Machine Learning

Uniform Convergence

Fabio Vandin

November 8^{th}, 2022

When is an Hypothesis Class PAC Learnable?

Previously seen result: for binary classification with

- realizability assumption
- 0-1 loss

any finite hypothesis class is PAC learnable by ERM.

What about the more general PAC learning model we have seen last? Recall the (agnostic) PAC learnability for general loss:

Definition

A hypothesis class $\mathcal{H}$ is *agnostic PAC learnable* with respect to a set Z and a loss function $\ell : \mathcal{H} \times Z \rightarrow \mathbb{R}_+$ if there exist a function $m_{\mathcal{H}}$: $(0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm such that for every $\delta, \varepsilon \in (0,1)$, for every distribution $\mathcal{D}$ over Z, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ the algorithm returns a hypothesis h such that, with probability $\geq 1 - \delta$ (over the choice of the m training examples):

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

where $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$

Uniform Convergence and Learnability

Uniform convergence: the empirical risks (training error) of *all* members of $\mathcal{H}$ are good approximations of their true risk (generalization error).

Definition

A training set S is called ε*-representative* (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$) if

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

Proposition

Assume that training set S is $\frac{\varepsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ (i.e., any $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$) satisfies

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

Proof For any $h \in \mathcal{H}$:

$$\begin{aligned}
L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} && S \text{ is } \tfrac{\varepsilon}{2}\text{-representative} \\
&\leq L_S(h) + \frac{\varepsilon}{2} && \text{since } h_S \text{ is picked by ERM } (\Rightarrow L_S(h_S) \leq L_S(h)) \\
&\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} && S \text{ is } \tfrac{\varepsilon}{2}\text{-representative} \\
&\leq L_{\mathcal{D}}(h) + \varepsilon
\end{aligned}$$

This is true for $h = \arg\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h')$ as well

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

Proposition

Assume that training set S is $\frac{\varepsilon}{2}$-representative (w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function ℓ, and distribution $\mathcal{D}$). Then, any output of $\mathrm{ERM}_{\mathcal{H}}(S)$ (i.e., any $h_S \in \arg\min_{h \in \mathcal{H}} L_S(h)$) satisfies

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

Proof.

For every $h \in \mathcal{H}$:

$$\begin{aligned} L_{\mathcal{D}}(h_S) &\leq L_S(h_S) + \frac{\varepsilon}{2} \\ &\leq L_S(h) + \frac{\varepsilon}{2} \\ &\leq L_{\mathcal{D}}(h) + \frac{\varepsilon}{2} + \frac{\varepsilon}{2} \\ &= L_{\mathcal{D}}(h) + \varepsilon \end{aligned}$$

$\square$

Uniform convergence depends on training set: when do we have uniform convergence?

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ i.i.d. examples drawn from $\mathcal{D}$, then with probability $\geq 1-\delta$, S is ε-representative.

Uniform convergence depends on training set: when do we have uniform convergence?

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ i.i.d. examples drawn from $\mathcal{D}$, then with probability $\geq 1 - \delta$, S is ε-representative.

Proposition

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 $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. Furthermore, in that case the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

Uniform convergence depends on training set: when do we have uniform convergence?

Definition

A hypothesis class $\mathcal{H}$ has the *uniform convergence property* (w.r.t. a domain Z and a loss function ℓ) if there exists a function $m_{\mathcal{H}}^{UC} : (0,1)^2 \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and for every probability distribution $\mathcal{D}$ over Z, if S is a sample of $m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$ i.i.d. examples drawn from $\mathcal{D}$, then with probability $\geq 1 - \delta$, S is ε-representative.

Proposition

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 $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta)$. Furthermore, in that case the $\mathrm{ERM}_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

What classes of hypotheses have uniform convergence?

Finite Classes are Agnostic PAC Learnable

We prove that finite sets of hypotheses are agnostic PAC learnable under some restriction for the loss.

Proposition

Let $\mathcal{H}$ be a finite hypothesis class, let Z be a domain, and let $\ell : \mathcal{H} \times Z \to [0, 1]$ be a loss function. Then:

- $\mathcal{H}$ enjoys the uniform convergence property with sample complexity

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

- $\mathcal{H}$ is agnostically PAC learnable using the ERM algorithm with sample complexity

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

Idea of the proof:

1. prove that uniform convergence holds for a finite hypothesis class
2. use previous result on uniform convergence and PAC learnability

Useful tool: Hoeffding's Inequality

Hoeffding's Inequality

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

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

$\mathbb{E}$ [average of the observations] = expectation of each observation

Proof (see also the book) [Corollary 4.6]

Fix $\varepsilon, \delta \in (0,1)$. We need a sample size m such that, for any $\mathcal{D}$, with probability $\geq 1-\delta$ (on the choice of $S = (z_1, z_2, \dots, z_m)$, $z_i = (\vec{x}_i, y_i)\ \forall 1 \leq i \leq m$, sampled i.i.d from $\mathcal{D}$) we have:

for all $h \in \mathcal{H}$: $|L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon$.

That is: $\mathcal{D}^m(\{S : \forall h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| \leq \varepsilon\}) \geq 1-\delta$

Equivalently, we need to show:

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

We have:

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

Then

by union bound

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

Now we want to bound each term in (☆☆).

Recall: $L_{\mathcal{D}}(h) = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)]$

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

Important note: each z_i is sampled i.i.d. from $\mathcal{D}$

$$\Rightarrow \mathbb{E}[\ell(h, z_i)] = \mathbb{E}_{z \sim \mathcal{D}}[\ell(h, z)] = L_{\mathcal{D}}(h)$$

Therefore $\mathbb{E}[L_S(h)] \overset{\text{by def of } L_S(h)}{=} \mathbb{E}\left[\frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)\right]$

by linearity of expectation $\rightarrow$ $= \frac{1}{m} \sum_{i=1}^{m} \underbrace{\mathbb{E}[\ell(h, z_i)]}_{L_{\mathcal{D}}(h)}$

$$= \frac{1}{m} \sum_{i=1}^{m} L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h)$$

Let θ_i be the r.v. given by $\ell(h, z_i)$ → i-th point in S, $(\vec{x}_i, y_i)$
↳ because $z_i \sim \mathcal{D}$

Since h is fixed, z_i are sampled i.i.d. from $\mathcal{D}$
$\Rightarrow \theta_1, \theta_2, \dots, \theta_m$ are i.i.d. r.v.

Note that: $L_S(h) = \frac{1}{m}\sum_{i=1}^{m}\theta_i$; let's define $\mu = L_{\mathcal{D}}(h)$

Given assumption: $\ell : \mathcal{H} \times \mathcal{Z} \to [0,1] \Rightarrow \theta_i \in [0,1], \forall i = 1, \dots, m$

We can apply Hoeffding's inequality with $a_i = 0, b_i = 1 \; \forall i = 1, \dots, m$

$$\mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) = \Pr\left[\left|\frac{1}{m}\left(\sum_{i=1}^{m}\theta_i\right) - \mu\right| > \varepsilon\right]$$

by Hoeffding's inequality → $\leq 2 \cdot e^{-2m\varepsilon^2}$

Combining the above with (☆☆):

$$\mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}) \leq \sum_{h \in \mathcal{H}} 2 e^{-2m\varepsilon^2}$$

since $|\mathcal{H}|$ is finite → $= 2|\mathcal{H}| e^{-2m\varepsilon^2}$

By choosing $m \geq \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \frac{1}{(2\varepsilon^2)}$ then

$$\mathcal{D}^m\left(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\}\right) \leq 2|\mathcal{H}| e^{-\cancel{2\varepsilon^2} \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \frac{1}{\cancel{(2\varepsilon^2)}}}$$

$$= 2|\mathcal{H}| \cdot \frac{\delta}{2|\mathcal{H}|} = \delta$$

$\square$

for example:

$$m = \left\lceil \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \frac{1}{2\varepsilon^2} \right\rceil$$

Bibliography

[UML] Chapter 4