Lecture 5 - PAC-Bayes

Tuesday, January 31, 2023 2:28 PM

today: ~~finish~~ Occam's bound
- PAC Bayes
- probit loss

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

example of learning alg: $\hat{w}_n = \operatorname{argmin}_{w \in W}$ RHS of bound $\Rightarrow$ universally consistent!

$|w|_\pi = \log_2 \frac{1}{\pi(w)}$ if $\pi(w) \propto \exp(-\|w\|^2)$ → RHS is ℓ_2-regularized ERM

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

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

proof: use 3 things:

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

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

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

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

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

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

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

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

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

using union bound: (Kraft's ineq.)

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

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"

PAC-Bayes

Occam's bound → we linked $\hat{L}_n(w)$ with $L_p(w)$

uniformly over all $w \in W$ ← but countable

using complexity $|w|_\pi$ (for "prior")

PAC-Bayes: generalize this to
- arbitrary W
- general $\ell(y,y') \in [0,1]$

caveat: switch to a randomized predictor

i.e. instead of learning $\hat{w}$, predicting $y = h_{\hat{w}}(x)$

consider $\hat{q}$ distribution over W

predict: first $w \sim \hat{q}(w)$; $y = h_w(x)$

⇒ use $\mathbb{E}_{\hat{q}}[L(w)]$ as the "generalization error" for $\hat{q}$

i.e. $\mathbb{E}_{(x,y)\sim p}\, \mathbb{E}_{w\sim\hat{q}}[\ell(y, h_w(x))]$

idea: use empirical version

$\mathbb{E}_q[\hat{L}_n(w)]$ ← optimize over q

→ on structured prediction will yield probit surrogate loss. (see soon)

PAC-Bayes thm. [McAllester 1999, 2003]

PAC-Bayes thm. [McAllester 1999, 2003]

(let $\ell(y,y') \in [0,1]$) ; for any fixed prior π over W
and any dist. p over $X \times Y$
then with $\geq 1-\delta$ over $D_n \sim p^{\otimes n}$
it holds that $\forall q$ dist. over W

$$\mathbb{E}_q[L_p(w)] \leq \mathbb{E}_q[\hat{L}_n(w)] + \frac{1}{\sqrt{2(n-1)}} \sqrt{KL(q\|\pi) + \ln \frac{n}{\delta}}$$

new complexity term

note: if W is countable ; let $q_{w_0} = \mathbb{1}\{w = w_0\}$
then $KL(q\|\pi) = \sum_w q(w) \log \frac{q(w)}{\pi(w)} = \log \frac{1}{\pi(w_0)} = (\log 2)\,|w_0|_\pi$

15h33

probit loss for structured prediction [NeurIPS 2011 McAllester & Keshet]

if $q_w(w') \triangleq N(w' \mid w, I)$

$w' = w + \varepsilon$ where $\varepsilon \sim N(0,I)$

then $\mathbb{E}_{q_w}[L(w')] = \mathbb{E}_{w' \sim q_w} \mathbb{E}_{(x,y)\sim p}[\ell(y, h_{w'}(x))]$

$$= \mathbb{E}_{(x,y)\sim p}\left[\mathbb{E}_{\varepsilon \sim N(0,I)}[\ell(y, h_{w+\varepsilon}(x))]\right]$$

$$\ell_{probit}(x,y;w)$$

why called probit?

binary classification : $y = \{-1,+1\}$ with 0-1 loss
$h_w(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle)$ let $\alpha = y\langle w, \varphi(x)\rangle$ (margin)

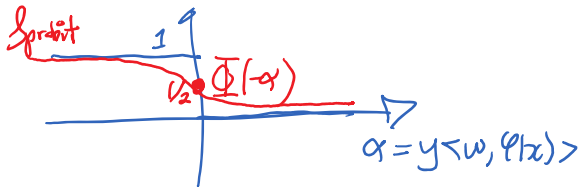
then $\ell_{probit}(x,y;w) = \mathbb{E}_{\varepsilon \sim N(0,I)} \mathbb{1}\{y \neq h_{w+\varepsilon}(x)\}$
↳ 1 when
i.e. $y\langle w+\varepsilon, \varphi(x)\rangle < 0$

$$\underbrace{y\langle w, \varphi(x)\rangle}_{\alpha} < -y\langle \varepsilon, \varphi(x)\rangle$$

1 when $-\alpha > \underbrace{\langle \varepsilon y, \varphi(x)\rangle}_{N(0, \|\varphi(x)\|^2)}$ (assume $\|\varphi(x)\| = 1$)

$$\ell_{probit} = \Pr\{\varepsilon_1 < -\alpha\} = \Phi(-\alpha)$$

↑ cdf of standard Gaussian



(*) define $\hat{w}_n^{(probit)} = \underset{w \in W}{\operatorname{argmin}} \; \hat{\ell}_{probit}(w) + \frac{\lambda_n}{2n}\|w\|^2$ (*)

McAllester showed the consistency of $\hat{w}_n^{(probit)}$

McAllester 2011 uses Catoni's PAC-Bayes' version

$$[\forall q, \; \mathbb{E}_q[L(w)] \le \left(\frac{1}{1-\frac{1}{2\lambda_n}}\right)\left[\mathbb{E}_q[\hat{L}_n(w)] + \frac{\lambda_n}{n}\left[KL(q\|\pi) + \ln\frac{1}{\delta}\right]\right]$$

if we use $\pi = N(0, I)$, $Q_w = N(w, I)$ } $\hat{\ell}_{probit}(w) + \frac{\lambda_n}{n}\frac{\|w\|^2}{2}$

Motivates $\hat{w}_n^{(probit)}$ (*)

thm 1 in paper Let $\lambda_n \nearrow \infty$ slowly enough so that $\frac{\lambda_n \ln n}{n} \to 0$

then $\ell_{probit}(\hat{w}_n) \xrightarrow[n\to\infty]{a.s.} L^* = \min_{w \in W} L_p(w)$

McAllester call this "consistency"

but true consistency would be $L(\hat{w}_n) \xrightarrow{a.s.} L^*$

[Lacoste-Julien unpublished result : if $L(w)$ is cts.]

then $\ell_{probit}(\hat{w}_n) \xrightarrow{a.s.} L^*$

$\Rightarrow L(\hat{w}_n) \xrightarrow{a.s.} L^* = L(w^*)$

proof idea : use Catoni's PAC Bayes bound

with prob $\ge 1-\delta_n$ $\ell_{probit}(\hat{w}_n) \le \left(\frac{1}{1-\frac{1}{2\lambda_n}}\right)\left[\hat{\ell}_{probit}(\hat{w}_n) + \frac{\lambda_n}{2n}\|\hat{w}_n\|^2 + \ln\frac{1}{\delta_n}\right]$

$\le$ (by def. of $\hat{w}_n$) $\ell_{probit}(\alpha w^*) + \frac{\lambda_n}{2n}\alpha^2\|w^*\|^2$

set $\delta_n = \frac{1}{n^2}$

$\leq \ell_{probit}(\alpha w^*) + \sqrt{\frac{\alpha n}{n}}$ using Chernoff bound for αw^*

with prob $\geq 1 - \frac{1}{n^2}$

get $\lim_{n \to \infty} \ell_{probit}(\hat{w}_n) \leq \ell_{probit}(\alpha w^*)$ with prob 1

* also show that $\lim_{\alpha \to 0} \ell_{probit}(\alpha w^*) \leq L(w^*)$ [see paper for details]

problem: $\ell_{probit}(x, y; w)$ is non-convex in w $\Rightarrow$ no optimization guarantees

now: convex surrogates ℓ