Machine Learning

Uniform Convergence

Fabio Vandin

November 10^{th}, 2023

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 $\varepsilon > 0$

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

average of the observations $\approx$> expectation

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 $\geq 1-\delta$ (over the choice of $S=(z_1, z_2, \dots, z_m)$, $z_i = (\vec{x}_i, y_i)$), we have:

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

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

where $S=(z_1, z_2, \dots, z_m)$, $z_i = (\vec{x}_i, y_i)$, i.i.d. from $\mathcal{D}$.

Equivalently, we need to show:

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

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: $(\star) \overset{\text{union bound}}{\leq} \sum_{h \in \mathcal{H}} \mathcal{D}^m(\{S : |L_S(h) - L_{\mathcal{D}}(h)| > \varepsilon\})$ $(\star\star)$

Now we want to bound each term in $(\star\star)$

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

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

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

by def. of $L_S(h)$

by linearity of expectation

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

$$= \frac{1}{m}\cdot m \cdot L_{\mathcal{D}}(h) = L_{\mathcal{D}}(h)$$

Let θ_i be the r.v. given by $\ell(h, z_i)$ i-th element of S

Since h is fixed and z_i is 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 that $\ell : \mathcal{H} \times Z \to [0,1]$

we have $\theta_i \in [0,1]$, $\forall i = 1, \dots, m$.

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

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

by Hoeffding's ineq. $\rightarrow$ $\leq 2 \cdot e^{-2m \cdot \varepsilon^2}$

Combining the inequality above with $(\star\star)$

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

$$= 2|\mathcal{H}|e^{-2m\varepsilon^2}$$

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

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

$$\leq 2|\mathcal{H}| e^{-\cancel{2\varepsilon^2} \lg\left(\frac{2|\mathcal{H}|}{\delta}\right) \cdot \cancel{\frac{1}{2\varepsilon^2}}}$$

$$= 2|\mathcal{H}| e^{-\lg\left(\frac{2|\mathcal{H}|}{\delta}\right)}$$

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

$\square$

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

Bibliography

[UML] Chapter 4