LEARNING VIA UNIFORM CONVERGENCE

Earlier we showed finite $\mathcal{H}$ class is PAC LEARNABLE.
What about AGNOSTIC PAC LEARNABILITY?
Here we will show via the tool of uniform Convergence.

We define below Condition on training set for which all members of $\mathcal{H}$ have their risks close to true risk more or less.

(ϵ-representative sample) A training set S is called ϵ-representative (w.r.t. $\mathcal{Z}, \mathcal{H}, \ell, D$) if

$$|L_S(h) - L_D(h)| \leq \epsilon \quad \forall h \in \mathcal{H}.$$

$\Downarrow$

LEMMA: If S is $\frac{\epsilon}{2}$-representative sample.

$$\Rightarrow L_D(h_S) \leq \min_{h \in \mathcal{H}} L_D(h) + \epsilon$$

h_S: output of ERM Algorithm

i.e. $h_S = \operatorname{argmin}_{h \in \mathcal{H}} L_S^{ERM}(h)$

Proof: ($\frac{\varepsilon}{2}$-representative sample)

$$L_D(h_S) \leq L_S(h_S) + \frac{\varepsilon}{2}$$

ERM Minimizer $= h_S$

$$\leq L_S(h) + \frac{\varepsilon}{2} \leq L_D(h) + \varepsilon$$

($\frac{\varepsilon}{2}$-representative sample)

$$\therefore L_D(h_S) \leq L_D(h) + \varepsilon \quad (\text{any } h \in \mathcal{H})$$

$$\Rightarrow L_D(h) \leq \min_{h \in \mathcal{H}} L_D(h) + \varepsilon.$$

NOTE: Thus in order to infer Agnostic PAC learnability, it suffices to show that i.i.d. samples are $\frac{\varepsilon}{2}$-representative w.r.t. $\mathcal{H}, Z, \ell, D$ with high probability ($\geq 1-\delta$)

formalize this for hypothesis class

(UNIFORM CONVERGENCE PROPERTY [UCP])

A hypothesis class $\mathcal{H}$ has uniform convergence property (w.r.t. Z, ℓ)

if $\forall \varepsilon, \delta \in (0,1)$ and $\forall D$

$\exists\ m_{\mathcal{H}}^{UC}: (0,1)^2 \to \mathbb{N}$ s.t. for any w.p. $\geq 1-\delta$, a sample sequence S of m

($m \geq m_{\mathcal{H}}^{UC}(\varepsilon, \delta)$) i.i.d. samples from D is ε-representative (w.r.t. $\mathcal{H}, Z, \ell, D$)

Consequence of UCP

If a class $\mathcal{H}$ has UCP with a function $m_{\mathcal{H}}^{UC}$ then the class is agnostic PAC learnable with

(1) Sample Complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}\left(\frac{\varepsilon}{2}, \delta\right)$.

(2) A is $ERM_{\mathcal{H}}$ paradigm

Next Big Question: Are finite hypothesis classes $\mathcal{H}$ agnostic PAC learnable?

(Equivalently: Does UCP hold for finite hypothesis class $\mathcal{H}$?)

CLAIM: Yes, UCP holds for finite hypothesis class $\mathcal{H}$.

PROOF: We have to find a sample size m such that (fix $\varepsilon, \delta \in (0,1)$)

$$\mathbb{P}_{X_1, Y_1 \sim D,\ S \sim D^m}\left(\{S : \forall h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \varepsilon\}\right) < \delta$$

$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}_{S \sim D^m}\left(\{S : |L_S(h) - L_D(h)| > \varepsilon\}\right).$$

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

$$L_D(h) = \mathbb{E}_{z\sim D}\left[\ell(h, z)\right]$$

(Hoeffding's Inequality) Let $\theta_1, \theta_2, \ldots \theta_m$ be a sequence of i.i.d. r.v.s and assume that $\forall i$, $\mathbb{E}[\theta_i] = \mu$ and $\mathbb{P}[a \le \theta_i \le b] = 1$. Then, $\forall \varepsilon > 0$

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

In our problem (w.l.o.g. after normalizing we can assume $\ell \in [0,1]$)

$\Rightarrow a = 0$, $b = 1$, $\mu = L_D(h)$, $\theta_i = \ell(h, z_i)$.

$$\therefore \mathbb{P}_{S\sim D^m}\left(\{S : |L_S(h) - L_D(h)| > \varepsilon\right)$$

$$\le 2\exp\left(\frac{-2m\varepsilon^2}{1}\right).$$

$$\Rightarrow \mathbb{P}_{S\sim D^m}\left(\{S : \forall h \in \mathcal{H},\ |L_S(h) - L_D(h)| > \varepsilon\}\right)$$

$$\le 2|\mathcal{H}|\exp\left(-2m\epsilon^2\right)$$

$\therefore$ ($\le \delta$ if $m \ge \log(2|\mathcal{H}|/\delta)/2\epsilon^2$

$\Rightarrow$ $\mathcal{H}$ (finite hypothesis class) w.r.t to domain $\mathcal{Z}$ and $\ell(\in [0,1])$ has UCP with

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

$\Rightarrow$ $\mathcal{H}$ is agnostic PAC learnable and $ERM_{\mathcal{H}}$ is the learning algorithm with sample complexity

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

Note: A general $\mathcal{H}$ with d parameters can be represented (and approximated) after quantizing each parameter into 64 bits as a finite class with 2^{64d} members.

It is then Agnostic PAC learnable with sample complexity $\leq \left\lceil \frac{128d + 2\log 2/\delta}{\epsilon^2} \right\rceil$

(drawback depends on d and machine representation)