Lecture 4 - scribbles

Thursday, January 17, 2019 13:25

today: theory basics

theory basics

decision theory setup

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

task loss

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

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

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

learning alg. : $\hat{w}_n = A(D_n)$ (A = algorithm)

from statistics / prob. theory

$$\hat{L}_n(w) \overset{\text{a.s.}}{\underset{n\to\infty}{\to}} L_P(w) \quad \forall w$$

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

$[\ X_n \overset{\text{a.s.}}{\to} X$ ...

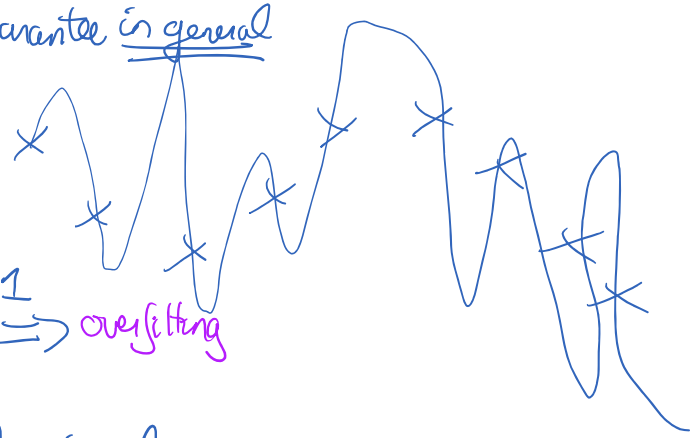
$$P\{\omega \in \Omega \text{ s.t. } X_n(\omega) \not\to X(\omega)\} = 0$$

note: minimizing training error gives no guarantee in general

e.g. polynomial regression

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

$\Rightarrow$ overfitting



in learning theory, study properties of learning alg

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

different approaches:

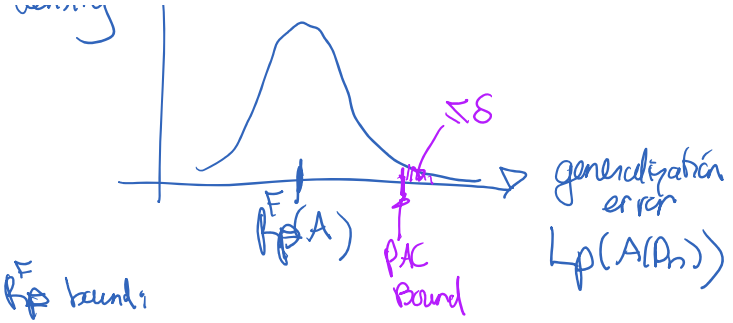
a) "frequentist risk" $R^F_{p,n}(A) \triangleq \mathbb{E}_{D_n \sim p^n}[L_p(A(D_n))]$ (D_n is random)

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

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

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

↳ generalization error bound

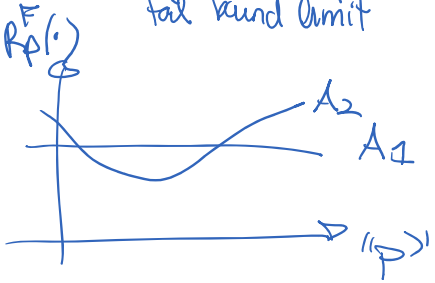


from PAC bound to R_p^F bound:

$$\mathbb{E}X = \mathbb{E}X\,\mathbb{1}\{X \le B\} + \mathbb{E}X\,\mathbb{1}\{X > B\}$$

$\le B \cdot (1-\delta)$ ← tail bound limit; $\le B \cdot \delta$

issue with R_p^F → depends on P



weighted frequentist risk

$$\mathbb{E}_{\theta \sim p(\theta)}\left[R_{p_\theta}^F(A)\right]$$

c) "Bayesian posterior risk"

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

- prior $p(\theta)$ over distributions
- observation model $p(D_n \mid \theta)$

posterior ⇒ $p(\theta \mid D_n)$

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

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

14h20

No free lunch

frequentist risk analysis of learning algorithm A

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

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 $\sup_{P \in \mathcal{P}} [R^E_P(A;n) - L_P(h^*_P)] < \varepsilon$

"uniform result"

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

terminology:

- A is consistent for dist. P

  if $\lim_{n \to \infty} R^E_P(A;n) - L_P(h^*_P) = 0$

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

  if $\lim_{n \to \infty} \left[\sup_{P \in \mathcal{P}} [R^E_P(A;n) - L_P(h^*_P)] \right] = 0$

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 $\mathcal{X}$ is infinite

no free lunch thm
(for binary class with 0-1 loss)

for any n and any learning alg. A

then $\sup_{P \text{ all distn}} [R^F_P(A;n) - L_P(h^*_P)] \geq \frac{1}{2}$

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

NFLT II:

[thm. 7.2 in Devroye et al. 1996]

Let $\varepsilon_n \leq \frac{1}{16}$ be any non-increasing sequence converging to zero (could be arbitrarily slowly e.g. $\frac{1}{\log(\log(\ldots(n))}$ 10^{23})

then $\exists P$ s.t. $R^F_P(A;n) - L_P(h^*_P) \geq \varepsilon_n \ \forall n$

consequence: we need assumptions on P to say anything

Occam's generalization error bound

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

Let's define a prior prob. 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

$\triangleq \log_2 \frac{1}{\pi(w)}$

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

(*) bound is only useful for distribution P s.t. $|w_P^*|_\pi$ is small

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

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

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

note: 0-1 loss assumption appears in constants of Chernoff bound

proof: use 3 things

1) Chernoff bound. (concentration inequality)

$$P\{D_n : \hat{L}_n(w) \leq L(w) - \varepsilon\} \leq \exp(-2n\varepsilon^2) \quad \forall \varepsilon \geq 0$$

2) union bound

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

3) "Kraft's inequality"

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

"bad" ... $\Rightarrow L - \varepsilon > \hat{L}_n$

$\sum_w$

we say w is "bad" naughty if bound fails

$$\text{bad}(w) = \mathbb{1}\{ L(w) > \hat{L}_n(w) + \underbrace{\frac{1}{\sqrt{2n}} \Omega_q(w;\delta)}_{\varepsilon_n(w)} \} \quad \rightarrow \quad L - \varepsilon > \hat{L}_n$$

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

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

$$= \delta 2^{-|w|_q}$$

using union bound

$$P\{\exists w : \text{bad}(w)\} \leq \sum_w P\{\text{bad}(w)\} \leq \sum_w \delta 2^{-|w|_q} \leq \delta$$

Surrogate loss :

We need to minimize $\hat{L}_n(w)$; replace with $\hat{\mathcal{L}}_n(w)$ which is "surrogate"

eg. hinge loss
· log-loss
...