Lecture 12. Concentration of measure.

* Recap lecture 11.

Claim. ERM is agnostic PAC learner with

$$m_{\mathcal{H}}(\varepsilon,\delta) \le \left\lceil \frac{\log\frac{|\mathcal{H}|}{\delta}}{2\varepsilon^2} \right\rceil$$

Corollary 1: $(\frac{\varepsilon}{2},\delta)$-Uniform convergence $\longrightarrow$ PAC

$\longleftrightarrow$ $m_{\mathcal{H}}(\varepsilon,\delta) \le m_{\mathcal{H}}^{U.C}(\varepsilon/2,\delta)$

left to show: $m \ge \left\lceil \frac{\log\frac{|\mathcal{H}|}{\delta}}{2\varepsilon^2} \right\rceil \longrightarrow$ Uniform Convergence w.p $1-\delta$

$$\mathbb{P}_S\left(\bigcup_{h\in\mathcal{H}} \{ |L_S(h) - L_D(h)| > \varepsilon \}\right) \le \delta$$

n.t.s. $|\mathcal{H}| \cdot \mathbb{P}_S(|L_S(h) - L_D(h)| > \varepsilon) \le \delta$

$m \ge \left\lceil \frac{\log\frac{2|\mathcal{H}|}{\delta}}{2\varepsilon^2} \right\rceil$ n.t.s $\longrightarrow$ $\mathbb{P}_S\left(\left|\frac{1}{m}\sum_{i=1}^{m} \ell(h,z_i) - L_D(h)\right| > \varepsilon\right) \le \frac{\delta}{|\mathcal{H}|}$

assuming $0 \le \ell(h,z) \le 1$. apply Hoeffding's inequality.

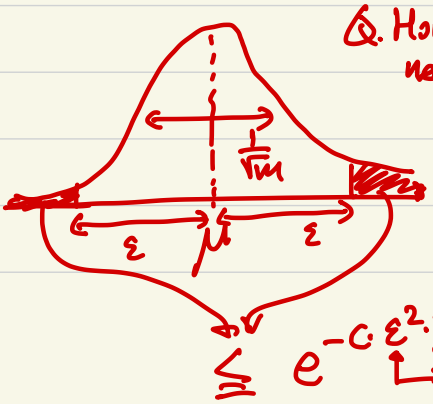
* Hoeffding's inequality.

Let $\theta_1, \theta_2, \dots, \theta_m$ be a set of iid R.V.s with $\mu = \mathbb{E}[\theta_i]$, with $a \le \theta_i \le b$, $\forall i \in [m]$. 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.

for fixed m, varying ε.

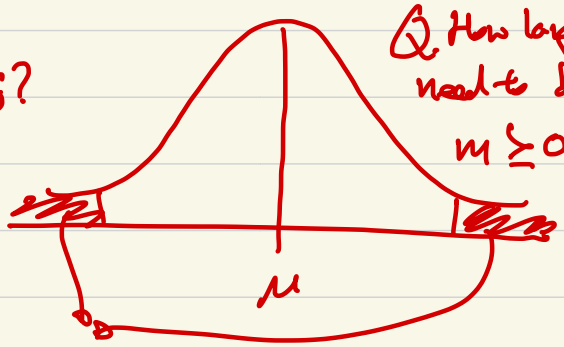


Q. How large does ε need to be for tail $\le \delta$?

$\varepsilon = O(\frac{1}{\sqrt{m}}\log\frac{1}{\delta})$

$\le e^{-c\cdot\varepsilon^2 \cdot m}$: sub Gaussian tail

for varying m



Q. How large does m need to be for tail $\le \delta$

$m \ge O(\frac{1}{\varepsilon^2}\log\frac{1}{\delta})$

$\le \delta$

letting $\theta_i = \ell(h, z_i)$

$0 \le \ell(h, z_i) \le 1$

$a \qquad\qquad b$

$\mu = \mathbb{E}[\ell(h, z)] = L_D(h)$

Hoeffding's $\Rightarrow$ $\mathbb{P}(|L_S(h) - L_D(h)| > \varepsilon) = \mathbb{P}(|\frac{1}{m}\sum_{i=1}^{m} \ell(h, z_i) - L_D(h)| > \varepsilon)$

$\le 2 \cdot e^{-2m\varepsilon^2}$ need $\le \frac{\delta}{|\mathcal{H}|}$

need: $e^{-2m\varepsilon^2} \le \frac{\delta}{2|\mathcal{H}|}$

need: $m \ge \left\lceil \frac{\log(\frac{2|\mathcal{H}|}{\delta})}{2\varepsilon^2} \right\rceil$

---

* Concentration of measure.
  - The more you know about the R.V. θ_i, the tighter concentration you can prove.

① [least information , Markov's Inequality]

if $\theta_i \ge 0$, $\mathbb{E}[\theta_i] = \mu < \infty$ then $\mathbb{P}(\theta_i \ge \varepsilon) \le \frac{\mu}{\varepsilon}$

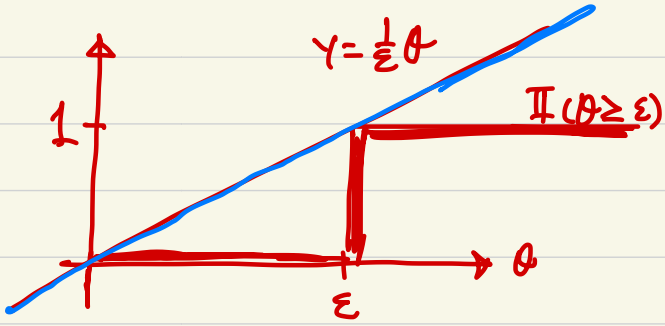
and $\mathbb{P}(\frac{1}{m}\sum_{i=1}^{m} \theta_i \ge \varepsilon) \le \frac{\mu}{\varepsilon} = O(1)$: no concentration

proof> $\mathbb{P}(\theta \ge \varepsilon) = \int_{\varepsilon}^{\infty} f_\theta(t)\,dt$

$= \mathbb{E}[\mathbb{I}_{\{\theta \ge \varepsilon\}}]$

$\theta \ge 0 \rightarrow \le \mathbb{E}[\frac{1}{\varepsilon}\theta]$

$\mathbb{E}[\theta] = \mu \quad = \frac{\mu}{\varepsilon}$



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

If $\mathbb{E}[\theta]=\mu$, $\mathrm{Var}[\theta]\leq\sigma^2$ then $\mathbb{P}(|\theta_i-\mu|\geq\varepsilon)\leq\frac{\sigma^2}{\varepsilon^2}$.

and $\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m}\theta_i-\mu\right|\geq\varepsilon\right)\leq\frac{\sigma^2}{m\varepsilon^2}=O\left(\frac{1}{m}\right)$ Concentration

why?

Proof>

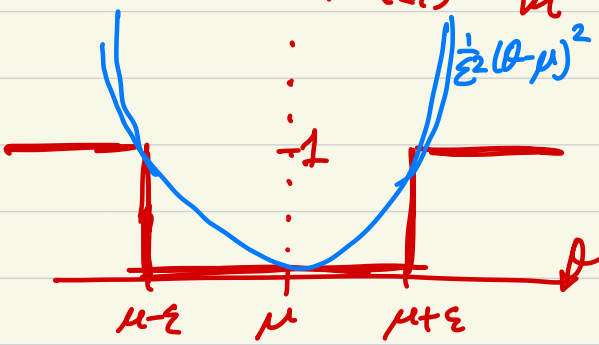
$\mathbb{P}(|\theta_i-\mu|\geq\varepsilon)=\mathbb{E}[\mathbb{I}_{\{\theta\leq\mu-\varepsilon \text{ or } \theta\geq\mu+\varepsilon\}}]$

$\leq\mathbb{E}\left[\frac{1}{\varepsilon^2}(\theta-\mu)^2\right]$

$\mathrm{Var}(\theta)\leq\sigma^2\rightarrow=\frac{\sigma^2}{\varepsilon^2}$

$\bar{\theta}_i=\frac{1}{m}\sum_{i=1}^{m}\theta_i\rightarrow\mathbb{E}[\bar{\theta}_i]=\mu$

$\mathrm{Var}(\bar{\theta}_i)=\frac{\sigma^2}{m}$



---

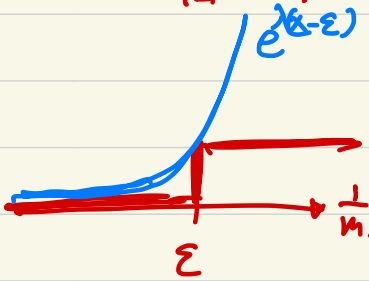
③ [bounded domain , Hoeffding's inequality]

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

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

Generic Recipe $\Longrightarrow$ Hoeffding's.

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



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

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

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

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

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

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

minimize over $\lambda \geq 0$

$$\frac{\partial}{\partial\lambda} = -\varepsilon + \frac{\lambda(b-a)^2}{4m} \rightarrow \leq \exp\left\{-\frac{2\varepsilon^2 m}{(b-a)^2}\right\}$$

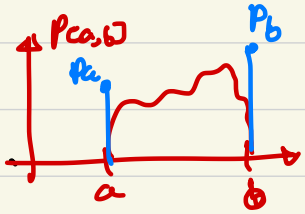
$$\lambda^* = \frac{4\varepsilon m}{(b-a)^2}$$

*Repeat for $\mathbb{I}\{+\frac{1}{m}\sum_{i=1}^{m}\theta_i < \mu - \varepsilon\}$

---

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

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



Convexity of $\exp(\cdot)$

$$\leq \max_{P_a, P_b} P_a \cdot e^{\frac{\lambda(a-\mu)}{m}} + P_b \cdot e^{\frac{\lambda(b-\mu)}{m}}$$

s.t. $P_a + P_b = 1$

$a P_a + b P_b = \mu$

$$\leq e^{\frac{-\lambda\mu}{m}} \cdot \left\{\max_{P_a, P_b} P_a e^{\frac{\lambda a}{m}} + P_b e^{\frac{\lambda b}{m}}\right\}$$) omit proof.

$$\leq e^{\frac{-\lambda\mu}{m}} \cdot \left\{\max_{P_a, P_b} \exp\left\{\frac{\lambda a}{m}P_a + \frac{\lambda b}{m}P_b + \frac{(b-a)^2\lambda^2}{8m^2}\right\}\right\}$$

$$= e^{\frac{-\lambda\mu}{m}} \cdot \left\{\max_{P_a, P_b} \exp\left\{\frac{\lambda}{m}\mu + \frac{(b-a)^2\lambda^2}{8m^2}\right\}\right\}$$

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

---

④ [Bernoulli dist, Chernoff bound]

$\theta_i = \begin{cases} 1 & \text{w.p } p_i \\ 0 & 1-p_i \end{cases}$

$$\mathbb{P}\left(\frac{1}{m}\sum_{i=1}^{m}\theta_i > (1+\delta)\underbrace{\frac{1}{m}\sum_{i=1}^{m}p_i}_{\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\theta_i\right]}\right) \leq e^{-\underbrace{\frac{\delta^2}{2+\frac{2}{3}\delta}}_{\rightarrow O(\delta^2) \text{ or } O(\delta)} \cdot \underbrace{\sum_{i=1}^{m}p_i}_{O(m)}}$$

multiplicative bound vs. additive bound $\frac{1}{m}\sum_{i=1}^{m}p_i + \varepsilon$

independent but not necessarily identically distributed.

$$\mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\theta_i\right] = \frac{1}{m}\sum_{i=1}^{m}p_i$$

Recap.

lecture 10. (ch 2 & 3)

- PAC learning, ← realizable, $Y=\{0,1\}$, $|\mathcal{H}|<\infty$
- technique: union bound.

lecture 11. (ch 4)

- Agnostic PAC learning ← non-realizable, any loss (bounded), $|\mathcal{H}|<\infty$
- technique: Hoeffding's ineq.

Q 3.3 $\mathcal{X}=\mathbb{R}^2$, $Y=\{0,1\}$, $\mathcal{H}=\{h_r : r\in\mathbb{R}_+\}$, $|\mathcal{H}|=\infty$.

$h_r = \mathbb{1}\{\|x\|\le r\}$

$$m_{\mathcal{H}} = \frac{\log \frac{1}{\delta}}{\varepsilon}$$ → How?

Chapter 5.

the reason the above problem has low sample complexity for PAC learning is that we had a <u>prior knowledge</u>.

- realizable: ground truth D, f, has a circular boundary
  $\mathcal{H}$ is a collection of circular decision boundaries.

Q. is such prior knowledge necessary for PAC learning? Yes.

Learning Problem

fix $\mathcal{X}, Y$

nature D → S → Learner H, H_{fo}

$\Downarrow$

h

Goal: $L_D(h)\downarrow$

Q. Can there be a universal Algorithm that is PAC learnable for all $\mathcal{D}$?

Q. Is prior knowledge necessary for PAC learning?

- lec 10: realizable w.r.t. $\mathcal{H}$.
- lec 11: $\min_{h \in \mathcal{H}} L_{\mathcal{D}(H)}$ is small (implicit in Agnostic PAC)
- Q3.3: parametric family $\mathcal{D}$

Q. How much prior knowledge is good to assume?

a lot

- learner knows $\mathcal{D}$ $\rightarrow h^*$: Bayes optimal predictor
- learner knows a class
  - lec 10: realizable, $|\mathcal{H}| < \infty$ $\rightarrow$ PAC $m \approx \frac{\log \frac{|\mathcal{H}|}{\delta}}{\varepsilon}$
  - lec 11: Q3.3, $f \in \mathcal{H} = \{h_r\}$ $\rightarrow$ PAC $m \approx \log r_{\delta}/\varepsilon$
    non-realizable, $|\mathcal{H}| < \infty$ $\rightarrow$ Agnostic PAC $m \approx \log \frac{|\mathcal{H}|}{\delta}/\varepsilon^2$
- learner knows nothing $\rightarrow$ [No free lunch theorem]

NO