Learning Theory: Rademacher Complexity and Infinite Hypotheses.

Friday, March 24, 2017 8:31 AM

▷ Last time
- Defined PAC learning model
- Defined notion of learnability
- Showed Learnability for finite hypotheses via uniform convergence.

▷ <u>This time:</u>
- Infinite hypotheses
- Learnability via bounded Rademacher Complexity.

▷ We will still be looking at the ERM algorithm over a hypothesis space $\mathcal{H}$: given dataset S

$$h_S = \underset{h \in \mathcal{H}}{\operatorname{argmin}} \underbrace{\hat{\mathcal{L}}_S(h)}_{\text{loss on samples}} \equiv \frac{1}{m} \sum_{i=1}^{m} \ell(h, z_i)$$

▷ We want to measure performance w.r.t. true loss: $\mathbb{E}_{z \sim D}[\ell(h, z)] = \mathcal{L}_D(h)$

▷ Our goal is to show if m is large enough
(1) In expectation:

$$\mathbb{E}[\mathcal{L}_D(h)] \leq \inf \mathcal{L}_D(h) + \varepsilon$$

$$E_S[L_D(h_S)] \leq \inf_{h \in H} L_D(h) + \varepsilon$$

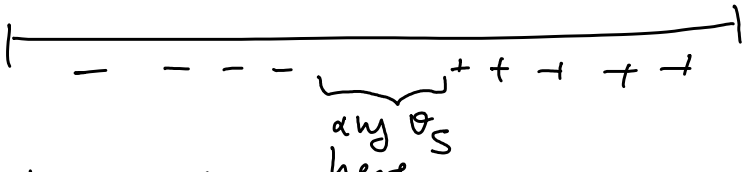
(2) with high probability $1-\delta$ over S

$$L_D(h_S) \leq \inf_{h \in H} L_D(h) + \varepsilon.$$

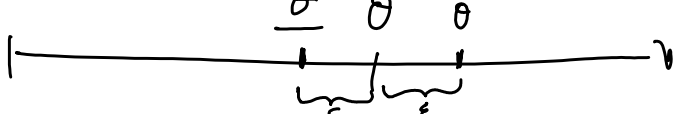
▷ Last time we show on example on 1-dimension

If H was all functions then we are in trouble.

- What if we restrict H to threshold functions. Then ERM picks the best threshold on sample.



- So the only misclassifications happen in the region between true θ and θ_S.
- But if you consider any region around θ



$$Pr(X \in [\underline{\theta}, \theta]) = \varepsilon = P(x \in [\theta, \bar{\theta}])$$

then θ_S has loss more than ε if it either falls below $\underline{\theta}$ or above $\bar{\theta}$. But it fall below $\underline{\theta}$ if there are no x's drawn in the region $[\underline{\theta}, \theta]$

Each of them falls in there w.p. ε.

So $\{\text{prob of no } x \text{ in } [\underline{\theta}, \theta]\} = (1-\varepsilon)^m \leq e^{-\varepsilon m}$

So if $m \geq \frac{\log(2/\delta)}{\varepsilon} \Rightarrow \{\text{above}\} \leq \frac{\delta}{2}$

similarly $\theta_S \in [\theta, \bar{\theta}]$ w.p. at most $\frac{\delta}{2}$.

- So sample complexity: $m(\varepsilon, \delta) \leq \frac{\log(2/\delta)}{\varepsilon}$! !

▷ We will present a general theory of when we can make such claims for infinite hypotheses!

▷ The reason for non-learnability in the unconstrained H was that there were hypotheses (the dark memoization one) that had a zero loss on sample but high loss on distribution, i.e.

$$\sup_h \left(L_D(h) - L_S(h) \right) \text{ was large!}$$

▷ Last time we saw that if we can bound $\sup_{h \in H} |L_D(h) - L_S(h)| \leq \varepsilon$ i.e. uniform convergence, then we get learning.

▷ Now we will see that the absolute value is not really needed, i.e. we only need to have that the loss on the distribution cannot be much larger than the loss on sample, uniformly over H.

(the other direction is not necessary. to be guaranteed uniformly)

Learnability via Representativeness

▷ The representativeness of a sample S is defined as:

$$Rep(S, D) = \sup_{h \in H} \left(L_D(h) - L_S(h) \right)$$

Lemma Representativeness implies good expected loss

$$\mathop{\mathbb{E}}_{S}[L_D(h_S)] \leq L_D(h^*) + \mathop{\mathbb{E}}_{S}[Rep(S, D)]$$

Pf

$$\mathbb{E}[L_D(h_S) - L_S(h_S)] \leq \mathbb{E}_S\left[\sup_h \left(L_D(h) - L_S(h)\right)\right]$$

$$\underset{S}{E}\left[L_D(h_S) - L_S(h_S) \right] \le \underset{S}{E}\left[\sup_h \left(L_D(h) - L_S(h) \right) \right]$$
$$= E_S[\text{Rep}(S,D)]$$

$$E_S[L_S(h^*)] = L_D(h^*) \quad \text{(because } h^* \text{ is ind. of } S\text{)}$$

$$\Rightarrow E_S[L_D(h_S)] \le E_S[L_S(h_S)] + E_S[\text{Rep}(S,D)]$$

$$\left(\begin{matrix}\text{since} \\ h_S \text{ max} \\ L_S\end{matrix}\right) \le E_S[L_S(h^*)] + E_S[\text{Rep}(S,D)]$$

$$= L_D(h^*) + E_S[\text{Rep}(S,D)]$$

It also implies an easy h.p. bound:

- By Markov's ineq.

$$P\left(L_D(h_S) - L_D(h^*) > \frac{E[L_D(h_S) - L_D(h^*)]}{\delta} \right) \le \delta$$

$\Rightarrow$ w.p. $1-\delta$

$$L_D(h_S) \le L_D(h^*) + \frac{E[\text{Rep}(S,D)]}{\delta}.$$

- Using McDiarmid's Inequality we can actually get: w.p. $1-\delta$

$$L_D(h_S) \le L_D(h^*) + E[\text{Rep}(S,D)] + 5\sqrt{\frac{\log(3/\delta)}{m}}$$

(Look at Chapter 26 of "Understanding Machine Learning")

Bounding Representativeness

- Intuitively what if I wanted to bound

representativeness by using sample data.

- Split into train set S_1, validation set S_2
- Use:

$$\sup_h \left(L_{S_1}(h) - L_{S_2}(h) \right)$$

as a proxy for representativeness.

- Split sample representativeness is exactly equal to: how much larger than the training error, can my test error be when I run ERM on train?

- Now if $|S_1| = |S_2| = \frac{m}{2}$ then:

$$\sup_h \left(L_{S_1}(h) - L_{S_2}(h) \right) = \frac{2}{m} \sup_h \sum_{t=1}^{m} \sigma_t \ell(h, z_t)$$

where $\sigma_t = \begin{cases} 1 & \text{if } t \in S_1 \\ -1 & \text{if } t \in S_2 \end{cases}$

- Then the Rademacher complexity is equal to the Expected train-test split error gap, over a random split, where each data point is in S_1 or S_2 w.p. $\frac{1}{2}$.

$$\mathcal{R}(H, S) = \frac{2}{m} \mathop{\mathbb{E}}_{\vec{\sigma}} \left[\sup_{h \in H} \sum_{t=1}^{m} \sigma_t \ell(h, z_t) \right]$$

- It captures: If I randomly split my data and run ERM on one set how good is the test error.

Thm

$$\mathop{\mathbb{E}}_{S}\left[Rep(D, S)\right] \leq \mathop{E}_{S}\left[R(H, S)\right]$$

Pf

- For every fixed h: true loss of h, $L_D(h)$, can be seen as expected loss of a newly sampled dataset S', i.e. $L_D(h) = \mathop{E}_{S'}\left[L_{S'}(h)\right]$
- So: $Rep(S, D) = \sup_{h \in H} \left(L_D(h) - L_S(h)\right)$

$$= \sup_{h \in H} \left(\mathop{E}_{S'}\left[L_{S'}(h)\right] - L_S(h)\right)$$

$$\leq \mathop{E}_{S'}\left[\sup_{h \in H} \left(L_{S'}(h) - L_S(h)\right)\right]$$

- So:

$$\mathop{E}\left[Re(S, D)\right] \leq \mathop{E}\left[\sup \left(L_{S'}(h) - L_S(h)\right)\right]$$

$$E_S[Re(S,D)] \le E_{S,S'}\Big[\underbrace{\sup_{h\in H}\left(L_{S'}(h) - L_S(h)\right)}_{\text{split sample representativeness on a sample of size } 2m.}\Big]$$

- Since: S, S' contain iid data I can swap any data point between them:

$$E_{S,S'}\left[\sup_{h\in H}\frac{1}{m}\sum_{t=1}^{T}\left(\ell(h,z_t') - \ell(h,z_t)\right)\right]$$

$$\| $$

$$E_{S,S'}\left[\sup_{h\in H}\frac{1}{m}\sum_{t}\sigma_t\left(\ell(h,z_t') - \ell(h,z_t)\right)\right]$$

For any $\sigma_t \in \{-1, 1\}$ (since this just corresponds to swapping z_t and z_t' in the samples. But z_t, z_t' are i.i.d.)

$\|$ if we draw $\sigma_t = \begin{cases} 1 & \text{w.p. } 1/2 \\ -1 & \text{w.p. } 1/2 \end{cases}$

$$E_{S,S',\vec{\sigma}}\left[\sup_{h}\frac{1}{m}\sum_{t}\sigma_t\left(\ell(h,z_t') - \ell(h,z_t)\right)\right]$$

$\le$

$$E_{S,S',\vec{\sigma}}\left[\sup_{h}\frac{1}{m}\sum_{t}\sigma_t\ell(h,z_t') + \sup_{h}\frac{1}{m}\sum_{t}(-\sigma_t)\ell(h,z_t)\right]$$

$$E_{S,S',\vec{\sigma}}\left[\sup_h \frac{1}{m}\sum_t \sigma_t \ell(h,z_t') + \sup_h \frac{1}{m}\sum_t (-\sigma_t)\ell(h,z_t)\right]$$

$$\| $$

$$E_{S',\vec{\sigma}}\left[\sup_h \frac{1}{m}\sum_t \sigma_t \ell(h,z_t')\right] + E_{S,\vec{\sigma}}\left[\sup_h \frac{1}{m}\sum_t (-\sigma_t)\ell(h,z_t)\right]$$

$$\| \quad \text{since } \sigma_t, -\sigma_t \text{ are identically dist and } S, S' \text{ too}$$

$$2\, E_{S,\vec{\sigma}}\left[\sup_h \frac{1}{m}\sum_t \sigma_t \ell(h,z_t)\right]$$

$$\|$$

$$E_S\left[R(H,S)\right]$$

Bounding the Rademacher complexity

▷ We want to bound for any data set of size m:

$$\frac{2}{m}\, E_{\vec{\sigma}}\left[\sup_{h\in H} \sum_{t=1}^{m} \sigma_t \ell(h,z_t)\right]$$

▷ Nice property of Rademacher:

- for any fixed h, variables:

$$Y_t = \sigma_t \ell(h, z_t) \quad \text{are mean zero.}$$

- So $\frac{1}{m}\sum_{t=1}^{m} \sigma_t \ell(h, z_t)$ is the sum of mean zero random vars.
- By Hoeffding's inequality we expect this to be of order $\sqrt{\frac{\log 1/\delta}{m}}$ w.p. $1-\delta$
- So then by a union bound we expect the $\sup_h \frac{1}{m}\sum_t \sigma_t \ell(h, z_t)$ to be of order

$$\sqrt{\frac{\log |H|/\delta}{m}} \quad \text{w.p. } 1-\delta. \Rightarrow \underset{\text{value}}{\text{Expected}} = O\left(\sqrt{\frac{\log |H|}{m}}\right)$$

- This works for finite $|H|$. What about infinite?
- Well we don't really need to take union bound over all H. Just over all possible distinct functions H_S that can occur on a sample S of size m.

Formally:

- Let H_S be all possible distinct hypotheses

on S, i.e. $\forall h \in H$, $\exists h' \in H_S$ s.t.

$$\forall z_t \in S : \ell(h', z_t) = \ell(h, z_t)$$

i.e. I cannot distinguish h, h' by just looking at their outputs on S.

Example:

So

If H is the space of threshold functions then how many distinct hypotheses can I create on m samples?

Answer: At most $(m+1)!$

So then by the above reasoning I should expect:

$$R(H, S) = O\left(\sqrt{\frac{\log(m+1)}{m}}\right)$$

Even if H is unbounded!

Growth Rate | The growth rate of a hypothesis space H is the largest size of H_S for any data set of size m, as a function of m:

$$\tau_H(m) = \sup_{S : |S| \le m} |H_S|$$

$\#$ $\quad S : |S| \le m$

Lemma For any H:

$$R(H, S) \le 2\sqrt{\frac{2\log(\tau_H(m))}{m}}.$$

Pf

$$R(H, S) = 2\, E_{\vec{\sigma}}\left[\sup_{h \in H} \frac{1}{m} \sum_{t=1}^{m} \sigma_t \ell(h, z_t)\right]$$

$$= 2\, E_{\vec{\sigma}}\left[\sup_{h \in H_S} \frac{1}{m} \sum_{t=1}^{m} \sigma_t \ell(h, z_t)\right]$$

(We use the following in-expectation analogue of the Hoeffding Analysis)

Massart's Lemma Let $H = \{h_1, \ldots, h_N\}$ be a finite hypothesis space. Then

$$E_{\vec{\sigma}}\left[\max_{h \in H} \frac{1}{m} \sum_{t=1}^{m} \sigma_t \ell(h, z_t)\right] \le \sqrt{\frac{2\log N}{m}}$$

Pf

$$E_{\vec{\sigma}}\left[\max_{h \in H} \frac{1}{m} \sum_{t=1}^{m} \sigma_t \ell(h, z_t)\right] =$$

$$\frac{1}{\lambda} E_{\vec{\sigma}}\left[\max_{h \in H} \log\left(\exp\left(\frac{\lambda}{m} \sum_{t} \sigma_t \ell(h, z_t)\right)\right)\right]$$

$$\frac{1}{\lambda} E_{\vec{\sigma}}\left[\log \max_{h \in H} \exp\left(\frac{\lambda}{m} \sum_t \sigma_t \ell(h, z_t)\right)\right]$$

$\wedge\vee$ Jensen's

$$\frac{1}{\lambda} \log E_{\vec{\sigma}}\left[\max_{h \in H} \exp\left(\frac{\lambda}{m} \sum_t \sigma_t \ell(h, z_t)\right)\right]$$

$\wedge\vee$

$$\frac{1}{\lambda} \log E_{\vec{\sigma}}\left[\sum_{h \in H} \exp\left(\frac{\lambda}{m} \sum_t \sigma_t \ell(h, z_t)\right)\right]$$

$\vee\vee$

$$\frac{1}{\lambda} \log \sum_{h \in H} E_{\vec{\sigma}}\left[\prod_t \exp\left(\frac{\lambda}{m} \sigma_t \ell(h, z_t)\right)\right]$$

$\vee\vee$

$$\frac{1}{\lambda} \log \sum_{h \in H} \prod_t E_{\sigma_t}\left[\exp\left(\frac{\lambda}{m} \sigma_t \ell(h, z_t)\right)\right]$$

$\wedge\vee$

$$\frac{1}{\lambda} \log \sum \prod^{m} \exp\left(-\frac{\lambda^2}{}\right)$$

$$\frac{1}{\lambda} \log \sum_{h \in H} \prod_{t=1}^{n} \exp\left(-\frac{\lambda^2}{2m^2}\right)$$

$$\frac{1}{\lambda} \log \sum_{h \in H} \exp\left(\frac{\lambda^2}{2m}\right)$$

$$\frac{1}{\lambda} \log N \exp\left(\frac{\lambda^2}{2m}\right)$$

$$\frac{1}{\lambda} \log N + \frac{\lambda}{2m} \overset{\lambda = \sqrt{\frac{2m}{\log N}}}{=\!=} 2\sqrt{\frac{\log N}{2m}}$$

$$\leq \sqrt{\frac{2 \log N}{m}}$$