Lecture 5 - scribbles - PAC-Bayes

Tuesday, January 22, 2019 14:33

today: PAC-Bayes
probit loss
review surrogate losses

PAC-Bayes

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

uniformly over all $w \in W$ (countable)

using complexity $\ln \frac{1}{q}$ (q "prior")

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

caveat: switch to a randomized predictor

ie. instead of $\hat{w}$ $\quad y = h_{\hat{w}}(x)$

consider q distribution over W

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

use $\mathbb{E}_{\hat{q}}[L(w)]$ as the generalization error for $\hat{q}$ ie. $\mathbb{E}_{(x,y)\sim p} \mathbb{E}_{w\sim\hat{q}} \ell(y, h_w(x))$

$\mathbb{E}_{\hat{q}}[L(w)]$ → empirical version $\hat{\mathbb{E}}_{\hat{q}}[\hat{L}(w)]$

$\mathbb{E}_{\hat{q}}[L_{L_n(w)}]$ → structured prediction, this will yield probit surrogate loss (see soon)

PAC-Bayes thm. [McAllester 1999, 2003]

(let $\ell(y,y') \in [0,1]$) for any fixed prior π over $\mathcal{W}$

and any dist. p on $\mathcal{X} \times \mathcal{Y}$

then with $\geq 1-\delta$ over $D_n \sim p^{\otimes n}$

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

new complexity term

note: if $\mathcal{W}$ is countable; let $q_{w_0} = \mathbb{1}\{w = w_0\}$

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

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

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

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

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

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

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

Why name probit?

binary classification $\mathcal{Y} = \{-1, +1\}$ with 0-1 loss

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

Let margin $a = y\langle w, \varphi(x)\rangle$

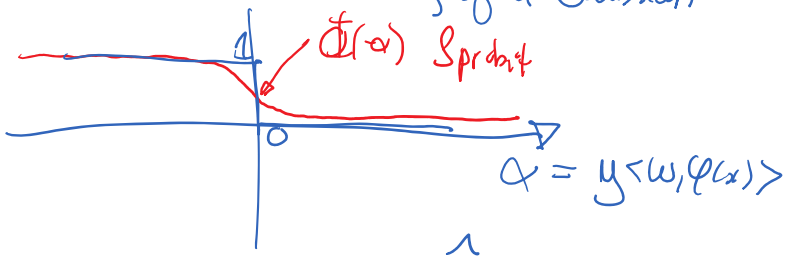
then $\ell_{\text{probit}}(x, y; w) = \mathbb{E}_{\varepsilon\sim N(0,I)}\, \mathbb{1}\{y \neq h_{w+\varepsilon}(x)\}$

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

(supposing $\|\varphi(x)\| = 1$) $\Rightarrow$ $a < -y\langle \varepsilon, \varphi(x)\rangle$

$= P\{\varepsilon_1 \geq a\} = \Phi(-a)$

Φ: cdf of a Gaussian



McAllester showed the consistency of the $\hat{w}_n^{\text{probit}}$ which minimizes a ℓ_2-reg. $\hat{\ell}_{\text{probit}}$

McAllester 2011 uses Catoni's PAC-Bayes version:

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

$\mathbb{E}_q[\hat{L}_n(w)] \Rightarrow \hat{\ell}_{\text{probit}}(w)$; $KL(q\|\pi) \Rightarrow \frac{1}{2}\|w\|^2$

if we use $\pi = N(0, I)$, $q_w = N(w, I)$

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

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

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

McAllester calls this "consistency"

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

(Lacoste-Julien unpublished fix: if $L(w)$ is cts.

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

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

proof idea: use Catoni's PAC-Bayes bound

with prob $\ge 1 - \delta_n$

$$\ell_{\text{probit}}(\hat{w}_n) \le \left(\frac{1}{1-\frac{1}{2\lambda n}}\right)\left(\underbrace{\hat{\ell}_{\text{probit}}(\hat{w}_n) + \frac{\lambda_n}{2n}\|\hat{w}_n\|^2}_{\text{pick } \alpha} + \ln\frac{1}{\delta_n}\right)$$

$$\le \hat{\ell}_{\text{probit}}(q_{\alpha w^*}) + \lambda_n \alpha^2 \|w^*\|^2$$

$$\leq \underbrace{\ell_{\text{probit}}(\alpha w^*)}_{} + \frac{\partial}{\partial \alpha}\frac{1}{2}\alpha^2 \|w^*\|^2$$

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

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

$\therefore$

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

15h31

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

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

Review of convex surrogates mentioned so far:

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

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

let $m(\tilde{y}) \triangleq s(y) - s(\tilde{y})$

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

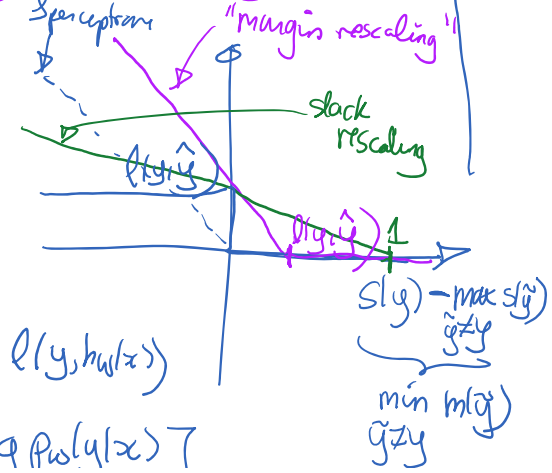
(structured SVM)

"margin rescaling" vs. "slack rescaling"

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

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

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



$$\ell_{\text{CRF}}(\text{---}) = \frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))\right) - s(y) \qquad \lceil -\log P_w(y|x) \rceil$$

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

(log-loss for CRF)

$\min_{\tilde{y} \neq y} m(\tilde{y})$

$\beta \to \infty \Rightarrow$ perceptron loss $> \frac{1}{\beta}\log\left(\sum_{\tilde{y}} \exp(\beta\, m(\tilde{y}))\right)$

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

[e.g. Pletscher & al. 2010]

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

but more computationally costly

what theoretical properties could we look at?

a) generalization error bounds [today]

b) consistency properties & calibration [next class]

why structured score functions? $s(x,y) = \sum_{c \in \mathcal{C}} s_c(x, y_c)$

motivations similar to graphical models

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

$\Rightarrow$ easier to learn (generalization guarantees) [see Cortes & al. NIPS 2016] [beg. of next class]

2) computational ": compute $\operatorname{argmax}_{\tilde{y} \in \mathcal{Y}} s(\tilde{y})$

BUT compare to what happens for Hamming loss:

BUT compare to what happens for Hamming loss:

given true conditional $q_x(y) \triangleq p(y|x)$ generating data $y = (y_1, \dots, y_p, \dots, y_L)$

expected error when using $\tilde{y}$ as prediction is $\mathbb{E}_{y \sim q_x(y)}[\ell(y, \tilde{y})] \overset{\text{def}}{=} \ell_{q_x}(\tilde{y})$

for Hamming loss: $\ell_{q_x}(\tilde{y}) = \mathbb{E}_{q_x(y)}\left[\sum_p \underbrace{\mathbb{I}\{y_p \neq \tilde{y}_p\}}_{1 - \mathbb{I}\{y_p = \tilde{y}_p\}}\right] = \sum_p (1 - q_x(\tilde{y}_p))$

$q_x(\tilde{y}_p)$: marginal on y_p i.e. $\triangleq \sum_{y' : y'_p = \tilde{y}_p} q_x(y')$

$\Rightarrow$ best decision $y^* = \underset{\tilde{y} \in \mathcal{Y}}{\operatorname{argmin}}\, \ell_x(\tilde{y})$

is just independent predictions $\tilde{y}_p^* = \underset{\tilde{y}_p}{\operatorname{argmax}}\, p(\tilde{y}_p|x)$ (marginal of $p(y|x)$)

max marginals decoding

⊛ thus; if there is no constraints, then can just train independently models for each part marginal $p(y_p|x)$ i.e. $s_p(y_p|x; w_p)$ e.g. $p(y_p|x) \propto \exp(s_p(y_p; x, w))$

but a) this function might be too complicated

b) statistically, could be beneficial to share learning together "transfer learning" between parts

16h25