Lecture 4 - theory basics

Thursday, January 16, 2020 13:28

today: theory basics

theory basics

decision theory setup

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

task loss

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

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 $\underbrace{(x^{(i)}, y^{(i)})_{i=1}^n}_{\triangleq D_n \text{ training data}} \overset{\text{i.i.d.}}{\sim} P$ → 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 fixed w (pt. wise)

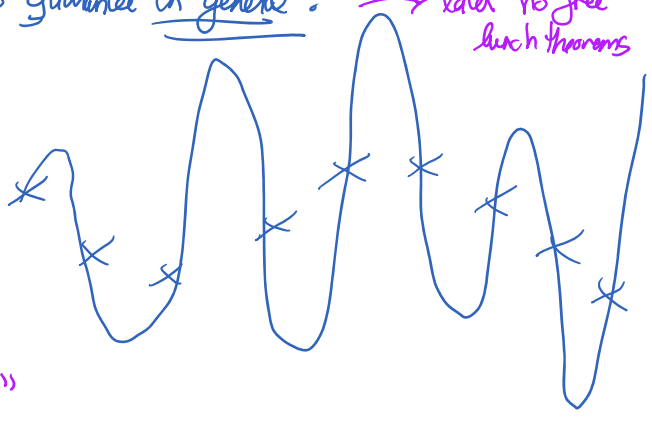
(LLN)

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

note: minimizing the training error gives no guarantee in general? → later no free lunch theorems

e.g. polynomial regression

for n points, can get zero training error with polynomial of deg. $n-1$



⇒ "overfitting"

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

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

different approaches

D_n is random

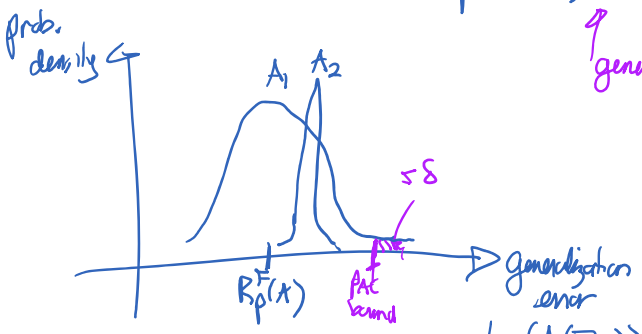
a) "frequentist risk" $R^F_{p,n}(A) \triangleq \mathbb{E}_{D_n \sim p^n}[L_p(A(D_n))]$

b) PAC framework ("tail bound")
"probably approximately correct"

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

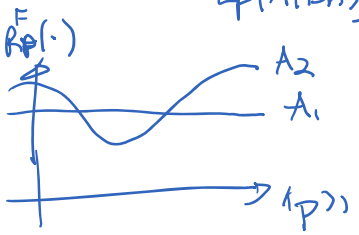
ie. $L_p(A(D_n)) \leq$ " " with prob. $\geq 1-\delta$

↑ 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 def.
- observation model $p(D_n|\theta)$

⇒ posterior $p(\theta|D_n)$

Bayesian estimate $\hat{w}_n^{Bayes} = \underset{w}{\text{argmin}}\, R^{post}(w|D_n)$

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

14h20

No free lunch?

frequentist risk analysis learning algorithm 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 $\sup_{P \in \mathcal{P}} [R_P^F(A;n) - L_P(h_P^*)] < \varepsilon$

"uniform result"

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

terminology:

- A is consistent for 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 $Y = \{-1, +1\}$

I) If 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 $X \times Y$

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

II) if X is infinite

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

for any n and any learning alg. A

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

i.e. ∃ always a dist. P s.t. your alg A is worse than random prediction (?)

NFLT 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 slowly e.g. $\frac{1}{\log(\log(\log(\ldots \log(n) \ldots)))}$ with 10^{23} logs)

then ∃ P s.t. $R_P^F(A;n) - L_P(h_P^*) \geq \varepsilon_n \ \forall n$

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

Occam's generalization error bound

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

Let's define a prior prob 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 set $D_n \sim p^{\otimes n}$

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

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

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

$|w|_\pi = \log_2 \frac{1}{\pi(w)}$ if $\pi(w) \propto \exp(-\|w\|^2)$ then $|w|_\pi \sim \|w\|^2 + \text{cst.}$

note: 0-1 loss ass. appears in constants of Chernoff 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 $P\{\exists x \text{ s.t. } prop(x) \text{ is true}\} \leq \sum_x P\{prop(x) \text{ is true}\}$
3) "Kraft's inequality" $\sum_w 2^{-|w|_\pi} \leq 1$

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

we say w is naughty if bound fails

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

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

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

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

using union bound

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

//

Surrogate loss NP hard to minimize $\hat{L}_n(w)$; replace with $\hat{\ell}_n(w)$ which is "surrogate"
eg. • hinge loss
• log-loss