Lecture 4 - theory basics

Thursday, January 18, 2024 9:29 AM

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 $w^* = \operatorname{argmin}_{w \in \mathcal{W}} L_P(w)$

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

suppose $\underbrace{(x^{(i)}, y^{(i)})_{i=1}^n}_{\triangleq D_n \text{ training dataset}}$ i.i.d. P ⇝ we could look at

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

from statistics / prob. theory

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

(LLN)

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 in general !?

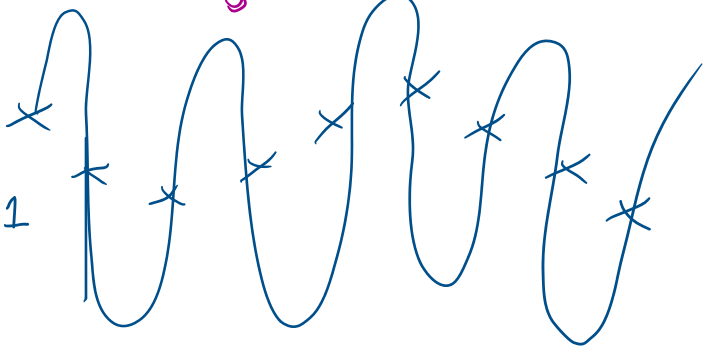
$L_P(\hat{w}_n)$

⇝ later: no free lunch theorems

e.g. polynomial regression

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

⇒ "overfitting"



in learning theory: study properties of learning alg. $A : \mathcal{D}_n \to \mathcal{W}$

in particular, what can we say about $L_P(\underbrace{A(D_n)}_{\hat{w}})$

in particular, what can we say about $L_p(\underbrace{A(D_n)}_{\hat{\omega}_n})$

different approaches:

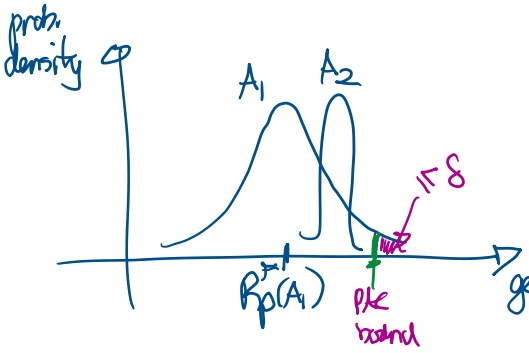
a) "frequentist risk" $R^F_{p,n}(A) \triangleq \mathbb{E}_{D_n \sim p^{\otimes n}}[L_p(A(D_n))]$ (D_n is random)

b) PAC framework "probably approximately correct"

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

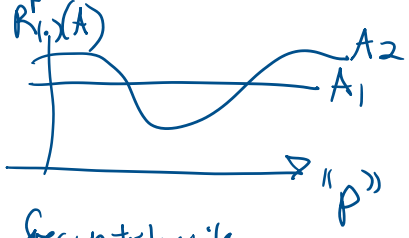
i.e. $L_p(A(D_n)) \leq$ ‖ ‖ with prob $\geq 1-\delta$

(‖ ‖ = 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"

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

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

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

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

No free lunch!

frequentist risk analysis learning alg. A

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

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 that $\sup_{p \in \mathcal{P}} [R^F_p(A; n) - L_p(h^*_p)] \leq \varepsilon$

we have that $\sup_{p \in \mathcal{P}} [R_p^F(A;n) - L_p(h_p^*)] \leq \varepsilon$

"uniform result"

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

terminology:

- A is consistent for fixed dist. p if $\lim_{n \to \infty} R_p^F(A;n) - L_p(h_p^*) = 0$
- A is uniformly consistent for a family $\mathcal{P}$ if $\lim_{n \to \infty} \left[\sup_{p \in \mathcal{P}} [R_p^F(A;n) - L_p(h_p^*)] \right] = 0$

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 on $\mathcal{X} \times \mathcal{Y}$

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

10h28

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

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

for any n and any learning alg. A

then $\sup_{p \text{ all dist.}} [R_p^F(A;n) - L_p(h_p^*)] \geq \frac{1}{2}$

i.e. $\exists$ always a dist $p_{A,n}$ s.t. your A is worse than random prediction for $p_{A,n}$!

NFL II: [thm. 7.2 in Devroye et al. 1996] — $\varepsilon_1 \leq \frac{1}{16}$

Let ε_n be any non-increasing seq. converging to 0 (could be arbitrarily slow)

for any A, then $\exists p_A$ s.t. $[R_{p_A}^F(A;n) - L_p(h_p^*)] \geq \varepsilon_n \ \forall n$

e.g. $\frac{1}{\lg(\lg(\lg(\ldots(n))))}$ (10^{23})

(⊗⊗) Consequence: we need assumptions on P to say anything useful

Occam's generalization error bound

- binary class & 0-1 loss
- consider W to be a countable set

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

$|w|_\pi$ = "description length" of $w \triangleq \log_2 \frac{1}{\pi(w)}$

$\Rightarrow \sum_w 2^{-|w|_\pi} \leq 1$ "Kraft's inequality"

Occam's bound

for any fixed p: with prob $\geq 1-\delta$ over training $D_n \sim p^{\otimes n}$

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

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

(*) minimizing bound $\rightarrow$ universally consistent alg.

(*) bound is useful only for dist. p s.t. $|w^*_p|_\pi$ is small
($\hookrightarrow$ $\operatorname{argmin}_{w \in W} L_p(w)$)

note: 0-1 loss assumption appears in constant of bound

proof: use 3 things

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

2) union bound $P\{\exists x \text{ s.t. } prop(x) \text{ is true}\} \leq \sum_{x \in X} P\{prop(x) \text{ is true}\}$ (X countable)

3) "Kraft's ineq." $\sum_w 2^{-|w|_\pi} \leq 1$

we say that w is "naughty" if bound fails ("bad")

$L - \varepsilon > \hat{L}_n$

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

using Chernoff, $\hat{L}_n(w) \leq L(w) - \varepsilon$ with small prob

$P\{bad(w) = 1\} \leq \exp(-2n\varepsilon_n(w)^2) = \exp(-2n / (\ldots$

using Chernoff, $L_n(w) \nearrow L(w) - \varepsilon$ with small prob

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

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

using union bound

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

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

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

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

e.g. • hinge loss
• log-loss

next: countable → uncountable
"PAC-Bayes"