Lecture 2

RECAP

Definition (PAC Learnability):

A hypothesis class $\mathcal{H}$ is PAC-learnable if $\exists$ a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, every distribution D over X and every $h \in \mathcal{H}$, when the algorithm is given $n_{\mathcal{H}}(\epsilon, \delta)$ samples drawn from D & labelled by h, the alg. produces a hypothesis $\hat{h}$ s.t. with probability $1-\delta$, $R(\hat{h}) \leq \epsilon$.

(The probability is over randomness in training set, and any internal algorithmic randomness.

TODAY:

* PAC bound for finite hypothesis classes
* Impossibility of PAC-learning classes which are too rich (No-free lunch theorem)
* Agnostic PAC learning
* Uniform convergence

Theorem (PAC bound for finite hypothesis class):

Let $\mathcal{H}$ be a hypothesis class with finite size $|\mathcal{H}|$. Then $\mathcal{H}$ is PAC-learnable with $n_{\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log(|\mathcal{H}|/\delta)}{\epsilon}\right)$ samples.

Proof: We will show that an alg. which finds ERM PAC-learns $\mathcal{H}$.

[Recall, $R(h) = \mathbb{E}_{(x,y)\sim D}\, \mathbb{1}(h(x) \neq y)$

$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(x) \neq y)$]

$(\hat{R}_S(h^*) = 0)$

First, due to **realizability**, $\exists\, h^* \in \mathcal{H}$ s.t. $R(h^*) = 0$.

Let $\mathcal{H}_{bad} = \{h \in \mathcal{H} \mid R(h) > \epsilon\}$ (bad hypothesis)

$S_{bad} = \{S \in (X \times Y)^n \mid \exists h \in \mathcal{H}_{bad}, \hat{R}_S(h) = 0\}$

(bad training sets)

Goal: Upper bound probability of getting training set from S_{bad}.

$$S_{bad} = \bigcup_{h \in \mathcal{H}_{bad}} \{ S \in (\mathcal{X} \times Y)^n \mid \hat{R}_S(h) = 0 \}$$

Now, $\mathbb{P}_S [S \in S_{bad}] = \mathbb{P}_{S \sim D} \left[\bigcup_{h \in \mathcal{H}_{bad}} \{ \hat{R}_S(h) = 0 \} \right]$

$$\leq \sum_{h \in \mathcal{H}_{bad}} \mathbb{P}_{S \sim D} [\hat{R}_S(h) = 0] \quad \text{(union bound)}$$

$$= \sum_{h \in \mathcal{H}_{bad}} \mathbb{P}_{S \sim D} [\{ \forall i \in \{1, \dots n\} \; h(x_i) = h^*(x_i) \}]$$ (realizability)

$$= \sum_{h \in \mathcal{H}_{bad}} \prod_{i=1}^{n} \mathbb{P}_{x_i \sim D} [h(x_i) = h^*(x_i)] \quad \text{(i.i.d.)}$$

$$\leq \sum_{h \in \mathcal{H}_{bad}} \prod_{i=1}^{n} (1 - \varepsilon) \leq |\mathcal{H}| (1 - \varepsilon)^n$$

$$\leq |\mathcal{H}| e^{-\varepsilon n}$$

(since $1 - x \leq e^{-x}$

)

$\therefore$ If $n \geq \left\lceil \frac{\log(|\mathcal{H}|/\delta)}{\varepsilon} \right\rceil$, prob. of failure is at most δ.

■

Bias - Complexity Tradeoff

$$R(h_{ERM}) - R(h^*) = \underbrace{R(h_{ERM}) - \min_{h \in \mathcal{H}} R(h)}_{\text{Estimation error}} + \underbrace{\min_{h \in \mathcal{H}} R(h) - R(h^*)}_{\text{Approximation error}}$$

Can we always make approximation error small?
Can we learn arbitrarily rich classes?

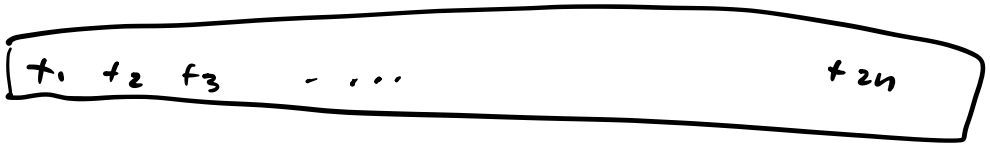
Thm: (No-free-lunch theorem):

Let A be a learning alg. for binary classification over $\mathcal{X}$ & let $n \leq |\mathcal{X}|/2$.
Then there exists a dist. D over $\mathcal{X} \times \{0,1\}$ s.t.

1) $R(h^*) = 0$.
2) w.p. atleast $1/7$ over training set S of size n, we have $R_D(A(S)) \geq 1/8$.

Proof:

set C of size $2n$, subset of domain, $C \in X^{2n}$



Consider all possible labelling functions

f_1	0	0	0	–	–	–	–	0
f_2	0	0	0	–	–	–	–	1
⋮								
f_T	1	1	1	–	–	–	–	1

$$T = 2^{2n}$$

For each $f_i \rightarrow$ dist. D_i

D_i : uniform over C, labels given by f_i

D_i^n : dist. of n i.i.d. samples from D_i

$$R_{D_i}(f_i) = 0$$

Goal: Show that

$$\max_{i \in [T]} \mathbb{E}_{S \sim D_i^n} \left[R_{D_i}(A(S)) \right] \geq 1/4 \quad \text{— (1)}$$

$([T] = \{1, \ldots, T\})$

Exercise: Show that (1) implies $\exists\, i \in [T]$ s.t. w.p. atleast $1/7$ over training set $S \sim D_i^n$ we have $R_{D_i}(A(S)) \geq 1/8$.

Idea: Take random variable $R_{D_i}(A(S)) = Z$

If for any r.v. Z if $\mathbb{E}(Z) \geq \ldots$

$\Rightarrow Z \geq \ldots$ w.p. $\ldots$

Let $S_1, \ldots, S_K$ be set of all possible training examples $(K = (2n)^n)$

f_i ⊕ S_1 S_2 $\ldots$ S_K

$\rightarrow$ S_1^i S_2^i $\ldots$ S_K^i

$$\max_{i \in [T]} \mathbb{E}_{S_j}\left[R_{D_i}(A(S_j^i))\right]$$

$$\geq \frac{1}{T}\sum_{i=1}^{T} \mathbb{E}_{S_j}\left(R_{D_i}(A(S_j^i))\right)$$

$$= \mathbb{E}_{S_j}\frac{1}{T}\sum_{i=1}^{T}\left(R_{D_i}(A(S_j^i))\right)$$

$$\geq \min_{S_j \in e^n} \frac{1}{T}\sum_{i=1}^{T} R_{D_i}(A(S_j^i))$$

Fix any S_j of size n.

$\Rightarrow$ $p \geq n$ samples $v_1, \ldots, v_p \in \ell$ that do not appear in S_j.

$$R_{D_i}(h) = \frac{1}{2n} \sum_{x \in C} \mathbb{1}(h(x) \neq f_i(x))$$

$$\geq \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}(h(v_r) \neq f_i(v_r)).$$

Thus,

$$\frac{1}{T} \sum_{i=1}^{T} R_{D_i}(A(S_j^i))$$

$$\geq \frac{1}{T} \sum_{i=1}^{T} \frac{1}{2p} \sum_{r=1}^{p} \mathbb{1}(A(S_j^i)(v_r) \neq f_i(v_r))$$

$$\geq \frac{1}{2} \min_{r \in [p]} \left(\frac{1}{T} \sum_{i=1}^{T} \mathbb{1}\{A(S_j^i)(v_r) \neq f_i(v_r)\} \right)$$

Partition f_i into $T/2$ pairs, each pair $(f_i, f_{i'})$ agrees on everything except v_r. Every pair $(f_i, f_{i'})$ produces same labelled dataset $S_j^i, S_j^{i'}$.

$$\mathbb{1}(A(S_j^i)(v_r) \neq f_i(v_r)) + \mathbb{1}(A(S_j^{i'})(v_r) \neq f_{i'}(v_r)) = 1$$

$$\frac{1}{T}\sum_{i=1}^{T} R_{D_i}(A(S_j^i)) \geq \frac{1}{2}\cdot\frac{1}{T}\cdot\frac{T}{2}$$

$$\geq \frac{1}{4}.$$

∎

corollary: Let X be an infinite domain set ($\mathbb{R}^d$) & let $\mathcal{H}$ be all possible functions from $X \to \{0,1\}$. Then $\mathcal{H}$ is not PAC-learnable.

Agnostic PAC learning

PAC learning
=> Needs realizability
=> Cannot handle noise in labels

Agnostic PAC-learning:

Arbitary dist. D over $X \times Y$

There could be noise, i.e. $D(y|x) = \begin{cases} 1 & \text{w.p. } \eta(x) \\ 0 & \text{w.p. } 1-\eta(x) \end{cases}$

Definitions are preserved.

$$R(h) = \mathbb{E}_{(x,y)\sim D} \ell(h(x), y) = \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

$$\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(h(x_i) = y_i)$$

Digression (Bayes-Optimal Predictor)

$$\min_{h: X \to Y} \mathbb{P}_{(x,y)\sim D}[h(x) \neq y]$$

$$= \sum_{x \in X} D(x) \left(\min_{h(x) \in \{0,1\}} \begin{cases} \mathbb{P}_D(Y=1 \mid x=x)\, \mathbb{1}(h(x)=0) \\ \mathbb{P}_D(Y=0 \mid x=x)\, \mathbb{1}(h(x)=1) \end{cases} \right.$$

$$= \sum_{x \in \mathcal{X}} D(x) \min \begin{cases} Pr_D(Y=1 \mid X=x) \\ Pr_D(Y=0 \mid X=x) \end{cases}$$

The predictor (Bayes-optimal predictor) $h^*(x)$

$$h^*(x) = \mathbb{1}\left(\eta(x) \geq 1/2\right)$$

$$\eta(x) = Pr_D(Y=1 \mid X=x)$$

Bayes-optimal risk

$$R(h^*) = \mathbb{E}_{x \sim D(x)} \left[\min \{ \eta(x), 1-\eta(x) \} \right]$$