More General PAC learning

$\mathcal{X}$: Domain set

$\mathcal{Y}$: Label set

Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$. ~~Let D_x be the marginal distribution over $\mathcal{X}$~~

We will assume that samples are drawn $(x,y) \sim D$

Let the risk be given by

$$L_D(h) := \underset{(x,y)\sim D}{\mathbb{P}} \left(h(x) \neq y \right) = D(\{(x,y) : h(x) \neq y\})$$

__Lemma__ (optimal Bayes predictor)
The classifier w/ minimum risk is

$$f_D(x) = \begin{cases} 1 & \text{if } \mathbb{P}(y=1 \mid x) \geq \frac{1}{2} \\ 0 & \text{else} \end{cases}$$

__pf__

Note

$$L_D(f_D) = \underset{(x,y)\sim D}{\mathbb{E}} \mathbb{1}_{\{f_D(x) \neq y\}}$$

$$(*) \quad = \underset{x \sim D_x}{\mathbb{E}} \Bigg(\underbrace{\underset{y \sim D_{y|x}}{\mathbb{E}} \left(\mathbb{1}_{\{f_D(X) \neq y\}} \mid X = x \right)}_{\text{Study this} \;=\; \mathbb{P}(f_D(X) \neq y \mid X = x)} \Bigg)$$

Let $\alpha_x = \mathbb{P}(Y = 1 \mid X = x)$

$$\mathbb{P}(f_D(X) \neq y \mid X = x) = \underbrace{\overset{\mathbb{1}_{\{\alpha_x \geq 1/2\}}}{\cancel{\mathbb{P}(f(X) = 0 \mid X = x)}}}_{\in \{0,1\}} (1 - \alpha_x) + \underbrace{\overset{\mathbb{1}_{\{\alpha_x < \frac{1}{2}\}}}{\cancel{\mathbb{P}(f(X) = 1 \mid X = x)}}}_{\in \{0,1\}} \alpha_x$$

$$\geq \min\left((1-\alpha_x), \alpha_x \right)$$

But also, suppose g is any other classifier, possibly stochastic. Then

$$\mathbb{P}(g(X) \neq Y \mid X = x) = \mathbb{P}(g(X) = 1 \mid X = x)\,\mathbb{P}(Y = 0 \mid X = x)$$
$$+ \mathbb{P}(g(X) = 0 \mid X = x)\,\mathbb{P}(Y = 1 \mid X = x)$$

$$= \mathbb{P}(g(X) = 1 \mid X = x)(1-\alpha_x) + \mathbb{P}(g(X) = 0 \mid X = x)\,\alpha_x$$

$$\geq \min\{(1-\alpha_x), \alpha_x\}$$

Hence, ~~$\mathbb{E}(1_{f_D(X)})$~~

going back to $(*)$ we see that $\mathbb{E}_{X \sim D_x}(1_{\{f_D(X) = y\}} \mid X = x) \leq \mathbb{E}(1_{\{g(X) \neq y\}} \mid X = x)$ $\forall x$

$$\cancel{(*) = \mathbb{E}_{X \sim D_x} \min(\alpha_x, 1-\alpha_x)}$$

$$(*) \leq \mathbb{E}_{X \sim D_x}\left(\mathbb{E}_{Y \sim D_{y|X}}\left(1_{\{g(X) \neq y\}} \mid X = x\right)\right)$$

$$= \mathbb{E}_{(x,y) \sim D}\left(1_{g(X \neq y)}\right)$$

∎

Def (Agnostic PAC learnability)

Hypothesis class $\mathcal{H}$ is agnostic PAC learnable if $\exists$ $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ & learning algo A s.t. : $\forall$ $\varepsilon, \delta \in (0,1)$ & every distribution D over $\mathcal{X}, \mathcal{Y}$, running A on $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ i.i.d samples from D, A returns a hypothesis h s.t. w.p. at least $1-\delta$

$$L_D(h) \leq \min_{h' \in \mathcal{H}} L_D(h') + \varepsilon,$$

where $L_D(h) = \mathbb{E}_{z \sim D} \lambda(h, z)$

Note: Above, we consider general loss function ℓ. e.g.

- $$\lambda_{0-1}(h, (x,y)) := \begin{cases} 0 & \text{if } h(x) = y \\ 1 & h(x) \neq y \end{cases}$$

- $$\ell_{sq}(h, (x,y)) := (h(x) - y)^2$$

Uniform Convergence

Def (ε-rep. sample) A training set S is called ε-representative if

$$|L_S(h) - L_D(h)| < \varepsilon \quad \forall h \in \mathcal{H}$$

Lemma Assume S is $\varepsilon/2$-rep. Then any $h_S \in ERM_{\mathcal{H}}(S)$ satisfies

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

pf

$$\forall h, \quad L_D(h_S) \overset{\varepsilon\text{-rep}}{\leq} L_S(h_S) + \frac{\varepsilon}{2} \overset{h_S \in ERM}{\leq} L_S(h) + \frac{\varepsilon}{2} \leq L_D(h) + \varepsilon.$$

Minimizing over h completes the proof. ■

<u>Def</u> (uniform convergence) We say a hypo class $\mathcal{H}$ has the <u>uniform conv property</u> if $\exists$ $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ s.t. $\forall \varepsilon, \delta \in (0,1)$ & $\forall$ prob distribution D over Z, if S is a sample of $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ iid draws, then w.p. at least $(1-\delta)$ S is ε-rep.

<u>Note</u> If $\mathcal{H}$ has uniform conv property w/ func $m_{\mathcal{H}}^{UC}$, then $\mathcal{H}$ is agnostic PAC learnable w/ sample complexity $m_{\mathcal{H}}(\varepsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\frac{\varepsilon}{2}, \delta)$. Also, the ERM paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

<u>Claim</u> Let $\mathcal{H}$ be a finite hypothesis class, Z be a domain, & let ~~$f: \mathcal{H} \times Z \to \mathbb{R}$~~ $\ell : \mathcal{H} \times Z \to [0,1]$ be a loss function. Then $\mathcal{H}$ has the uniform conv. property w/ sample complexity

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

Fix ε, δ. W.t.s.

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

equivalently,

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

write

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

union bd gives

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

Now, just want to bound terms inside sum. Note that

$$L_S(h) = \frac{1}{m} \sum_{i=1}^{m} \underbrace{l(h, z_i)}_{\substack{\text{has} \\ \text{mean } \mathbb{E}_{z \sim D}(l(h,z)) =: L_D(h)}}$$

So,

$$\mathbb{E}(L_S(h)) = L_D(h).$$

Recall Hoeffding ineq: (θ_i) iid w/ $\mathbb{E}\,\theta_i = \mu$ &
$\mathbb{P}(a \le \theta_i \le b) = 1$. Then $\forall \varepsilon > 0$,

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

Applying Hoeffding we see

$$D^m\left(\{ S : |L_S(h) - L_D(h)| > \varepsilon \} \right) \le 2 \exp\left(\frac{-2m\varepsilon^2}{1} \right)$$

Hence, w/ (*) this implies

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

$$\le 2|\mathcal{H}|\, e^{-2m\varepsilon^2} \quad \blacksquare$$