COL774
Nov 3rd, 2021

Last Class:- Learning Theory

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ $y^{(i)} \in \{0, 1\}$ (classification)

① $\mathcal{H}$:- finite
② $\mathcal{H}$:- infinite sized → (VC-Dim)

i.i.d ← $(x^{(i)}, y^{(i)}) \sim$ Dist.

Goal:- $h \in \mathcal{H}$ → hypothesis space

$\forall i$ $h(x^{(i)}) \approx y^{(i)}$ (minimizing training error)

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$$:- training error

Eventually..

$$\varepsilon(h) = E_{(x,y)\sim \text{Dist.}}\left[1\{h(x) \neq y\}\right]$$
→ generalization error

[$\hat{h}$:- best training data
→ h^* :- "optimal" on entire data]

$\forall h \in \mathcal{H}$
⇒ $|\hat{\varepsilon}(h) - \varepsilon(h)|$

for any given hypothesis $h \in \mathcal{H}$
$$\Pr\left(|\hat{\varepsilon}(h) - \varepsilon(h)| > \gamma\right) \leq 2e^{-2\gamma^2 m}$$

→ PAC bound:-
Probably Approximately Correct (Bound)

→ $\Pr(\exists h \in \mathcal{H} :- |\hat{\varepsilon}(h) - \varepsilon(h)| > \gamma)$ — ①
$\leq |\mathcal{H}| \cdot 2e^{-2\gamma^2 m}$ margin of error

→ Property P:-

⇒ equivalent
$$\boxed{\Pr(\exists h \in \mathcal{H} : |\hat{\varepsilon}(h) - \varepsilon(h)| > \gamma) \leq \delta}$$ — ② ≈ ②

$\delta \geq |\mathcal{H}| \cdot 2e^{-2\gamma^2 m}$ confidence PAC probably ... bounds

$$\Pr\left[\forall h \in \mathcal{H} : \left|\hat{\varepsilon}(h) - \varepsilon(h)\right| \leq \gamma\right) \geq 1-\delta$$

$$\Pr\left[\exists h \in \mathcal{H} : \left|\hat{\varepsilon}(h) - \varepsilon(h)\right| > \gamma\right) \geq 1-\delta$$

↳ Aside:

Coin Toss Experiment: $z^{(i)} \in \{0,1\}$

ϕ :- Bernoulli parameter

$\{z^{(i)}\}_{i=1}^{m}$ $\quad z^{(i)} \sim \text{Bernoulli}(\phi)$

$$\hookrightarrow \hat{\phi} = \frac{1}{m}\sum_{i=1}^{m} z^{(i)}$$

$\Rightarrow \hat{\phi}$:- random variable

↳ estimator for ϕ

$$E[\hat{\phi}] = E\left[\frac{1}{m}\sum_{i=1}^{m} z^{(i)}\right]$$

$\{z^{(i)}\}_{i=1}^{m}$

$\{z^{(i)}\} \sim \text{Bernoulli}(\phi)$

$$= \frac{1}{m}\sum_{i=1}^{m} E[z^{(i)}]$$

$$= \frac{1}{\cancel{m}} \times \cancel{m} \times [\phi]$$

$$= \phi$$

$\{z^{(i)}\}_{i=1}^{m}$ $\quad m = 100$

(1) #h :- 57 #t :- 43

$\hat{\phi} = .57$

(2) #h :- 53 #t :- 47

$\hat{\phi} = .53$

(3) #h :- 67 #t :- 33

$\hat{\phi} = .67$

$\Rightarrow \hat{\phi}$:- unbiased estimator of ϕ

↳ $E[\hat{\phi}] = \underline{\phi}$

$$Var[\hat{\phi}] = Var\left[\frac{1}{m}\sum_{i=1}^{m} z^{(i)}\right]$$

$$= \frac{1}{m^2} Var\left[\sum_{i=1}^{m} z^{(i)}\right]$$

$$= \frac{1}{m^2} \sum_{i=1}^{m} \underbrace{Var[z^{(i)}]}$$

$(\because z^{(i)}$'s are i.i.d.)

$z \sim Bernoulli(\phi)$

$\hookrightarrow Var[z] = E[z^2] - (E[z])^2$

$= \phi(1-\phi)$

$$Var[\hat{\phi}] = \frac{1}{m^2} \times m \times \phi(1-\phi)$$

$$= \frac{\phi(1-\phi)}{m}$$

↳ Goes down as # of examples is increased

Hoeffding inequality / Chernoff Bound:-

$$Pr(|\phi - \hat{\phi}| > \gamma) \leq 2e^{-2\gamma^2 m}$$

ϕ ↓ true parameter, $\hat{\phi}$ ↓ estimator

↳ PAC bounds for supervised learning setting

$\{x^{(i)}, y^{(i)}\}_{i=1}^{m}$ $\quad (x^{(i)}, y^{(i)}) \sim Dist$

$$z^{(i)} = 1\{h(x^{(i)}) \neq y^{(i)}\} \quad h \in \mathcal{H}$$

↳ error of i^{th} example

$$\hat{\varepsilon}(h) = \frac{1}{m} \sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\} \quad z^{(i)} \sim Bernoulli(\phi)$$

$(x^{(i)}, y^{(i)}) \sim Dist$

$$1\{h(x^{(i)}) \neq y^{(i)}\} = z^{(i)}$$

$z^{(i)} \sim Bernoulli(\phi)$ $\quad \phi = \varepsilon(h)$

$$\underset{(x^{(i)}, y^{(i)}) \sim Dist}{E}\left[1\{h(x^{(i)}) \neq y^{(i)}\}\right] = ??$$

$$= \underset{(x,y) \sim Dist}{E}\left[1\{h(x) \neq y\}\right]_{z^{(i)}} \because \text{ Generalization error}$$

$$\frac{1}{m} \sum_{i=1}^{m} \overbrace{1\{h(x^{(i)}) \neq y^{(i)}\}}^{z^{(i)}}$$

$$\Rightarrow \hat{\varepsilon}_u(h) = \frac{1}{m} \sum_{i=1}$$

$\hookrightarrow$ estimator of $\varepsilon(h)$

$$\text{s.t } E[\hat{\varepsilon}_u(h)] = \frac{1}{m} \sum_{i=1}^{m} E[\underbrace{1\{h(x^{(i)}) \neq y^{(i)}\}}]$$

estimator for an underlying Bernoulli distribution $= \underbrace{\varepsilon_u(h)}$:-

$\rightarrow$ Chernoff bound

$$\Pr(\ |\underbrace{\hat{\varepsilon}_u(h)}_{\hat{\phi}} - \underbrace{\varepsilon_u(h)}_{\phi}| > \gamma) \le 2e^{-2\gamma^2 m}$$

$\hookrightarrow$ $h_e :- \in \mathcal{H}$

$$\underbrace{\Pr(\ |\hat{\varepsilon}_u(h_e) - \varepsilon_u(h)| > \gamma)} \le \underbrace{2e^{-2\gamma^2 m}}$$

$|\mathcal{H}| = K$

$$A_e :- \quad |\hat{\varepsilon}_u(h_e) - \varepsilon_u(h_e)| > \gamma$$

$$\Rightarrow P(\underbrace{A_1 \cup A_2 \cdots A_K}_{\Downarrow}) \le \sum_{e=1}^{K} \underbrace{P(A_e)}$$

$$\left[\Pr(\underbrace{\exists h_e \in \mathcal{H}}\ |\hat{\varepsilon}_u(h_e) - \varepsilon_u(h_e)| > \gamma) \le \cancel{\phi}\, K \cdot \underbrace{2e^{-2\gamma^2 m}} \right]$$

Uniform convergence bound