Lecture 5 - PAC-Bayes

Tuesday, January 21, 2020 14:30

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 $w \in \mathcal{W}$ (but countable)

using complexity $|w|_\pi$
↳ "prior"

PAC-Bayes : generalizes this to
• arbitrary $\mathcal{W}$
• 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 $\hat{q}$ distribution over $\mathcal{W}$

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

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

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

empirical version

$\mathbb{E}_{\hat{q}}[\hat{L}_n(w)]$ ⇝ structured prediction
this will yield probit surrogate loss (see soon)

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

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

$$\frac{1}{\sqrt{2(n-1)}} \sqrt{KL(q\|\pi) + \ln\frac{n}{\delta}}$$

new complexity

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

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

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

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

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

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

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

why name probit?

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

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

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

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

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

(supposing $\|\varphi(x)\| = 1$)

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

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

$\ell_{probit} = P\{\varepsilon_1 < -\alpha\} = \Phi(-\alpha)$ $\quad -\alpha > \langle \varepsilon, \varphi(x)\rangle$

Φ: cdf of a Gaussian

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

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

McAllester 2011 uses Catoni's PAC-Bayes version:

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

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

if we use $\pi = N(0, \frac{1}{\lambda} I)$, $q_w = N(w, I)$ $\Rightarrow$ $\hat{\ell}_{probit}(w)$, $\frac{\lambda_n}{n}$, $\frac{1}{2}\|w\|^2$

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

thm 1

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

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

McAllester call this "consistency"

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

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

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

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

proof idea: use Catoni's PAC Bayes bound

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

$\hat{\ell}_{probit}(\hat{w}_n)$ picked

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

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

* also use $\lim_{\alpha\to 1} \ell_{probit}(\alpha w^*) \le L(w^*)$

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

15h37

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

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

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

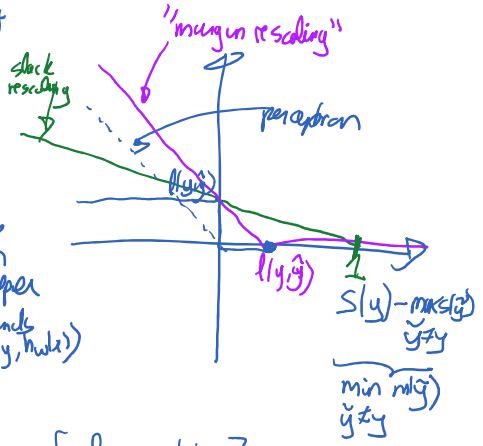
ℓ_{hinge} (structured svm) $= \max_{\tilde{y}} [s(\tilde{y}) + \ell(y,\tilde{y})] - s(y)$

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

vs.

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

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



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

$\frac{1}{\beta} \log \left(\sum_{\tilde{y}} \exp(-\beta m(\tilde{y})) \right)$ $\xrightarrow{\beta \to \infty}$ ⇒ perceptron loss

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

[e.g. Pletscher et 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 [next class]

b) consistency properties & calibration fct. [next² class]
↳ relationship between $L(w)$ & $\ell(w)$

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)

⇒ easier to learn (generalization guarantees) [see Cortes & al. NIPS 2016] next class

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

But consider what happens for Hamming loss $\quad y = (y_1, \dots, y_p, \dots, y_L)$

given true conditional $q_x(y) \triangleq p(y|x)$ generating data

Expected error when using $\tilde{y}$ as prediction is $\mathbb{E}_{y \sim q_x(y)}[\ell(y, \tilde{y})] \triangleq \ell_{q_x}(\tilde{y})$

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

marginal on $\tilde{y}_p$

ie. $\sum_{y' : y'_p = \tilde{y}_p} q_x(y')$

$\Rightarrow$ best decision $y^* = \operatorname{argmin}_{\tilde{y}} \ell_{q_x}(\tilde{y})$

is just independent predictions

max marginal decoding

$y_p^* = \operatorname{argmax}_{y_p} p(\tilde{y}_p | x)$

marginal of $p(y|x)$

(*) thus: if there is no constraints, then can just train independently models for each part marginal $p(y_p|x)$

ie. $s_p(y_p | x; w_p)$

eg. $p(y_p|x)$ or $\exp(s_p(y_p))$

but a) this function might be too complicated

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