Lecture 5 - PAC-Bayes

Thursday, January 28, 2021 13:33

today: • PAC-Bayes
• probit loss
• review surrogate loss

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 $|w|_\pi$ ← "prior"

PAC-Bayes: generalize this to

- arbitrary W
- general $\ell(y, y') \in [0,1]$

caveat: switch to a randomized predictor

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

↓ empirical version

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

↝ or structured prediction will yield probit surrogate loss (see soon)

PAC-Bayes theorem [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 prob. $\geq 1-\delta$ over $D_n \sim p^{\otimes n}$

it holds that $\forall q$ dist. over W $\quad \mathbb{E}_q[L_p(w)] \leq \mathbb{E}_q[\hat{L}_n(w)]$

it holds that $\forall q$ dist. over W $\quad \mathbb{E}_q[L_p(w)] \le \mathbb{E}_q[\hat{L}_n(w)] + \frac{1}{\sqrt{2(n-1)}}\sqrt{KL(q\|\pi) + \ln\frac{n}{\delta}}$

note: if W is countable; let $q_{w_0} = \mathbb{1}\{w = w_0\}$ — new complexity term

then $KL(q\|\pi) = \sum_w q(w) \log\frac{q(w)}{\pi(w)} = \log\frac{1}{\pi(w_0)} = (\ln 2)\,|w_0|_{bit}$

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

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

then $\mathbb{E}_{q_w}[L(w')] = \mathbb{E}_{w' \sim q_w}\mathbb{E}_{(x,y)\sim P}[\ell(y, h_{w'}(x))]$ where $w' = w + \varepsilon$, $\varepsilon \sim N(0, I)$

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

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

why name probit?

binary classi.: $\mathcal{Y} = \{-1, +1\}$ with 0-1 loss

$h_w(x) = \mathrm{sgn}(\langle w, \varphi(x)\rangle)$

let margin $\alpha = y\langle w, \varphi(x)\rangle$

then $\ell_{probit}(x, y; w) = \mathbb{E}_{\varepsilon\sim N(0,I)}\mathbb{1}\{y \ne h_{w+\varepsilon}(x)\}$

$y\langle w+\varepsilon, \varphi(x)\rangle < 0$

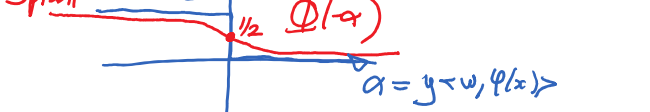
$\underbrace{y\langle w, \varphi(x)\rangle}_{\alpha} < -y\langle \varepsilon, \varphi(x)\rangle$

$-\alpha > \langle \varepsilon y, \varphi(x)\rangle$

same prob. as $-\alpha > \langle \varepsilon, \varphi(x)\rangle$

(suppose $\|\varphi(x)\| = 1$) $\hookrightarrow P\{\varepsilon_1 < -\alpha\}$

$\ell_{probit} = P\{\varepsilon_1 < -\alpha\} = \Phi(-\alpha)$ — cdf of standard Gaussian



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

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

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

14h23

McAllester 2011 uses Catoni's PAC-Bayes version

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

if we use $\pi = N(0, I)$, $q_w = N(w, I)$ } $\mathbb{E}_q[\hat{L}_n(w)] \to \hat{\ell}_{\text{probit}}(w)$, $\frac{\lambda_n}{n} KL(q\|\pi) \to \frac{\lambda_n}{n} \frac{\|w\|^2}{2}$

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

Thm. 7

in paper: Let $\lambda_n \to \infty$ slowly enough so that $\frac{\lambda_n}{n/\ln n} \to 0$

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

McAllester calls 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 $\hat{\ell}_{\text{probit}}(\hat{w}_n) \to L^*$
$\Rightarrow L(\hat{w}_n) \xrightarrow{a.s.} L^* = L(w^*)$]

proof idea: use Catoni's PAC Bayes bound

with prob. $\geq 1 - \delta_n$ $\hat{\ell}_{\text{probit}}(\hat{w}_n) \leq \left(\frac{1}{1 - \frac{1}{2\lambda_n}}\right) \left(\hat{\ell}_{\text{probit}}(\hat{w}_n) + \frac{\lambda_n}{2n} \|\hat{w}_n\|^2 + \ln \frac{1}{\delta_n} \right)$ by def. of $\hat{w}_n$

$\leq \hat{\ell}_{\text{probit}}(\alpha w^*) + \frac{\lambda_n}{2n} \alpha^2 \|w^*\|^2$ pick α

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

* also use $\lim_{\alpha \to \infty} \ell_{\text{probit}}(\alpha w^*) \leq L(w^*)$

$\lim_{n\to\infty} \ell_{\text{predict}}(\hat{w}_n) = L(w^*)$ [see paper for details]

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

how: convex surrogates $s(\tilde{y}) \triangleq s(x, \tilde{y}; w)$ i.e. x & w is implicit

Review of convex surrogates mentioned so far:

$\ell_{\text{perceptron}}(x, y; w) = \max_{\tilde{y}\in\mathcal{Y}} s(\tilde{y}) - s(y)$ let $m(\tilde{y}) = s(y) - s(\tilde{y})$

$= \max_{\tilde{y}\in\mathcal{Y}} [-m(\tilde{y})] = [\max_{\tilde{y}\neq y} -m(\tilde{y})]_+$

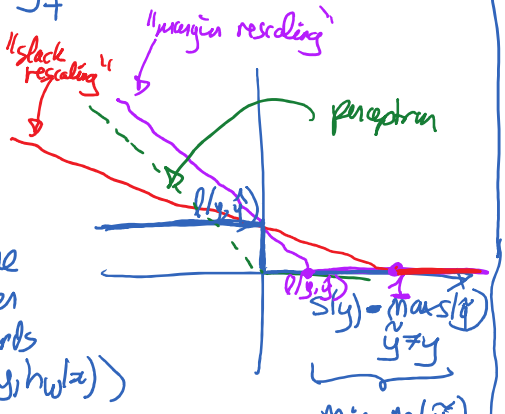
$\ell_{\text{hinge}}(\ )$ (structured SVM) $= \max_{\tilde{y}} [s(\tilde{y}) + \ell(y,\tilde{y})] - s(y)$

"margin rescaling" $\rightarrow$ $= \max_{\tilde{y}} [\ell(y,\tilde{y}) - m(\tilde{y})]$

vs.

"slack rescaling" $= \max_{\tilde{y}} \ell(y,\tilde{y}) [1 - m(\tilde{y})]$

} are upper bounds $\ell(y, h_w(x))$



$\ell_{\text{CRF}}(\ ) = \frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))\right) - s(y)$ $[-\log p_w(y|x)]$

$\beta \to \infty \Rightarrow$ perceptron loss

$\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(-\beta m(\tilde{y}))\right)$ suggests "smoothed hinge loss" $\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta[\ell(y,\tilde{y}) - m(\tilde{y})])\right)$

[e.g. Pletscher & al. 2010]

note: slack rescaling more robust when have small $\ell(y,\hat{y})$ [e.g. 0]

but more computionally costly

what theoretical properties could we look at?

a) generalization error bounds [next class]

b) consistency properties & calibration fct. [next^2 class]
↳ relationship between $L(w)$ & $\ell(w)$

why structural score functions?

$$s(x,y) = \sum_{c \in \mathcal{C}} s_c(x, y_c)$$

Motivation similar to graphical models

1) statistical efficiency: less # of parameters (simpler score fct. s_c)

$\Rightarrow$ easier to learn (generalization guarantees) [see Cortes & al. NIPS 2006 next class]

2) computational ?: compute argmax $s(\vec{y})$ $\vec{y} \in \mathcal{Y}$