Generalization Theory

2020年10月12日 14:15

there is no universal learner.

THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*
2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

Here the hypothesis class is the whole function space!

All possible functions f — distribution D

$\Rightarrow L_D(f) = 0$

All kinds of sampled training set S

bound the expected loss $\max_i \mathbb{E}_{S\sim D_i} [L_{D_i}(A(S))]$ by $\frac{1}{4}$

by avg loss of f class

Proof. Assume $m = |\mathcal{X}|/2$

$T = 2^{2m}$ functions $\mathcal{X} \to \{0,1\}$

$\{f_1, \cdots, f_T\}$.

def. $D_i(\{x,y\}) = \begin{cases} \frac{1}{|\mathcal{X}|} & \text{if } y = f_i(x) \\ 0 & \text{otherwise} \end{cases}$ (distribution)

$\Rightarrow L_{D_i}(f_i) = 0.$

There are $k = (2m)^m$ kinds of sequence of sampling result $\Rightarrow \{S_1, \cdots, S_k\}$

$S_j = (x_1, \cdots, x_m)$

$S_j^i = ((x_1, f_i(x_1)), \cdots, (x_m, f_i(x_m)))$

$\Rightarrow \mathbb{E}_{S\sim D_i^m} [L_{D_i}(A(S))] = \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$

$A(S)$: function derived by algo. A.

$\max_{i\in[T]} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{k} \sum_{j=1}^{k} L_{D_i}(A(S_j^i))$

max ≥ Avg ≥ min $\quad \geq \min_{j\in[k]} \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i))$

Let $v_1, \cdots, v_p$ be examples in $C - S_j$

$\Rightarrow p \geq m$

$\Rightarrow \forall$ function h

$L_{D_i}(h) = \frac{1}{2m} \sum_{x\in C} [h(x) \neq f_i(x)]$

$\geq \frac{1}{2m} \sum_{r=1}^{p} [h(v_r) \neq f_i(v_r)]$

$\geq \frac{1}{2p} \sum_{r=1}^{p} [h(v_r) \neq f_i(v_r)]$

$\Rightarrow \frac{1}{T} \sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} [A(S_j^i)(v_r) \neq f_i(v_r)]$

$= \frac{1}{2p} \sum_{r=1}^{p} \frac{1}{T} \sum_{i=1}^{T} [A(S_j^i)(v_r) \neq f_i(v_r)]$

$$\Rightarrow \frac{1}{T}\sum_{i=1}^{T} L_{D_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T} \frac{1}{2p}\sum_{r=1}^{p} [A(S_j^i)(v_r) \neq f_i(v_r)]$$

$$= \frac{1}{2p}\sum_{r=1}^{p} \frac{1}{T}\sum_{i=1}^{T} [A(S_j^i)(V_r) \neq f_i(V_r)]$$

$$\geq \frac{1}{2} \cdot \min_{r\in[p]} \frac{1}{T}\sum_{i=1}^{T} [A(S_j^i)(V_r) \neq f_i(V_r)]$$

partition T functions into $T/2$ pairs. $(f_i, f_{i'})$ only different on $f_i(V_r)$, $f_{i'}(V_r)$.

$$\Rightarrow \frac{1}{T}\sum_{i=1}^{T} [A(S_j^i)(V_r) \neq f_i(V_r)] = \frac{1}{2}$$

$$\Rightarrow \boxed{\begin{array}{l} \forall \text{ algo } A',\ \exists f: X \to \{0,1\} \text{ and distribution } D. \\ \text{s.t. } \mathbb{E}_{S\sim D^m}[L_D(A'(S))] \geq \frac{1}{4} \end{array}}$$

Hypothesis Class

- the class of functions
- infinite / finite

e.g. $H = \{f: f(x) = w^T x,\ \|w\|_2 \leq 1\}$ (infinite)

e.g. decision tree (finite)

Empirical risk memorization.

↳ $ERM_H(S) \in \underset{h\in H}{\operatorname{argmin}}\ L_S(h)$

training dataset (S)

If H is infinite, ERM_H may simply memorize S (overfit)

Def. 2.1 Realizability Assumption

There exists $h^* \in \mathcal{H}$, s.t. $L_{(D,f)}(h^*) = 0$ $\qquad$ h: hypo. function.

$\forall S$ sampled from D, labled by f. $L_S(h^*) = 0$.

Corollary 2.3 $\mathcal{H}$ finite hypothesis class, $\delta \in (0,1)$. $\varepsilon > 0$.

let integer m: $m \geq \dfrac{\log(|\mathcal{H}|/\delta)}{\varepsilon}$.

Then $\forall D, f$. that realizability assumption holds.

w.h.p of $1-\delta$. over i.i.d. sample S of size m, we have

$\forall$ ERM hypo. h_S: $L_{(D,f)}(h_S) < \varepsilon$

Proof. want to upper bound:

$$\bar{S} = \{S: L_{(D,f)}(h_S) > \varepsilon\}$$

bad hypothesis: $\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{(D,f)}(h) > \varepsilon\}$

misleading samples: $M = \{S:\ \exists h \in \mathcal{H}_B.\ L_S(h) = 0\}$

bad hypothesis: $\mathcal{H}_B = \{h \in \mathcal{H} \mid L_{(D,f)}(h) > \varepsilon\}$

misleading samples: $M = \{S : \exists h \in \mathcal{H}_B.\ L_S(h) = 0\}$

By realization assumption, $L_S(h) = 0$

$\Rightarrow$ $L_{(D,f)}(h_S) > \varepsilon$ only if S is misleading

$$\Rightarrow \quad \bar{S} \subseteq M = \bigcup_{h \in \mathcal{H}_B} \{S \mid L_S(h) = 0\}$$

$$D^m(\bar{S}) \leq \sum_{h \in \mathcal{H}_B} D^m(\{S \mid L_S(h) = 0\}) \quad \text{Union Bound}$$

$\downarrow$ $S = \{x_1, \cdots, x_m\}$

$$= \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} D(\{x_i : h(x_i) = f(x_i)\})$$

$$= \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} 1 - L_{(D,f)}(h)$$

$$\leq \sum_{h \in \mathcal{H}_B} \prod_{i=1}^{m} (1-\varepsilon)$$

$$\leq \sum_{h \in \mathcal{H}_B} e^{-\varepsilon m} \leq |\mathcal{H}| e^{-\varepsilon m}$$

Def 3.1 PAC Learnability

A hypo. class $\mathcal{H}$ is PAC learnable if:

$\exists$ function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$, Learning algo. A.

$\forall$ D, f if realization assumption holds, run A on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ examples (iid)

$\Rightarrow$ A returns h w.p. $\geq 1-\delta$. $L_{(D,f)}(h) \leq \varepsilon$

e.g. ERM for finite $\mathcal{H}$ is PAC-learnable

Realizability too strong

- Sometimes $L_{D,f}(h^*)$ cannot reach 0
- f may have multiple labels.

weaker version

the Bayes optimal predictor

$$f_D(x) = \begin{cases} 1 & \text{if } \Pr[y=1 \mid x] \geq \frac{1}{2}. \\ 0 & \text{otherwise.} \end{cases}$$

$\Rightarrow$ expect $\underline{L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \varepsilon}$ (instead of $\leq \varepsilon$)

when realizability holds, degrade to

use this to replace the requirement in PAC-learnability

$\boxed{\text{Agnostic PAC Learnability}}$

Error Decomposition

$L_D(h_S) = \varepsilon_{app} + \varepsilon_{est}$

$L_D(h_S) = \mathcal{E}_{app} + \mathcal{E}_{est}$

$\mathcal{E}_{app}$: approximation $= \min_{h \in \mathcal{H}} L_D(h) = \underbrace{L_D(BO)}_{\text{base opt.}} + \mathcal{E}'_{app}$

due to inherence of D.

$\mathcal{E}_{est}$: estimation

Infinite $\mathcal{H}$?

Lemma 6.2. $\mathcal{H}$: class of thresholds (1D)

ERM rule & $m_{\mathcal{H}}(\varepsilon, \delta) \le \lceil \log(2/\delta)/\varepsilon \rceil$ PAC learnable.

Generally $\Rightarrow$ VC dimension

Def. Restriction of $\mathcal{H}$ to C.

Let $C = \{c_1, \cdots, c_m\} \subset \mathcal{X}$

$\Rightarrow \mathcal{H}_C = \{ (h(c_1), \cdots, h(c_m)) : h \in \mathcal{H} \}$

Def. If all function $h: C \to \{0,1\}$ is in $\mathcal{H}_C$, then $\mathcal{H}$ shatters C.

Def. VC dimension

$VCdim(\mathcal{H})$: maximal size of $C \subset X$ that can be shattered by X.

THEOREM 6.8 (The Fundamental Theorem of Statistical Learning – Quantitative Version) *Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0,1\}$ and let the loss function be the $0-1$ loss. Assume that $\mathrm{VCdim}(\mathcal{H}) = d < \infty$. Then, there are absolute constants C_1, C_2 such that:*

1. *$\mathcal{H}$ has the uniform convergence property with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m^{UC}_{\mathcal{H}}(\epsilon, \delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
2. *$\mathcal{H}$ is agnostic PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon^2} \le m_{\mathcal{H}}(\epsilon, \delta) \le C_2 \frac{d + \log(1/\delta)}{\epsilon^2}$$
3. *$\mathcal{H}$ is PAC learnable with sample complexity*
$$C_1 \frac{d + \log(1/\delta)}{\epsilon} \le m_{\mathcal{H}}(\epsilon, \delta) \le C_2 \frac{d\log(1/\epsilon) + \log(1/\delta)}{\epsilon}$$

Def. ε - Representative Sample

A training set S.

$$\sup_{h \in \mathcal{H}} |L_D(h) - L_S(h)| \le \varepsilon.$$

$\mathcal{F} = \ell \circ \mathcal{H}$ (loss, hypo)

Def. Representativeness of S w.r.t. $\mathcal{F}$:

$\mathrm{Rep}_D(\mathcal{F}, S) \overset{def}{=} \sup (L_D(f) - L_S(f))$

Def. Representativeness of S w.r.t. $\mathcal{F}$:

$$\mathrm{Rep}_D(\mathcal{F}, S) \stackrel{\text{def}}{=} \sup_{f \in \mathcal{F}} \left(L_D(f) - L_S(f) \right)$$

$$L_D(f) = \mathop{\mathbb{E}}_{z \sim D} f(z)$$

Split S into two parts to estimate this:

Rademacher variable $\sigma_i = \begin{cases} \frac{1}{2} & p = 0.5 \\ -\frac{1}{2} & p = 0.5 \end{cases}$

Rademacher complexity of $\mathcal{F}$ w.r.t. S:

$$R(\mathcal{F} \circ S) \stackrel{\text{def}}{=} \frac{1}{m} \mathop{\mathbb{E}}_{\sigma \sim \{\pm 1\}^m} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i) \right]$$

Lemma.

$$\mathop{\mathbb{E}}_{S \sim D^m} \left[\mathrm{Rep}_D(\mathcal{F}, S) \right] \leq 2 \mathop{\mathbb{E}}_{S \sim D^m} R(\mathcal{F} \circ S)$$

Proof. $L_D(f) - L_S(f) = \mathop{\mathbb{E}}_{S'} \left(L_{S'}(f) - L_S(f) \right)$

$$\mathop{\mathbb{E}}_{S} \left[\sup_f \left(L_D(f) - L_S(f) \right) \right] \leq \mathop{\mathbb{E}}_{S} \left[\mathop{\mathbb{E}}_{S'} \left[\sup_f \left(L_{S'}(f) - L_S(f) \right) \right] \right]$$

$$= \frac{1}{m} \mathop{\mathbb{E}}_{S,S'} \left[\sup_{f \in \mathcal{F}} \left(f(z_i') - f(z_i) \right) \right]$$

z_j', z_j are iid & indep.

$$\mathop{\mathbb{E}}_{S,S'} \sup_f \left[f(z_j') - f(z_j) + \sum_{i \neq j} f(z_i') - f(z_i) \right] \quad ①$$

$$= \mathop{\mathbb{E}}_{S,S'} \sup_f \left[f(z_j) - f(z_j') + \sum_{i \neq j} f(z_i') - f(z_i) \right] \quad ②$$

$$\Rightarrow \mathop{\mathbb{E}}_{S,S',\sigma_j} \sup_f \left(\left[f(z_j') - f(z_j) \right] \sigma_j + \sum_{i \neq j} f(z_i') - f(z_i) \right)$$

$$= \frac{1}{2} ① + \frac{1}{2} ②$$

$$= ①$$

$$\mathop{\mathbb{E}}_{S,S'} \sup_f \left(\sum_{i=1}^{m} f(z_i') - f(z_i) \right) = \mathop{\mathbb{E}}_{S,S',\sigma} \sup_f \left[\sum_{i=1}^{m} \sigma_i \left(f(z_i') - f(z_i) \right) \right]$$

$$\sup_f \left[\sum_{i=1}^{m} \sigma_i \left(f(z_i') - f(z_i) \right) \right]$$

$$\leq \sup_f \sum_{i=1}^{m} \sigma_i f(z_i') + \sup_f \sum_{i=1}^{m} \sigma_i f(z_i)$$

$$\Rightarrow \mathop{\mathbb{E}}_{S} \left[\mathrm{Rep}_D(\mathcal{F}, S) \right] \leq \frac{2}{m} \mathop{\mathbb{E}}_{S,S',\sigma} \left(\sup \sum_{i=1}^{m} \sigma_i f(z_i) \right)$$

$$= 2 \mathop{\mathbb{E}}_{S} \left[R(\mathcal{F} \circ S) \right]$$

THEOREM 26.5 *Assume that for all z and $h \in \mathcal{H}$ we have that $|\ell(h,z)| \leq c$. Then,*

1. *With probability of at least $1-\delta$, for all $h \in \mathcal{H}$,*

$$L_{\mathcal{D}}(h) - L_S(h) \;\leq\; 2 \underset{S' \sim D^m}{\mathbb{E}} R(\ell \circ \mathcal{H} \circ S') + c\sqrt{\frac{2\,\ln(2/\delta)}{m}}.$$

   In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$.

2. *With probability of at least $1-\delta$, for all $h \in \mathcal{H}$,*

$$L_{\mathcal{D}}(h) - L_S(h) \;\leq\; 2\,R(\ell \circ \mathcal{H} \circ S) + 4\,c\sqrt{\frac{2\,\ln(4/\delta)}{m}}.$$

   In particular, this holds for $h = \mathrm{ERM}_{\mathcal{H}}(S)$.

3. *For any $h^\star$, with probability of at least $1-\delta$,*

$$L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h^\star) \;\leq\; 2\,R(\ell \circ \mathcal{H} \circ S) + 5\,c\sqrt{\frac{2\,\ln(8/\delta)}{m}}.$$