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 $l : \mathcal{H} \times Z \to R_+$, if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to N$ 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 $m_{\mathcal{H}}^{UC} : (0,1)^2 \to N$ 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

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