End lecture 2.

Relaxing the realizability assumption

Data generating distribution $\mathcal{D}$ is over X, Y : $z=(x,y)\sim\mathcal{D}$

True error: $L_{\mathcal{D}}(h) \triangleq P_{\mathcal{D}}(h(x)\neq y)$

Empirical error: $L_S(h) \triangleq \frac{|\{i\in[m]: h(x_i)\neq y_i\}|}{m}$

Goal: to find h that minimizes $L_{\mathcal{D}}(h)$

Bayes Optimal Predictor has lower error than any other predictor [HW1 Prob2]

$$f_{\mathcal{D}}(x) = \begin{cases} 1 & \text{if } P_{\mathcal{D}}(y=1|x) \geq 1/2 \\ 0 & \text{otherwise} \end{cases}$$

but requires knowledge of $\mathcal{D}$.

Agnostic PAC Learnability.

A class of hypotheses, $\mathcal{H}$, is **Agnostic PAC Learnable** if $\exists m_{\mathcal{H}}$ and a learning algorithm s.t. for $\varepsilon, \delta, \mathcal{D}$ running Algo on $m \geq m_{\mathcal{H}}(\varepsilon,\delta)$ samples returns h s.t. w.p $\geq 1-\delta$

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

Relaxing Binary Classification

- Multi-class classification $\mathcal{Y}=[K]$
- Regression, $\mathcal{Y}\subseteq\mathbb{R}$

loss $\ell_h(x_i,y_i) = \mathbb{I}(h(x_i)\neq y_i)$

loss $\ell_h(x_i,y_i) = (h(x_i)-y_i)^2$

Risk function $L_{\mathcal{D}}(h) \triangleq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell(h,(x,y))]$

Empirical Risk function $L_S(h) \triangleq \frac{1}{m}\sum_{i=1}^{m}\ell(h,(x_i,y_i))$

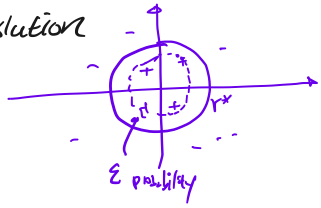
Q3.3 $X=\mathbb{R}^2$, $\mathcal{Y}=\{0,1\}$, $\mathcal{H}=\{h_r : r\in\mathbb{R}_+\}$

$$h_r(x) = \mathbb{I}_{\{\|x\|\leq r\}}$$

Prove that $\mathcal{H}$ is PAC learnable (assuming realizability), with

$$m_{\mathcal{H}}(\varepsilon,\delta) \leq \left\lceil \frac{\log 1/\delta}{\varepsilon} \right\rceil$$

solution



$\forall X, \mathcal{D}, f_{r^*}$, define $\mathcal{H}_{Bad} = \{r : r\leq \tilde{r}, \mathcal{D}(\{x : r\leq\|x\|\leq r^*\}) \geq \varepsilon\}$.

Algo: return smallest circle that includes all $+$.

$$P_S(\text{Algo} < r^*) \leq (1-\varepsilon)^m \leq e^{-\varepsilon m} \leq \delta.$$

for $m \geq \frac{1}{\varepsilon}\log\frac{1}{\delta}$

---

chapter 4 Uniform Convergence is sufficient for Learnability

Definition. (ε-representative sample)

Training set S is ε-representative if

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

Lemma. Assume we are given a training set S that is $\frac{\varepsilon}{2}$-representative. Then, any output of ERM $h_S \in \arg\min_{h\in\mathcal{H}} L_S(h)$, satisfies

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

Proof strategy: show S ε-representative $\Rightarrow$ use lemma to show ERM is

Proof. by triangular inequality, i.e. $a \le b + |a-b|$

$$L_D(h_S) \le |L_D(h_S) - L_S(h_S)| + L_S(h_S)$$
$$\le \frac{\epsilon}{2} + \underbrace{L_S(h_S) - L_S(h)}_{\le 0 \text{ by ERM}} + L_S(h)$$
$$\le \frac{\epsilon}{2} + 0 + |L_S(h) - L_D(h)| + L_D(h)$$
$$\le \frac{\epsilon}{2} + 0 + \frac{\epsilon}{2} + L_D(h), \quad \text{for } {}^\forall h \in \mathcal{H}|$$

To show ERM is agnostic PAC Learner, we are left to show that with probability $1-\delta$, S is ϵ-representative.

Definition (Uniform Convergence)
A hypothesis class $\mathcal{H}$ has the uniform convergence property, if $\boxed{\exists\, m_{\mathcal{H}}^{UC}}$ s.t for every ϵ, δ and every D, if $m \ge m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ the S is ϵ-representative with probability at least $1-\delta$.

Corollary 1. $\mathcal{H}$ is agnostically PAC learnable with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \le m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right)$$
And ERM is the agnostic PAC Learner.

Next, we want to show finite $\mathcal{H}$ is agnostic PAC Learnable.
↑ show uniform convergence.

Claim. Any finite $\mathcal{H}$, i.e., $|\mathcal{H}| < \infty$, is uniformly convergent with
$$m_{\mathcal{H}}^{UC}(\epsilon, \delta) \le \left\lceil \frac{\log \frac{2|\mathcal{H}|}{\delta}}{\epsilon} \right\rceil$$

Corollary 2. ERM is agnostic PAC Learner with sample complexity
$$m_{\mathcal{H}}(\epsilon, \delta) \underset{\text{Corollary 1.}}{\le} m_{\mathcal{H}}^{UC}\left(\frac{\epsilon}{2}, \delta\right) \underset{\text{claim}}{\le} \left\lceil \frac{2\log \frac{2|\mathcal{H}|}{\delta}}{\epsilon} \right\rceil$$

We are left to prove the Claim.
= Q. What is the sample size that guarantees, assuming $0 \le \ell(h, z_i) \le 1$
$|L_S(h) - L_D(h)| \le \epsilon$ w.p $1-\delta$?

n.t.s (need to show) $\quad \mathbb{P}_S(\{S : {}^\forall h \in \mathcal{H}, |L_S(h) - L_D(h)| \le \epsilon\}) \ge 1-\delta$

n.t.s $\quad \mathbb{P}_S(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_D(h)| > \epsilon\}) < \delta$

$$\underset{\text{union bound}}{\le} \sum_{h \in \mathcal{H}} \mathbb{P}_S(\{S : |L_S(h) - L_D(h)| > \epsilon\}) \qquad (*)$$

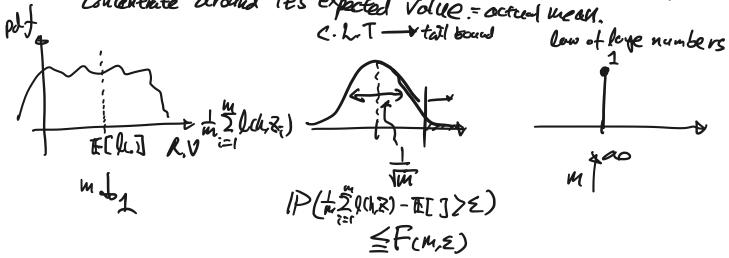
$L_S(h) = \frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i)$ ← average of iid R.V. $\quad z_i = (x_i, y_i)$

$L_D(h) = \mathbb{E}_{z \sim D}[\ell(h,z)] = \mathbb{E}[L_S(h)]$. ← expectation of each term

We need to show that the measure (= distribution) of R.V. $L_S(h)$ concentrates around its mean.

* most important tool in statistical analysis for learning. = concentration of measure = tail bound

Q. How fast (in terms of # of samples) does the measure of empirical mean concentrate around its expected value = actual mean.



C.L.T → tail bound

$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\ell(h,z) - \mathbb{E}[\,] > \epsilon\right) \le F(m, \epsilon)$

law of large numbers

---

* Hoeffding's Inequality

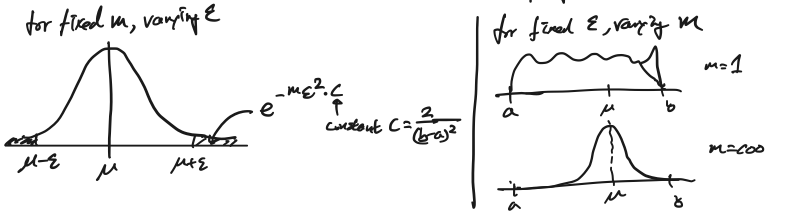
Let $\theta_1, \dots, \theta_m$ be a set of i.i.d random variables with $\mu \triangleq \mathbb{E}[\theta_i]$, satisfying

$$\mathbb{P}(a \le \theta_i \le b) = 1, \text{ for all } i \in [m] \qquad \text{[Boundedness]}$$

then for any $\varepsilon > 0$,

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

first example of < concentration of measure / tail bound



Letting $\theta_i = \ell(h, z_i)$, and assuming $a \le \ell(h, z_i) \le 1 = b$, and $\mu = L_D(h)$

$$\mathbb{P}_S\left(|L_S(h) - L_D(h)| > \varepsilon \right) = \mathbb{P}\left(\left| \frac{1}{m} \sum_{i=1}^{m} \theta_i - \mu \right| > \varepsilon \right)$$

Hoeffding's $\rightarrow \le 2 \cdot e^{-2m\varepsilon^2}$

$$(*) \Rightarrow \mathbb{P}_S\left(\{\exists h \in \mathcal{H} : |L_S(h) - L_D(h)| > \varepsilon\} \right) \le \sum_{h \in \mathcal{H}} 2e^{-2m\varepsilon^2}$$

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

Letting $m(\varepsilon, \delta) \ge \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\varepsilon^2}$, $\le \delta$

Summary

if $m \ge \frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2\varepsilon^2}$ then $\mathcal{H}$ has uniform convergence

Lemma: Uniform Convergence $\Leftrightarrow$ $m_{\mathcal{H}}^{UC}(\varepsilon, \delta) \ge m_{\mathcal{H}}(\varepsilon, \delta)$

(ε, δ)

$\downarrow$

(ε, δ)-PAC learnable with ERM

$$\text{if } m \ge \left\lceil \frac{2 \cdot \log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{\varepsilon^2} \right\rceil \ge m_{\mathcal{H}}^{UC}(\varepsilon/2, \delta) \ge m_{\mathcal{H}}(\varepsilon, \delta)$$

then $\mathcal{H}$ is (ε, δ)-PAC learnable with ERM.

End of lecture 3.

Concentration of measure

The more you know about the R.V. θ_i, the tighter concentration you can prove.

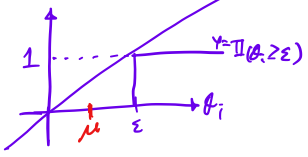
① [Markov's inequality]

If $\theta_i \ge 0$ w.p.1 then $\mathbb{P}(\theta_i \ge \varepsilon) \le \frac{\mu}{\varepsilon}$ — no concentration, why? tight?

and $\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m} \theta_i \ge \varepsilon\right) \le \frac{\mu}{\varepsilon} = O(1)$

Proof: $\mathbb{P}(\theta_i \ge \varepsilon) = \mathbb{E}[\mathbb{I}_{(\theta_i \ge \varepsilon)}]$

$= \int_{\varepsilon}^{\infty} f_{\theta_i}(t)\, dt$



$\theta_i \ge 0 \rightarrow \le \mathbb{E}\left[\frac{1}{\varepsilon}\theta_i\right]$

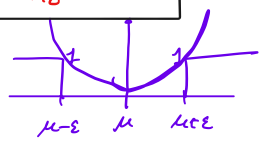
$= \frac{\mu}{\varepsilon}$

② [2nd order statistics, Chebyshev's inequality]

If $\mathrm{Var}(\theta_i) \le \sigma^2$ then $\mathbb{P}(|\theta_i - \mu| \ge \varepsilon) \le \frac{\sigma^2}{\varepsilon^2}$

and $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| \ge \varepsilon\right) \le \frac{\sigma^2}{m\varepsilon^2} = O\left(\frac{1}{m}\right)$ — variance $\times \frac{1}{m}$

Proof: $\mathbb{P}(|\theta_i - \mu| \ge \varepsilon) = \mathbb{E}[\mathbb{I}_{(\theta_i \le \mu-\varepsilon \text{ or } \theta_i \ge \mu+\varepsilon)}]$



Guided by the fact that $\mathrm{Var} \le \sigma^2$ $\rightarrow \le \mathbb{E}\left[\frac{1}{\varepsilon^2}(\theta_i - \mu)^2\right]$

$\le \frac{\sigma^2}{\varepsilon^2}$

③ [bounded domain, Hoeffding's inequality]

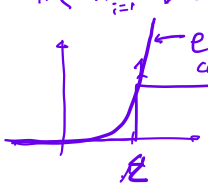
If $a \le \theta_i \le b$ then $\mathbb{E}\left[e^{\lambda(\theta_i - \mu)}\right] \le e^{\frac{\lambda^2(b-a)^2}{8m^2}}$ (**)

and $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} \theta_i - \mu\right| \ge \varepsilon\right) \le 2 \cdot \exp\left(-\frac{2m\varepsilon^2}{(b-a)^2}\right) = O(e^{-m\varepsilon^2})$

Generic Recipe

special to Hoeffding's

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \theta_i - \mu \geq \varepsilon\right) = \mathbb{E}\left[\mathbb{I}\left(\frac{1}{m}\sum_{i=1}^m \theta_i - \mu \geq \varepsilon\right)\right]$$



$$\leq \mathbb{E}\left[e^{\lambda\left(\frac{1}{m}\sum_{i=1}^m \theta_i - \mu\right) - \varepsilon}\right]$$

$$\leq e^{-\lambda\varepsilon} \cdot \prod_{i=1}^m e^{\left(\lambda \frac{\theta_i - \mu}{m}\right)}$$

$$\leq e^{-\lambda\varepsilon} \prod_{i=1}^m e^{\frac{\lambda^2 (b-a)^2}{8m^2}}$$

opt over λ $\rightarrow$ $\leq \exp\left(-\lambda\varepsilon + \frac{\lambda^2 (b-a)^2}{8m}\right)$

$-\varepsilon + \frac{2\lambda (b-a)^2}{8m} = 0$ $\leq \exp\left(\frac{-2m\varepsilon^2}{(b-a)^2}\right)$

$\lambda^* = \frac{4\varepsilon m}{(b-a)^2}$

do the same for $\frac{1}{m}\sum_{i=1}^m \theta_i < \mu - \varepsilon$

(**) $\mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right] \leq e^{\frac{\lambda^2 (b-a)^2}{8m^2}}$

$\mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right] \leq \max_{[a,b]} \mathbb{E}\left[e^{\frac{\lambda(\theta_i - \mu)}{m}}\right]$

convexity of $\exp(\cdot)$ $\rightarrow$ $\leq \max_{p_a p_b} p_a e^{\frac{\lambda(a-\mu)}{m}} + p_b e^{\frac{\lambda(b-\mu)}{m}}$

$p_a + p_b = 1$, $a p_a + b p_b = \mu$.

$\leq e^{-\frac{\lambda\mu}{m}} \cdot \max_{p_a p_b} p_a e^{\frac{\lambda a}{m}} + p_b e^{\frac{\lambda b}{m}}$

$\leq e^{-\frac{\lambda\mu}{m}} \cdot \exp\left[\frac{\lambda a p_a + \lambda b p_b}{m} + \frac{(b-a)^2 \lambda^2}{8m^2}\right]$

$\leq \exp\left(\frac{(b-a)^2 \lambda^2}{8m^2}\right)$

$\Uparrow$

* Chernoff's Bound

$\theta_1 \cdots \theta_m$ independent but not identically distributed Bernoulli dist. $\theta_i \begin{cases} 1 & \text{w.p. } p_i \\ 0 & 1 - p_i \end{cases}$

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^m \theta_i > (1+\delta)\frac{1}{m}\sum_{i=1}^m p_i\right) \leq e^{-\overbrace{\mathcal{C}(\delta)}^{\geq (1+\delta)\ln(1+\delta) - \delta} \cdot \underbrace{\sum_{i=1}^m p_i}_{\Theta(m)}} \leq e^{-\frac{\delta^2}{2 + \frac{2}{3}\delta} \sum_{i=1}^m p_i}$$

multiplicative

$\geq \frac{\delta^2}{2 + \frac{2}{3}\delta}$

2 phase