Statistics Learning Theorem:General PAC model with finite class

2019.02.09

Agnostic PAC Learnability for General Loss Functions

A hypothesis class $\mathcal{H}$ is agnostic PAC learnable with respect to a set Z and a loss function $l : \mathcal{H} \times Z \rightarrow R_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow N$ and a learning algorithm with the following preperty: For every $\varepsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid samples generated by $\mathcal{D}$, the algorithm returns $h \in \mathcal{H}$ such that, with probability of at least $1 - \delta$

$$L_{\mathcal{D}} \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \varepsilon$$

where $L_{\mathcal{D}}(h) = E_{z \sim \mathcal{D}}[l(h, z)]$

Goal

In today's slide, we aim to show that finite hypothesis class is agnostic PAC learnable with general learnable under uniform covergence condition.

Some terminology

Definition (ε-representative sample)

A training set S is called ε-representative sample if

$$\forall h \in \mathcal{H}, \quad |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$$

Definition (Uniform Convergence)

We say that a hypothesis class $\mathcal{H}$ has the uniform convergence property, if there exists a function $m_{\mathcal{H}}^{UC} : (0, 1)^2 \rightarrow N$ such that for every $\varepsilon, \delta \in (0, 1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ examples drawn i.i.d, according to $\mathcal{D}$, then, with probability of at least $1 - \delta$, S is ε-representative sample

Main Theorem

Theorem

Every finite hypothesis class is agnostic PAC learnable

Proof

Before the starting the proof, we set up some lemma in the fist.

Lemma

Assume that a training set S is $\frac{\varepsilon}{2}$-representative sample. Then, any output of $ERM_{\mathcal{H}}(S)$,namely, any $h_S \in arg\min_{h\in\mathcal{H}} L_S(h)$, satisfies

$$L_{\mathcal{D}}(h_S) \leq \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \varepsilon$$

The proof is simple, it follows that

$$L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2} \leq L_S(h) + \frac{\varepsilon}{2} \leq L_{\mathcal{D}}(h) + \varepsilon$$

Proof

From the above lemma, we know that the ERM rule is agnostic PAC learnable if given ε, δ we can alway find ε-representative sample under $1 - \delta$ probaility.
That is, the definition of uniform convergence. Therefore, we can get the following corollary:

Theorem

If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostically PAC learnable with the sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta)$. Furthermore, in that case, the $ERM_{\mathcal{H}}$ paradigm is a sucessful agnostic PAC learner for $\mathcal{H}$

Proof:Framework

By above theorem, we only need to show that the finite class $\mathcal{H}$ satisfies the property of uniform convergence.
We follow a two step argument. The first step applies the union bound and the second step employs a measure concentration inequality.

Proof:First Step

Fix some ε, δ, we want to show that

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\}) < \delta$$

We know that

$$\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\} = \cup_{h \in \mathcal{H}} \{S : |L_S(h) - L_D(h)| > \varepsilon\}$$

Therefore,

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\})$$

$$\leq \sum_{h \in \mathcal{H}} D^m(\{S : |L_S(h) - L_D(h)| > \varepsilon\})$$

Proof:Second Step

Since $L_D(h) = E[l(h, z)]$, $L_S = \frac{1}{m}\sum_{i=1}^{m} l(h, z_i)$, the iid assumption and law of large number make sure the each summand would converge to zero. However, we want to estimate the sample complexity, it requires to estimate the upper bound of concentrated error under finite error. To achieve it, we use the Hoeffding's Inequality:

Lemma (Hoeffding's Inequality)

Let $\theta_1, \cdots, \theta_m$ be a sequence of i.i.d. random variables and assume that for all $i, E[\theta_i] = \mu$ and $P[a \leq \theta_i \leq b] = 1$. Then, for any $\varepsilon > 0$

$$P[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \varepsilon] \leq 2exp(-2m\varepsilon^2/(b-a)^2)$$

Proof:Second Step

Using the same notation in the lemma($L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\theta_i, L_D(h) = \mu$), we have that

$$D^m(\{S : |L_S(h) - L_D(h)| > \varepsilon\})$$

$$= P[|\frac{1}{m}\sum_{i=1}^{m}\theta_i - \mu| > \varepsilon] \leq 2exp(-2m\varepsilon^2)$$

Therefore, we have

$$D^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \varepsilon\})$$

$$\leq \sum_{h\in\mathcal{H}} 2exp(-2m\varepsilon^2)$$

$$= 2|\mathcal{H}|exp(-2m\varepsilon^2)$$

That is , we choose

$$m \geq \frac{\log(2|\mathcal{H}|/\delta)}{2\varepsilon^2}$$

No Free Lunch Theorem

Question: Does there exists the universal learner?
The answer to this question is negative!!

No Free Lunch Theorem

Theorem

Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain X. Let m be any number smaller than $|X|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $X \times \{0, 1\}$ such that:

1. *There exists a function $f : X \to \{0, 1\}$ 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 $L_{\mathcal{D}}(A(S)) \geq 1/8$*

No Free Lunch Theorem

This theorem states that for every learner, there exists a task on which it fails, even though that task can be successfully learned by another learner.

No Free Lunch Theorem:Proof

Let C be a subset of X of size $2m$.
The main idea of proof is that any learning algorithm that observes only half of the instances in C has no information on what should be the labels of the rest of the instances in C has no information on what should be the labels of the rest of instances in C.
Note that there are $T = 2^{2m}$ possible functions from C to $\{0, 1\}$.
Denote these functions by $f_1, \cdots, f_T$.

No Free Lunch Theorem:Proof

Let $\mathcal{D}_i$ be a distribution over $C \times \{0, 1\}$ defined by

$$\mathcal{D}_i(\{x, y\}) = \begin{cases} 1/|C| & \textit{if} \quad y = f_i(x) \\ 0 & \textit{otherwise} \end{cases}$$

That is, the distribution make f_i as a correct label function.
We will show that for every algorithm, A, that receives a training set of m examples from $C \times \{0, 1\}$ and returns a function $A(S) : C \to \{0, 1\}$, it holds that

$$\max_{i \in [T]} E_{S \sim \mathcal{D}_i^m}[L_{\mathcal{D}_i} A(S))] \geq 1/4$$

No Free Lunch Theorem:Proof

To prove the above claim, note that there are $k = (2m)^m$ possible sequence of m examples from C. Denote these sequence by $S_1, \cdots, S_k$.
If $S_j = (x_1, \cdots, x_m)$ then we denote

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

Then we can write the $E_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}A(S))]$ as

$$E_{S\sim\mathcal{D}_i^m}[L_{\mathcal{D}_i}A(S))] = \frac{1}{k}\sum_{j=1}^{k} L_{\mathcal{D}_i}(A(S_j^i))$$

No Free Lunch Theorem:Proof

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

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i))$$

$$\geq \min_{j\in[k]} \frac{1}{T}\sum_{i=1}^{T} L_{\mathcal{D}_i}(A(S_j^i))$$

No Free Lunch Theorem:Proof

Next fix $j \in [k]$ and let $v_1 \cdots v_p$ be the examples in C that do not appear in S_j. Therefore, for every function $h : C \rightarrow \{0, 1\}$ we have

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

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

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

Hence

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

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

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

No Free Lunch Theorem:Proof

Next, fix some $r \in [p]$, we can partition all the functions in $f_1, \cdots, f_T$ into $T/2$ disjoint paris, where for a pair $(f_i, f_{i'})$ we have that for every $c \in C$, $f_i(c) \neq f_{i'}(c)$ if and only if $c = v_r$. Therefore

$$I_{[A(S_j^i)(v_r) \neq f_i(v_r)]} + I_{[A(S_j^i)(v_r) \neq f_{i'}(v_r)]} = 1$$

which yield

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

No Free Lunch Theorem:Proof

Clearly, this means that there exist a function $f : X \to \{0, 1\}$ and a distribution $\mathcal{D}$ such that

$$E_{S \sim \mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq 1/4$$

By Markov ineqality

$$Pr[L_{\mathcal{D}}(A(S)) \geq 1/8] \geq \frac{1/4 - 1/8}{7/8} = \frac{1}{7}$$