Review of Chapter07

Understanding Machine Learning

Peng Lingwei

July 17, 2019

.

Uniform Convergence

Definition

A hypothesis class $\mathcal{H}$ is *Uniform Convergence* if:
$\exists m_{\mathcal{H}}^{UC}(\delta, \epsilon) \to \mathbb{N}$ satisfies:
For $\forall \epsilon, \delta \in (0,1)$,
The training set $\{S : |S| \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta), S \sim \mathcal{D}^m\}$ guarantees that
$\mathbb{P}\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\} \geq 1 - \delta$.

Theorem

$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \left\lceil \frac{\log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \right\rceil$$

Proof.

1. $L_{S=\{(x_1,y_1),\ldots,(x_m,y_m)\}\sim D^m}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x_i) \neq y_i\}$
2. $L_D = \mathbb{E}_{(x,y)\sim D}\{1\{h(x) \neq y\}\} = \mathbb{E}_{S\sim D^m}\{L_S(h)\}$
3. $\mathbb{P}\{\exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\} \leq \sum_{i=1}^{|\mathcal{H}|} \mathbb{P}\{h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon\} \leq 2|\mathcal{H}| \exp(-2m\epsilon^2)$

Nonuniformly Learnable

Definition

A hypothesis class $\mathcal{H}$ is *Nonuniform Learnable* if:
$\exists A : S \to h_S \in \mathcal{H}, m_{\mathcal{H}}^{NUL}(\delta, \epsilon, h) \to \mathbb{N}$ satisfies:
For $\forall \epsilon, \delta \in (0,1), h \in \mathcal{H}$,
The training set $\{S : |S| \geq m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h), S \sim \mathcal{D}^m\}$ gaurantees that

$$\mathbb{P}\{L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}(h) \leq \epsilon\} \geq 1 - \delta$$

Theorem

$\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$ (means countable sets' union), s.t. $\mathcal{H}_n$ is uniform convergence.
$\Rightarrow \mathcal{H}$ is nonuniformly learnable.

Theorem

$\mathcal{H}$ of binary classfiers is nonuniformly learnable.
$\iff \mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$, s.t. $\mathcal{H}_n$ is agnostic PAC learnable.

Structural Risk Minimization

Definition

$$\epsilon_n(m, \delta) = \min\{\epsilon \in (0, 1) : m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) \le m\}$$

Theorem

$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) - L_S(h) \le \min_{n:h\in\mathcal{H}_n} \epsilon_n(m, w(n)\delta)\} \ge 1 - \delta$

Proof.

$\mathbb{P}\{\forall h \in \mathcal{H}_n, |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon_n(m, w(n)\delta)\} \ge 1 - w(n)\delta$
$\mathbb{P}\{\exists h \in \mathcal{H}_n, |L_S(h) - L_{\mathcal{D}}(h)| \ge \epsilon_n(m, w(n)\delta)\} \le w(n)\delta$
$\mathbb{P}\{\exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \ge \epsilon_{n:h\in\mathcal{H}_n}(m, w(n)\delta)\} \le \sum_n w(n)\delta \le \delta$
$\mathbb{P}\{\forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \le \epsilon_{n:h\in\mathcal{H}_n}(m, w(n)\delta)\} \le 1 - \delta$
$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) - L_S(h) \le \min_{n:h\in\mathcal{H}_n} \epsilon_n(m, w(n)\delta)\} \le 1 - \delta$ $\square$

Definition

(Structural Risk Minimization)

1. **prior knowledge**:
   - $\mathcal{H} = \cup_n \mathcal{H}_n$ where $\mathcal{H}_n$ has uniform convergence with $m^{UC}_{\mathcal{H}_n}$.
   - $w : \mathbb{N} \to [0,1]$ s.t. $\sum_n w(n) \leq 1$
2. **define**: ϵ_n and $n(h) = \min\{n : h \in \mathcal{H}_n\}$
3. **input**: training set $S \sim \mathcal{D}^m$, confidence δ
4. **output**: $h \in \arg\min_{h \in \mathcal{H}} [L_S(h) + \epsilon_{n(h)}(m, w(n(h))\delta)]$

Theorem

$$m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq m_{\mathcal{H}_{n(h)}}^{UC}\left(\epsilon/2, \frac{6\delta/2}{(\pi n(h))^2}\right)$$

Proof.

Let $m \geq m_{\mathcal{H}_{n(h)}}^{UC}\left(\epsilon/2, \frac{6\delta/2}{(\pi n(h))^2}\right)$, then
$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) - L_S(h) \leq \epsilon_n(m, w(n(h))\delta)\} \geq 1 - \delta/2$
$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(A(S)) \leq L_S(h) + \epsilon/2\} \geq 1 - \delta/2$.
From uniform convergence property, we also can get:
$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{S}}(h) \leq L_D(h) + \epsilon/2\} \geq 1 - \delta/2$ Then we can guarantee nonuniformly learnable event happens:
$\mathbb{P}\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(A(S)) \leq L_D(h) + \epsilon\} \geq 1 - \delta$ $\square$

In chapter6:
If $VCdim(\mathcal{H}) = n$,then $m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) = C\frac{n+\log(1/\delta)}{\epsilon^2}$.

Theorem

$$
\begin{aligned}
m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) - m^{UC}_{\mathcal{H}_n}(\epsilon/2, \delta) \leq & m^{UC}_{\mathcal{H}_n}\left(\epsilon/2, \frac{3\delta}{(\pi n)^2}\right) - m^{UC}_{\mathcal{H}_n}(\epsilon/2, \delta) \\
\leq & \frac{4C}{\epsilon^2}\log\left(\frac{(\pi n)^2}{3}\right) \\
\leq & \frac{4C}{\epsilon^2} 2\log\left(\frac{\pi n}{\sqrt{3}}\right) \\
\leq & \frac{4C}{\epsilon^2} 2\log(2n)
\end{aligned}
$$

Consistency

Definition

A hypothesis class $\mathcal{H}$ is *Consistency* in probability distributions set $\mathcal{P}$ if:
$\exists A : S \to h_S \in \mathcal{H}, m_{\mathcal{H}}^{CON}(\delta, \epsilon, h, \mathcal{D}) \to \mathbb{N}$ satisfies:
For $\forall \epsilon, \delta \in (0,1), h \in \mathcal{H}, \mathcal{D} \in \mathcal{P}$,
The training set $\{S : |S| \geq m_{\mathcal{H}}^{CON}(\epsilon, \delta, h), S \sim \mathcal{D}^m\}$ gaurantees that

$$\mathbb{P}\{L_{\mathcal{D}}(A(S)) - L_{\mathcal{D}}(h) \leq \epsilon\} \geq 1 - \delta$$

Theorem

The algorithm **Memorize** *is consistency in countable $\mathcal{X}$. ($\mathcal{P}$ is the set of every distribution on $\mathcal{X}$)*

Proof.

Let an $\mathcal{X}$'s enumeration $\{x_i : i \in \mathbb{N}\}$ satisfies:

$$i \le j \Leftrightarrow \mathcal{D}(x_i) \ge \mathcal{D}(x_j).$$

It's easy to verify $\lim_{n\to\infty} \sum_{i=n}^{\infty} \mathcal{D}(x_i) = 0$, *which means that*

$$\forall \epsilon > 0, \exists N \in \mathbb{N}, \text{ such that } \mathcal{D}(i > N, x_i) < \epsilon$$

$$\begin{aligned}
\mathbb{P}\{\exists x \notin S, \mathcal{D}(x) \ge \epsilon\} \le & \mathbb{P}\{\exists i \le N, x_i \notin S\} \\
& \le \sum_{i=1}^{N} \mathbb{P}\{x_i \notin S\} = \sum_{i=1}^{N} (1 - \mathcal{D}(x_i))^m \\
& \le N(1-\epsilon)^m \le Ne^{-\epsilon m}
\end{aligned}$$