CS492(D) Computational Learning Theory. 08 / 29 / 2023

§1. Objectives.

- Study basic ML algorithms.
- Study math techniques for analysing those algorithms.

§2 Preview.

$\mathcal{X}$ ··· input space.

$\mathcal{Y}$ ··· output space.

Unknown distribution $D \sim Pr(\mathcal{X} \times \mathcal{Y})$

① Given loss function L, $R(h) = \mathbb{E}_{(x,y)\sim D}[L(h(x), y)]$

e.g. quadratic loss

↳ Generalization error (risk)

Population risk (?)

We want (*) $\operatorname{argmin}_{h \in \mathcal{H}} R(h)$, but we don't know D.

$\mathcal{H}$: $[\mathcal{X} \to \mathcal{Y}]$

② Then, do what?

Sample from D is available.

- $S = ((x_1, y_1), \cdots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$
- Proxy objective of $R(h)$, using S.

  Then solve optimization problem with this objective.
- E.g. $\hat{R}_S(h) := \frac{1}{m}\sum_{i=1}^{m}(h(x_i) - y_i)^2 = \mathbb{E}_{(x,y)\sim \mathrm{Unif}(S)}[(h(x) - y)^2]$

- Empirical risk optimization (**) $\operatorname{argmin}_{h \in \mathcal{H}} \hat{R}_S(h)$

08 / 29 / 2023

What is relationship btwn (*) & (**)?

Result for regression:

Theorem Under some conditions, $\exists c_1, c_2, c_3$ s.t.

$\forall \delta > 0$, $\mathbb{P}[\forall h \in \mathcal{H} \quad R(h) \le \hat{R}_S(h) + c_1\sqrt{\frac{\log c_2 m}{m}} + c_3\sqrt{\frac{\log(1/\delta)}{m}}] \ge 1-\delta$.

③ How do we solve (**) quickly?

- Design a good hyps. set $\mathcal{H}$.
  - $\mathcal{H}_1 = \{f \in [\underbrace{\mathbb{R}^I}_{\mathcal{X}} \to \mathbb{R}] \mid f(x) = \langle w, \Phi(x)\rangle \text{ for } w \in \mathbb{R}^F\}$ → Traditional regression
    $\Phi : \mathbb{R}^I \to \mathbb{R}^F$. ↕ Strong duality.
  - $\mathcal{H}_2 = \{f \in [\mathcal{X} \to \mathbb{R}] \mid f(x) = \langle \alpha, (y_i K(x_i, x))_{i\in[m]}\rangle, \ \alpha \in \mathbb{R}^m,$ → Kernel (Nonparametric approach).
    $K : \mathbb{R}^I \times \mathbb{R}^I \to \mathbb{R}\}$.

§3 Logistics

* Ch2-6 of the textbook.

*
- Final 40%
- Group Project 40%
  - (i) 2-4 people
  - (ii) Pick a paper COLT 2011-2023
  - (iii) Report ≤ 4 pages
  - (iv) 35-min talk with slides (only 4 teams present with extra +5 marks).
- HW (2-3 Problem sheet). 20%.

Ch2 PAC learning framework

08 / 31 / 2023

§1. Classification Problem

$\mathcal{X}$: Input Space ($\mathbb{R}^n$ most of the time)

$\mathcal{Y}$: Output Space ($\{0,1\}$)

$c \in [\mathcal{X} \to \mathcal{Y}]$

$D \in Pr(\mathcal{X})$, $S = ((x_1, y_1), \cdots, (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^m$.

$H \subseteq [\mathcal{X} \to \mathcal{Y}]$ $\quad x_i \overset{iid}{\sim} D$, $y_i = c(x_i)$.

Goal: Find a good $h \in H$ s.t. $h \approx c$ using $\mathcal{X}, \mathcal{Y}, S$

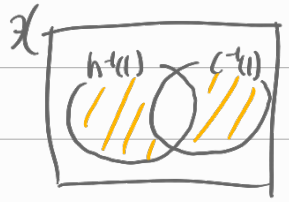
§2 Generalization Error, Empirical Error.

$R(h)$ $\quad$ $\hat{R}_S(h)$

Generalization Err

(i) For $h \in H$, $R(h) := \underset{x \sim D}{\mathbb{E}} [\mathbb{1}_{\{h(x) \neq c(x)\}}] = \underset{x \sim D}{\mathbb{P}}(h(x) \neq c(x))$

Can understand h, c as subset of input : $h^{-1}(1), c^{-1}(1)$



Empirical Err

(ii) $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} [\mathbb{1}_{h(x_i) \neq c(x_i)}]$

$= \underset{x_i \sim Unif(S)}{\mathbb{E}} [h(x_i) \neq c(x_i)]$

(iii) Prove $\underset{S \sim D^m}{\mathbb{E}} [\hat{R}_S(h)] = R(h)$.

By linearity of expectation, it suffices to show $\underset{S \sim D^m}{\mathbb{E}} [\mathbb{1}_{h(x_1) \neq c(x_1)}] = R(h)$

(*) =

$(*) = \mathop{\mathbb{E}}_{(x_i, y_i) \sim D} [\mathbb{1}_{h(x_i) \neq c(x_i)}] = R(h)$ by definition.

§3 PAC Learnability

PAC: Probably Approximately Correct

① Setup

$\mathcal{X}, \mathcal{Y}, \mathcal{H}, \mathcal{C} \subseteq [\underbrace{\mathbb{R}^n}_{\text{"}\mathcal{X}} \to \mathcal{Y}]$

- $\mathcal{C}$ is PAC-learnable by $\mathcal{H}$ if $\exists$ an algorithm $\mathcal{A}$ and a polynomial $p(\,,\,,\,)$ such that

$\forall c \in \mathcal{C}, \forall D \in Pr(\mathcal{X}), \forall \varepsilon, \delta > 0, \forall m \in \mathbb{N}$

$m \geq p(\frac{1}{\varepsilon}, \frac{1}{\delta}, n) \Rightarrow \mathop{\mathbb{P}}_{S \sim D^m} [R(h_{\mathcal{A},S}) \leq \varepsilon] \geq 1 - \delta$

(in polynomial scale)

i.e. If we see enough samples (m), with high probability, the algorithm produce good answer.

- $\mathcal{C}$ is efficiently PAC-learnable if $\exists$ polynomial-complexity algorithm $\mathcal{A}$ that makes $\mathcal{C}$ PAC-learnable.

Remarks

- This should work for all distribution D
- If $\mathcal{H} \not\supseteq \mathcal{C}$, then $\mathcal{C}$ is PAC is not PAC-learnable.

For any algorithm $\mathcal{A}$,

Take $c \notin \mathcal{H}$. $\forall\ h \in \mathcal{H}$, we have $c \neq h$.

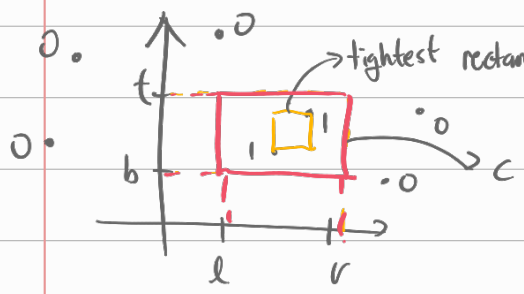
Then, $\exists x_i \in \mathcal{X}$ s.t. $c(x_i) \neq h(x_i)$.

Let

08 / 31 / 2023

§4. Examples - Learning Axis Aligned Rectangles

① Setup $\mathcal{X}=\mathbb{R}^2$, $\mathcal{Y}=\{0,1\}$, $\mathcal{H}=\mathcal{C}=\{(a,b)\mapsto \mathbb{1}_{[a,b)\in[l,r]\times[b,t]} \mid l\le r, b\le t\}$



② Polynomial p $P(\frac{1}{\varepsilon},\frac{1}{\delta})=\frac{4}{\varepsilon}\log\frac{4}{\delta}$

③ Why PAC Let $c\in\mathcal{C}$, $D\in P_1(\mathcal{X})$, $\varepsilon,\delta>0$ be given, $m\ge\frac{4}{\varepsilon}\log\frac{4}{\delta}$.

Goal: $\mathbb{P}_{S\sim D^m}[R(h_S)\le\varepsilon]\ge 1-\delta$.

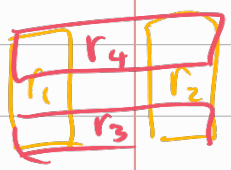
$\Leftrightarrow \mathbb{P}_{S\sim D^m}[R(h_S)>\varepsilon]<\delta$.

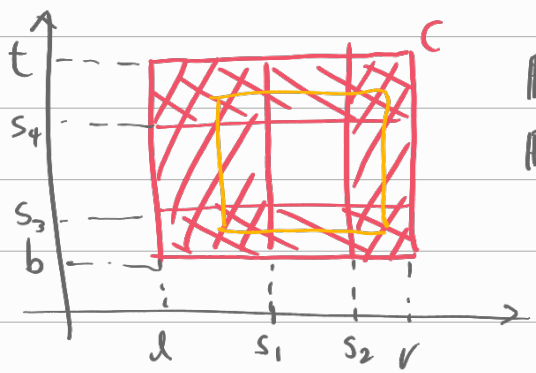
(pf) (Case 1) $\mathbb{P}_{S\sim D^m}[x\in c]\le\varepsilon$. As $h\in C$ (In sense $h^{-1}(1)\subseteq h^{-1}(c)$), ← By algorithm.

$h\triangle c\ (=(h\setminus c)\cup(c\setminus h)=(h^{-1}(1)\setminus c^{-1}(1))\cup(c^{-1}(1)\setminus h^{-1}(1)))$ has $P(h\triangle c)=R(h)<\varepsilon$.

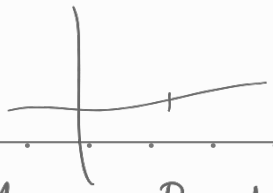
(Case 2) $\mathbb{P}_{S\sim D^m}[x\in c]>\varepsilon$.

$c\neq\emptyset$.





Take S_1 s.t.
$\mathbb{P}(x\in[l,s_1]\times[b,t])\ge\frac{\varepsilon}{4}$
$\mathbb{P}(x\in[l,s_1)\times[b,t])\le\frac{\varepsilon}{4}$.



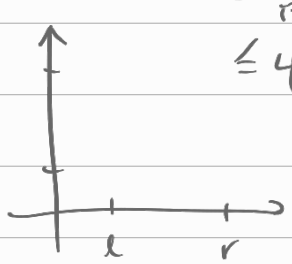
Find S_2, S_3, S_4 similarly. → Boundary of h is included in $///$
If h lives in $///$, then ($\Leftrightarrow h_S \cap r_i \neq \phi \ \forall i \in [4]$)

$$R(h) \leq \Pr(///) \leq 4 \cdot (\varepsilon/4) = \varepsilon$$

↳ Union bound.

Hence, it suffices to choose $\mathcal{A}$ produce h_S in $///$ with prob $\geq 1-\delta$.

$$\begin{aligned} \mathbb{P}(R(h_S) > \varepsilon) &\leq \mathbb{P}(\exists i \in [4] \text{ s.t. } h_S \cap r_i = \phi) \\ &\leq \sum_{i=1}^{4} \mathbb{P}(h_S \cap r_i = \phi) \leq \sum_{i=1}^{4} (1-\varepsilon/4)^m \\ &\leq 4e^{-\varepsilon m/4} = 4e^{-\log \frac{4}{\delta}} = \delta \end{aligned}$$



Theorem

Let $\mathcal{H} \subseteq [\mathcal{X} \to \mathcal{Y}]$ be finite with $\mathcal{C} \subseteq \mathcal{H}$.
Let $\mathcal{A}$ be an algorithm s.t. $\forall c \in \mathcal{C}$ and $S = ((x_1, y_1), \cdots, (x_m, y_m))$ for c, $\hat{R}_S(h_S) = 0$.

Then, $\forall \varepsilon, \delta > 0$, $\mathbb{P}_{S \sim D^m}[R(h_S) \leq \varepsilon] \geq 1-\delta$ if $m \geq \frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$.

Example

① Conjunction of Boolean literals.

$\mathcal{X} = \{0,1\}^n$. $\mathcal{C} = \mathcal{H} = \{x_1 \wedge x_2, x_2 \wedge \overline{x_{20}}, \cdots\} \cup \{\text{false}\}$.

$|\mathcal{H}| = 3^n \Rightarrow \log|\mathcal{H}| = n\log 3 \Rightarrow$ PAC learnable.

Efficiently PAC learnable. ($O(n)$ algorithm).

mn

② Universal Concept Class

$\mathcal{C} = \mathcal{H} = [\mathcal{X} \to \mathcal{Y}]$. Cannot classify using thm.

If $\mathcal{C}$ is PAC learnable, every $\mathcal{C}' \subseteq \mathcal{C}$ is also PAC learnable.

$|\mathcal{H}| = |\mathcal{Y}^{\mathcal{X}}|$. $\log|\mathcal{H}| = |\mathcal{X}|\log|\mathcal{Y}|$.

③ k-term DNF formulas

k-disjunction of conjugation $(\quad) \vee (\quad) \vee (\quad) \vee (\quad)$ (s, k)

$|\mathcal{H}| = (3^n)^k = 3^{nk}$ $(|\mathcal{H}| = \sum_{i=1}^{k} 3^{ni})$

$\log|\mathcal{H}| = nk \log 3 \Rightarrow$ PAC learnable using thm.

Not eff. pac learnable unless RP=NP

④ k-term CNF formulas.

$(\quad) \wedge \underbrace{(\quad)}_{k} \wedge (\quad) \cdots$

CNF $\Leftrightarrow$ DNF ?

2... $\Rightarrow (2n+1)^k$ variables

$|\mathcal{H}| = 3^{\#\text{ of var.}} \Rightarrow$ PAC learnable.

Efficiently PAC learnable.

(pf of theorem) Goal: $\mathbb{P}_{S\sim D^m}[R(h_S) > \varepsilon] \le \delta$.

(*) $\mathbb{P}_{S\sim D^m}[\exists h \text{ s.t. } \hat{R}_S(h) = 0 \wedge R(h) > \varepsilon]$

$\le \mathbb{P}_S\left[\bigvee_{\substack{h\in\mathcal{H} \\ R(h)>\varepsilon}} \hat{R}_S(h) = 0\right]$

$\le \sum_{\substack{h\in\mathcal{H} \\ R(h)>\varepsilon}} \mathbb{P}_{S\sim D^m}[R_S(h) = 0]$

$\le \sum_{\substack{h\in\mathcal{H} \\ R(h)>\varepsilon}} (1-\varepsilon)^m \le |\mathcal{H}|(1-\varepsilon)^m \le |\mathcal{H}| e^{-m\varepsilon}$

Thus, for $m \ge \frac{1}{\varepsilon}(\log|\mathcal{H}| + \log\frac{1}{\delta})$,

(*) $\le |\mathcal{H}| e^{-(\log|\mathcal{H}| + \log\frac{1}{\delta})} = \delta$.

/ /

Definition [Agnostic PAC Learnable]

" if

$$m \geq p(\tfrac{1}{\varepsilon}, \tfrac{1}{\delta}, n) \Rightarrow \mathbb{P}_{S\sim D^m}\left[R(h_S) \leq \inf_{h\in H} R(h)+\varepsilon\right] \geq 1-\delta.$$

Theorem A : algorithm s.t. $\forall C\in\mathcal{C}$ and $\forall S = ((x_1,y_1),\cdots,(x_m,y_m))\in C$,

$$\hat{R}_S(h_S) = \min_{h\in H} \hat{R}_S(h). \quad (\in ERM)$$

$\Rightarrow \forall C\in\mathcal{C}\ \forall D\in Pr(X)\ \forall \varepsilon,\delta>0,\ \forall m\in\mathbb{N}$,

if $m \geq \frac{2}{\varepsilon^2}\left(\log|H| + \log\frac{2}{\delta}\right)$, then $\mathbb{P}_{S\sim D^m}[R(h_S) \leq \min_{h\in H} R(h)+\varepsilon] \geq 1-\delta$

Lemma (c.f. proposition 4.1)

Denote h_S above as h^{ERM}.

$$\mathbb{P}[\exists_{h\in H}\ |R(h)-\hat{R}_S(h)| > \tfrac{\varepsilon}{2}] \quad \text{(=)}$$

$$\mathbb{P}_{S\sim D^m}[R(h^{ERM}) \geq \min_{h\in H} R(h)+\varepsilon] \leq \mathbb{P}_{S\sim D^m}\left[\max_{h\in H} |R(h)-\hat{R}_S(h)| > \tfrac{\varepsilon}{2}\right]$$

Thm 2.13' $m \geq \frac{1}{2\varepsilon^2}\left(\log|H| + \log\frac{2}{\delta}\right) \Rightarrow \mathbb{P}_{S\sim D^m}[\forall_{h\in H}\ |R(h)-\hat{R}_S(h)| \leq \varepsilon] \geq 1-\delta$

$\Updownarrow$

$\mathbb{P}[\exists_{h\in H}\ |R(h)-\hat{R}_S(h)| > \varepsilon] \leq \delta$

(pf) $m \geq \frac{1}{2\varepsilon^2}(\quad) \Rightarrow \mathbb{P}[\exists h\in H \text{ s.t. } |R(h)-\hat{R}_S(h)| > \tfrac{\varepsilon}{2}] \leq \delta$

$$\Rightarrow \mathbb{P}[R(h^{ERM}) \geq \min_{h\in H} R(h)+\varepsilon] \leq \delta. \blacksquare$$

Sep / 07 / 2023

Definition (weakly PAC learnable)

≡ Agnostic PAC learnable.

(pf of proposition 4.1) Let $h_0 \in \arg\min_{h \in \mathcal{H}} R(h)$.

Then, $R(h_S^{ERM}) - \min_{h\in\mathcal{H}} R(h) = R(h_S^{ERM}) - R(h_0)$

$\leq R(h_S^{ERM}) - R(h_0)$

$\leq R(h_S^{ERM}) - \hat{R}(h_S^{ERM}) + \hat{R}(h_S^{ERM}) - \hat{R}(h_0) + \hat{R}(h_0) - R(h_0)$

$\leq 2 \max_{h\in\mathcal{H}} |R(h) - \hat{R}(h)| + \hat{R}(h_S^{ERM}) - \hat{R}(h_0)$

$\hat{R}(h_S^{ERM}) - \hat{R}(h_0) \leq 0$ by def of h_S^{ERM}

Thus, $\mathbb{P}[R(h_S^{ERM}) - \min_{h\in\mathcal{H}} R(h) > \varepsilon] \leq \mathbb{P}[\max_{h\in\mathcal{H}} |R(h) - \hat{R}_S(h)| > \frac{\varepsilon}{2}]$ ∎

(pf of theorem 2.13(b)) Hoeffding's Inequality (theorem D.2)

If $X_1, \dots, X_m$ independent R.V. s.t. $X_i \in [a_i, b_i]$. $\forall i$.

For, $A := \frac{1}{m}\sum_{i=1}^m X_i$, $\varepsilon > 0$,

$$\mathbb{P}[A - \mathbb{E}[A] \geq \varepsilon] \leq e^{-\frac{2m^2\varepsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

$$\mathbb{P}[A - \mathbb{E}[A] \leq -\varepsilon] \leq e^{-\frac{2m^2\varepsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}$$

$$\Rightarrow \mathbb{P}[|A - \mathbb{E}[A]| \geq \varepsilon] \leq 2e^{-\frac{2m^2\varepsilon^2}{\sum_{i=1}^m (b_i - a_i)^2}}.$$

$$\mathbb{E}_{S\sim D^m}[\hat{R}_S(h)] = R(h).$$

$$\hat{R}_S(h) = \frac{1}{m}\sum_{i=1}^m \underbrace{\mathbb{1}_{h(x_i)\neq c(x_i)}}_{X_i} \qquad X_i \in [0,1] \text{ is a RV.}$$

Thus by Hoeffding, $\mathbb{P}[|\hat{R}_S(h) - R(h)| \geq \varepsilon] \leq 2e^{-\frac{2m\varepsilon^2}{\sum_{i=1}^m 1^2}} = 2e^{-2m\varepsilon^2}$.

For $m \geq \frac{1}{2\varepsilon^2}\log(|\mathcal{H}| + \log\frac{2}{\delta})$, $\leq \delta$. under union bd on $h \in \mathcal{H}$?

$$\mathbb{P}[\max |R(h) - \hat{R}_S(h)| > \varepsilon]$$
$$= \mathbb{P}[\exists h \in \mathcal{H} \text{ s.t. } |R(h) - \hat{R}_S(h)| \geq \varepsilon]$$
$$\leq \sum_{h \in \mathcal{H}} \mathbb{P}[|R(h) - \hat{R}_S(h)| \geq \varepsilon]$$
$$= |\mathcal{H}| \, 2e^{-2m\varepsilon^2} \leq \delta.$$

§5 Stochastic Scenarios

- $x \sim D$ $\quad y = c(x)$ $\quad \cdots$ stochastic scenario
- $D \in Pr(\mathcal{X} \times \mathcal{Y})$ $\quad (x, y) \sim D$. → y also given by probability.
- $S = ((x_1, y_1), \cdots, (x_m, y_m)) \sim D^m$.
- $\hat{R}_S(h) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{h_i(x_i) \neq y_i\}}$
- $R(h) = \mathbb{E}[\hat{R}_S(h)] = \mathbb{P}_{(x,y) \sim D}[h(x) \neq y]$.

Definition (Agnostic PAC-learning algo.) Stochastic Case

$\exists$ a poly p s.t. $\forall D \in Pr(\mathcal{X} \times \mathcal{Y})$ $\forall \varepsilon, \delta > 0$ $\forall m \in \mathbb{N}$

$m \geq p(\frac{1}{\varepsilon}, \frac{1}{\delta}, n) \Rightarrow \mathbb{P}_{S \sim D^m}[R(h_S) \leq \inf_{h \in \mathcal{H}} R(h) + \varepsilon] \geq 1 - \delta$.

§5.2 Bayes Classifier

- $D \in Pr(\mathcal{X} \times \mathcal{Y})$
- $h \in [\mathcal{X} \to \mathcal{Y}]$, $h \in \operatorname{argmin}_{h_0 \in [\mathcal{X} \to \mathcal{Y}]} R(h_0)$

$$h_{BC}(x) = \operatorname*{argmax}_{i \in \{0,1\}} (\mathbb{P}[y = i \mid x]).$$

Proposition $R(h_{BC}) \leq R(h)$ $\forall h \in \mathcal{H}$.

→ conditional prob.

$$\mathbb{E}_{(x,y) \sim D}[\mathbb{1}_{\{h(x) \neq y\}}] = \mathbb{E}_x[\mathbb{E}_y[\mathbb{1}_{\{h(x) \neq y\}} \mid x] \mid x]$$
$$= \mathbb{E}_x[\mathbb{E}_y[\mathbb{1}_{h(x)=0}\mathbb{1}_{y=1} + \mathbb{1}_{h(x)=1}\mathbb{1}_{y=0} \mid x]]$$
$$= \mathbb{E}_x[\mathbb{1}_{h(x)=0}\mathbb{E}_y[\mathbb{1}_{y=1} \mid x] + \mathbb{1}_{h(x)=1}\mathbb{E}_y[\mathbb{1}_{y=0} \mid x]]$$

need completion

$$\geq \mathbb{E}_x[\min(\mathbb{P}[y=0|x], \mathbb{P}[y=1|x])$$
$$= 1 - \mathbb{E}_x[\max(\mathbb{P}\quad, \mathbb{P}\quad))$$
$$= 1 - \mathbb{E}_x[h_{BC}(x) = y]$$

① Stochastic ⇔ deterministic.
② $\mathcal{H}$ infinite PAC learnable - theorems?
③ $C \subseteq \mathcal{H}$ necessary. why.

- Consistent / ERM 아닌 모드도 PAC-learnability 이야기 할 수 있는가?
- Non PAC-learnable 예시?
- Theorem 2.13 이 strictly strong statement 인가.
  ① for h_S vs $\forall h$. ② Bound $\frac{2}{\epsilon^2}(\log|\mathcal{H}| + \log\frac{2}{\delta})$ vs $\frac{1}{2\epsilon^2}($ " $)$
  ③ A need not be ERM.