesis Space $\mathcal{F} = \{f: \mathcal{X} \to \mathcal{Y}\}$ (suppose $\mathcal{F}$ normed space)
- complexity measure $\gamma(f)$, $\|f\|$
  1. Euclidean Norm of weights
  2. # of params
  3. # of G.D iterations

$\Rightarrow$ let $\mathcal{F}_\delta = \{f \in \mathcal{F} : \gamma(f) \le \delta\}$

ERM (empirical risk minimization)

empirical risk ⇓

with limited complexity

① $\min_{f \in \mathcal{F}_\delta} \hat{R}(f)$

risk

② $\min_{f \in \mathcal{F}} \hat{R}(f) + \lambda\gamma(f)$

risk complexity

③ $\min \gamma(f)$ complexity

when $\hat{R}(f) = 0$ $(f(x_i) = y_i)$

Basic Decomposition of risk

$$\mathrm{I}\quad R(\hat f) - \hat R(\hat f) + \hat R(\hat f) - \inf_{f\in F_\delta} \hat R(f)$$

optimization error

$$\mathrm{II}\quad + \inf_{f\in F_\delta} \hat R(f) - \inf_{f\in F_\delta} R(f) + \inf_{f\in F_\delta} R(f) - \inf_{f\in F} R(f)$$

generalization error

approximation error

$$\le 2\cdot \sup_{f\in F_\delta} \left| R(f) - \hat R(f) \right| + \text{opt} + \text{apx}$$

Note. II

$$\left| \inf_{F_\delta} \hat R(f) - \inf_{F_\delta} R(f) \right| \le \sup_{F_\delta} \left| R(f) - \hat R(f) \right|$$

(pf)

$$\begin{aligned} &\inf_{F_\delta} \hat R(f) - \inf_{F_\delta} R(f) \\ &= \inf_{F_\delta} \hat R(f) - R(f^*) \\ &\le \hat R(f^*) - R(f^*) \\ &\le \text{RHS} \end{aligned}$$

$$\begin{aligned} &\inf_{F_\delta} R(f) - \inf_{F_\delta} \hat R(f) \\ &= \inf_{F_\delta} R(f) - \hat R(\hat f) \\ &\le R(\hat f) - \hat R(\hat f) \\ &\le \text{RHS} \end{aligned}$$

Section 2. Generalization Error and the Curse of dimensionality

---

trade-off

$\delta \Uparrow$ $n \Downarrow$ then generalization error $\Uparrow$ vs $\delta \Downarrow$ then approximation error $\Uparrow$

---

idea ① $n \Uparrow$ then $\mathcal{E}_{gen} \Downarrow$

$$\underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)}_{\text{empirical risk}} - \underbrace{\mathbb{E}_{(x,y)} \left[\ell(\overset{\text{fixed}}{\cancel{f}}(x), y) \right]}_{\text{population risk}}$$

$$\underbrace{\mathrm{Var}\left[\hat{R}(f) - R(f) \right]}_{\text{generalization error}} = \underbrace{\frac{1}{n}}_{\text{데이터 } n\text{개}} \cdot \mathrm{Var}\left[\ell(\overset{\text{fixed.}}{\cancel{f}}(x_i), y_i) \right]$$

---

idea ② $d \Uparrow$ then $n \Uparrow$ (to keep $\mathcal{E}_{gen}$ stable)

let $\{ (\underbrace{x_i}_{\substack{N(0, I_d) \\ \in \mathbb{R}^d}}, \underbrace{f^*(x_i)}_{\text{target function}}) \}_{i=1,\dots,n}$

· minimize $\mathbb{E} \left| \hat{f}(x) - f^*(x) \right| \leq \varepsilon$

- · suppose $f^*(x) = \langle x, \theta^* \rangle$
- · then $\mathcal{F} = \{ f : f(x) = \langle x, \theta \rangle \} \approx \mathbb{R}^d$
- · $n = d$.

idea ③ curse of dimensity : $d \Uparrow$ then $n \Uparrow\Uparrow\Uparrow$ (to keep $\mathcal{E}_{gen}$ stable)

- f^* only locally linear & β-Lipschitz.

- $\|f\| = \mathrm{Lip}(f) + \|f\|_\infty$

  where $\mathrm{Lip}(f) = \inf \{ \beta' : |f(x) - f(x')| \leq \beta' \|x - x'\| \}$

- $\hat{f} = \underset{f \in \mathcal{F}}{\mathrm{argmin}} \left[\mathrm{Lip}(f) : f(x_i) = f^*(x_i) \right]$

- $$
\begin{aligned}
|\hat{f}(x) - f^*(x)| \leq\ & |\hat{f}(x) - \hat{f}(x_i)| \leq \mathrm{Lip}(\hat{f}) \|x - x_i\| \\
&+ \cancel{|\hat{f}(x_i) - f^*(x_i)|} \\
&+ |f^*(x_i) - f^*(x)| \leq \beta \|x - x_i\|
\end{aligned}
$$

- hence, $(MSE) = \mathbb{E}|\hat{f}(x) - f^*(x)|^2 \leq 4\beta \|x - x_i\|^2$

  $\sim W_2^2 (V, \hat{V}_n)$

  $\sim \frac{1}{n^{1/d}}$

- if $\|\hat{f} - f^*\|^2 \leq \varepsilon$

  then $n \sim (\underbrace{\tfrac{1}{\varepsilon}}_{\text{아주 큰 상수}})^d$

Section 3. Universial Approximation Thm.

UAT : $G(x) = \sum_{j=1}^{N} \alpha_j \, \sigma(w_j^T x + \theta_j)$

are dense in $C(I_n)$

σ : arbitary continuous sigmoid func.

$w_j \in \mathbb{R}^n$, $\alpha_j, \theta_j \in \mathbb{R}$

Note.

$\sigma : \mathbb{R} \to [0, 1]$

is called sigmoid

if ① $\sigma \in C'(\mathbb{R})$

② $\sigma(+\infty) = 1$

③ $\sigma(-\infty) = 0$

$f(x)$: raw system

μ : belief system

$d\mu(x) = \mu(dx)$: weight / importance

$\int f(x) \, d\mu(x)$: final evaluation

σ is discriminatory

for measure $\mu \in M(I_n)$

: $\int_{I_n} \sigma(w^T x + \theta) \, d\mu(x) = 0$

$\forall w \in \mathbb{R}^n$ $\forall \theta \in \mathbb{R}$ then $\mu = 0$

Lemma 1. any conti sigmoid

func is discriminatory for $\forall \mu \in M(I_n)$

Lemma 2. ① U : linear subspace of

normed linear space X.

② U is not dense : $\exists x_0 \in X$ s.t $d(x_0, U) \geq \delta > 0$

then 有 bounded

functional L on X

s.t
$$\begin{cases} ① \ L|_U = 0 \quad \text{i.e } L(u) = 0 \ \forall u \in U \\ ② \ L(x_0) = \delta \\ ③ \ \|L\| \leq 1 \end{cases}$$

$L \neq 0$

Note. bounded linear function

$$\|L\| = \sup_{x \neq 0} \frac{|L(x)|}{\|x\|}$$

Thm. RRT

- compact $K \subseteq \mathbb{R}^n$
- $C(K)$: set of real-valued conti func on K.
- F : bounded linear func on $C(K)$

$\Rightarrow$ 有 unique finite signed measure μ on K

s.t $F(f) = \int_K f(x)\, d\mu(x) \quad \forall f \in C(K)$

linear measure

Lemma 3.

U : linear, non-dense subspace of $C(I_n)$

$\Rightarrow$ 有 $\mu \in M(I_n)$ s.t $\int_U h\, d\mu = 0 \quad \forall h \in U$

$\mu \equiv L$ with $L|_U = 0$

$L \neq 0$

(pf) 有 bounded linear function L s.t $L|_U = 0$, $L \neq 0$ on $C(I_n)$ (lemma 2)

有 $\mu \in M(I_n)$ (RRT)

s.t $L(f) = \int_{I_n} h\, d\mu(x) \quad \forall h \in C(I_n)$

$$\begin{cases} L|_U = 0 & : \ L(f) = 0 \quad (\forall h \in U) \\ L \neq 0 & : \ \mu \neq 0 \end{cases}$$

Lemma 4.

σ be any conti. discriminating func.

then $U = \{ G : G(x) = \sum_{j=1}^{N} \alpha_j \cdot \sigma(w_j^T x + \theta_j) \}$

are dense in $C(I_n)$ $(w_j \in \mathbb{R}^d,\ \alpha_j, \theta_j \in \mathbb{R})$

(pf) suppose U not dense

· 有 measure $\mu \neq 0$ s.t $\int_{I_n} h \, d\mu = 0 \quad \forall h \in U$

$$\downarrow$$

$$\sum_{j=0}^{N} \alpha_j \int_{I_n} \sigma(w_j^T x + \theta_j) \, d\mu = 0 \qquad \forall w_j \in \mathbb{R}^n,\ \alpha_j, \theta_j \in \mathbb{R}$$

$$\downarrow$$

$$\int_{I_n} \sigma(w^T x + \theta) \, d\mu = 0 \qquad \forall w \in \mathbb{R}^n,\ \theta \in \mathbb{R}$$

since σ discriminatory, $\mu = 0 \longrightarrow$ contradict.

U is dense

$\Rightarrow \forall f \in C(I_n)$ and $\forall \varepsilon > 0$

有 $G(x)$ s.t $\underbrace{|G(x) - f(x)| < \varepsilon}_{L_\infty \text{ norm}}$ $(\forall x \in I_n)$