Lecture 4 - theory basics

Tuesday, January 26, 2021 14:36

today: theory basics

theory basics

decision theory setup

estimate $h_w : \mathcal{X} \to \mathcal{Y}$

generalization error = $L_P(w) \triangleq \mathbb{E}_{(x,y)\sim P}[\ell(y, h_w(x))]$ (task loss)

ultimate goal is to find $w^* = \underset{w \in \mathcal{W}}{\operatorname{argmin}}\, L_P(w)$

problem: do not know P ("true" distribution on (x,y))

suppose $(x^{(i)}, y^{(i)})_{i=1}^n \overset{\text{i.i.d.}}{\sim} P$ $\rightsquigarrow$ we could look at

$\triangleq D_n$ training dataset

$$\hat{L}_n(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(y^{(i)}, h_w(x^{(i)}))$$

from statistics / prob. theory.

$\hat{L}_n(w) \xrightarrow[n\to\infty]{\text{a.s.}} L_P(w)$ (LLN) for each fixed w (point wise)

this is weaker than $\sup_w |\hat{L}_n(w) - L_P(w)| \xrightarrow[n\to\infty]{\text{a.s.}} 0$

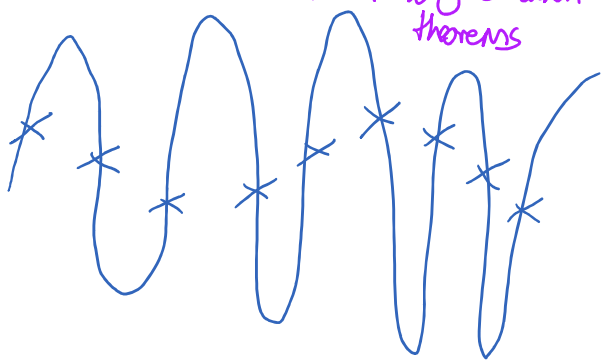
Note: minimizing the training error gives no guarantee on $L_P(\hat{w}_n)$ in general!

$\rightsquigarrow$ later no free lunch theorems

e.g. polynomial regression

for n points, can get zero training error with poly. of degree $n-1$

$\Rightarrow$ "overfitting"



in learning theory: $\rightarrow$ study properties of learning alg. $\hat{w}_n$

in particular, what can we say about $L_p(A(D_n))$

different approaches:

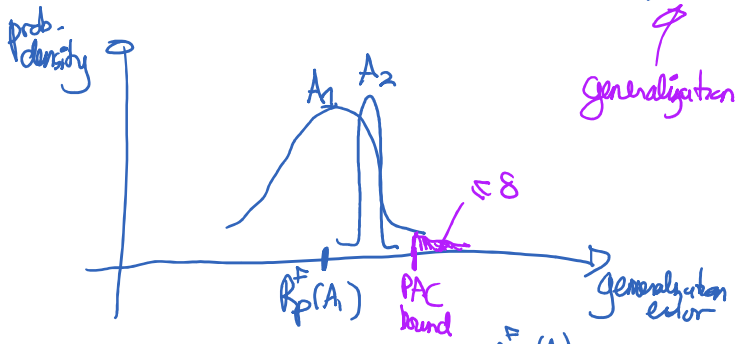
a) "frequentist risk" $R^F_{p,n}(A) \triangleq \mathbb{E}_{D_n \sim p^{\otimes n}}[L_p(A(D_n))]$ (D_n is random)

b) PAC framework "probably approximately correct"

$P\{L_p(A(D_n)) > \text{some bound}\} \leq \delta$ ("bad bound")

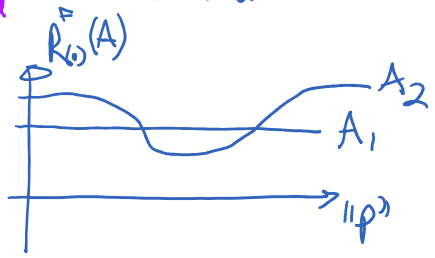
ie. $L_p(A(D_n)) \leq$ " " with prob. $\geq 1-\delta$

↑ generalization error bound



issue with R^F_p → depends on p

"risk profiles":



weighted frequentist risk $\mathbb{E}_{\theta \sim \pi(\theta)}[R^F_{p_\theta}(A)]$

c) "Bayesian posterior risk"

$R^{post}(w|D_n) \triangleq \mathbb{E}_{\theta \sim p(\theta|D_n)}[L_{p_\theta}(w)]$

- prior $p(\theta)$ over dist.
- observation model $p(D_n|\theta)$

$\Rightarrow$ posterior $p(\theta|D_n)$

Bayesian estimate $\hat{w}_n^{Bayes} = \operatorname{argmin}_w R^{post}(w|D_n)$

$A^{Bayesian}$ is optimal for weighted frequentist risk using $\pi(\theta) = p(\theta)$

15h25

No free lunch?

frequentist risk analysis learning alg. A

let $\mathcal{P}$ be a set of distributions on $\mathcal{X} \times \mathcal{Y}$

Sample complexity of A with respect to $\mathcal{P}$

is the smallest $n(\mathcal{P}, A, \varepsilon)$ s.t. $\forall n \geq n(\mathcal{P}, A, \varepsilon)$

we have $\sup_{P \in \mathcal{P}} [R_P^E(A;n) - L_P(h_P^*)] < \varepsilon$

"uniform result"

$h_P^* = \operatorname{argmin}_{h: \mathcal{X} \to \mathcal{Y}} L_P(h)$

terminology: • A is consistent for dist. P

if $\lim_{n \to \infty} R_P^E(A;n) - L_P(h_P) = 0$

• A is uniformly consistent for a family $\mathcal{P}$

if $\lim_{n \to \infty} \left[\sup_{P \in \mathcal{P}} [R_P^E(A;n) - L_P(h_P)] \right] = 0$

Binary classification $\mathcal{Y} = \{-1, +1\}$

I) if $\mathcal{X}$ is finite; then the "voting procedure" (assign the most frequent label to an input x)

is uniformly and universally consistent

↳ i.e. $\mathcal{P}$ is all distributions $\mathcal{X} \times \mathcal{Y}$

with (universal) sample complexity

$n(\mathcal{P}, \varepsilon, A_{voting}) \leq \frac{|\mathcal{X}|}{\varepsilon^2}$ (free lunch?!)

II) if $\mathcal{X}$ is infinite

no free lunch theorem (for binary with l the 0-1 loss)

for any n and any learning alg. A

then $\sup_{P \text{ all dist.}} [R_P^E(A;n) - L_P(h_P^*)] \geq \frac{1}{2}$

i.e. $\exists$ always a dist P_A s.t. your A is worse than random prediction for P_A

NEXT:

for p_A

NFL II:

[thm. 7.2 in Devroye & al. 1996]

$\varepsilon_1 \leq \frac{1}{16}$

let ε_n be any non-increasing seq. converging to 0 (could be arbitrarily slowly)

for any A,

then $\exists P_A$ s.t. $[R^F_{P_A}(A;n) - L_{P_A}(h^*_{P_A})] \geq \varepsilon_n \;\forall n$ e.g. $\frac{1}{\lg(\lg(\lg(\dots(n))))}$ ← 10^{23}

(*) (*) Consequence: we need assumptions on $\mathcal{P}$ to say anything useful

Occam's generalization error bound

- binary class. & 0-1 loss
- consider $\mathcal{W}$ to be a countable set

let's define a prior over $\mathcal{W}$: $\pi(w)$ i.e. $\sum_{w \in \mathcal{W}} \pi(w) = 1$ $\pi(w) \geq 0 \;\forall w$

$|w|_\pi$ = "description length" of $w \triangleq \log_2 \frac{1}{\pi(w)}$) ⇒ $\sum_w 2^{-|w|_\pi} \leq 1$ "Kraft's inequality"

Occam's bound

for any fixed P; with prob. $\geq 1-\delta$ over training set $D_n \sim P^{\otimes n}$

$$\forall w \in \mathcal{W} \quad L_P(w) \leq \hat{L}_P(w) + \frac{1}{\sqrt{2n}} \Omega_\pi(w; \delta)$$

where $\Omega_\pi(w;\delta) \triangleq \sqrt{(\ln 2)|w|_\pi + \ln\frac{1}{\delta}}$ (complexity measure)

* bound is useful only for dist P s.t. $|w^*_P|_\pi$ is small

↳ $\arg\min_{w \in \mathcal{W}} L_P(w)$

$|w|_\pi = \log_2 \frac{1}{\pi(w)}$)

if $\pi(w) \propto \exp(-\|w\|^2)$

then $|w|_\pi = \|w\|^2 + cst.$

note: 0-1 loss ass. appears in const. of Chernoff bound

proof: use 3 things

1) Chernoff bound $\quad P\{D_n : \hat{L}_n(w) \leq L(w) - \varepsilon\} \leq \exp(-2n\varepsilon^2) \;\forall \varepsilon > 0$

proof: use 3 things

1) Chernoff bound (concentration inequality) — appears on cost. of Chernoff bound

$$P\{P_n : \hat{L}_n(w) \le L(w) - \varepsilon\} \le \exp(-2n\varepsilon^2) \quad \forall \varepsilon > 0$$

2) union bound

$$P\{\exists x \text{ s.t. } prop(x) \text{ is true}\} \le \sum_x P\{prop(x) \text{ is true}\}$$

3) "Kraft's inequ."

$$\sum_w 2^{-|w|_\pi} \le 1$$

we say that w is naughty ("bad") if bound fails $\quad L - \varepsilon > \hat{L}_n$

$$bad(w) = \mathbb{1}\{L(w) > \hat{L}_n(w) + \underbrace{\frac{1}{\sqrt{2n}}\Omega_\pi(w,\delta)}_{\varepsilon_n(w)}\}$$

using Chernoff, $\hat{L}_n(w) \le L(w) - \varepsilon_n(w)$ with small prob.

$$P\{bad(w)\} \le \exp(-2n\varepsilon_n(w)^2) = \exp\left(-2n\frac{1}{2n}((\ln 2)|w|_\pi + \ln\tfrac{1}{\delta})\right)$$

$$= \delta 2^{-|w|_\pi}$$

using union bound

$$P\{\exists w : bad(w)\} \le \sum_w P\{bad(w)\} \le \sum_w \delta 2^{-|w|_\pi} \overset{\text{Kraft}}{\le} \delta$$ //

Surrogate loss NP hard to minimize $\hat{L}_n(w)$; replace with $\hat{g}_n(w)$ which is "surrogate"

eg. • hinge loss
• log-loss

next: countable → uncountable
"PAC-Bayes"