# Lecture 16 - scribbles - PAC-Bayes

Friday, March 10, 2017
14:58

- theory: PAC-Bayes → probit loss

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

(usually): $h_w(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}}\ s(x, y; w)$

$\langle w, \varphi(x, y) \rangle$ (linear score)

aspects of structured prediction
1) constraints on $\mathcal{Y}$ ] theory open?
2) $\ell(y, y')$
3) structured score function:

$$s(x, y; w) = \sum_c s_c(x, y_c; w)$$

$L_P(w) = \mathbb{E}_{(x,y) \sim P}[\ell(y, h_w(x))]$ generalization error
V-risk

($L(w) \to P$ from context)

$\hat{L}_n(w) = \hat{L}_{D_n}(w) \triangleq \frac{1}{n} \sum_{(x^{(i)}, y^{(i)}) \in D_n} \ell(y^{(i)}, h_w(x^{(i)}))$

Theory for structured prediction:

Last lecture: Linked $\hat{L}_n(w)$ with $L_P(w)$

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

using complexity $|w|_\pi$ ← prior

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

by using randomized predictor
i.e. instead of $\hat{w}$ $y = h_{\hat{w}}(x)$

Consider $\hat{q}$ distribution over $W$

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

then we work with $\mathbb{E}_q[L(w)]$ as the generalization of this process

$\mathbb{E}_q[\hat{L}_n(w)]$ empirical version

PAC-Bayes thm. [McAllester 1999, 2003]

($\ell(y, y') \in [0, 1]$) for any fixed prior $\pi$ over $W$
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

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

if $W$ is countable; let $q(w) = \mathbb{I}\{w = w_0\}$

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

Probit loss for structured prediction → NIPS 2011 McAllester & Keshet

if $q(w') = N(w'|w, I)$

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

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

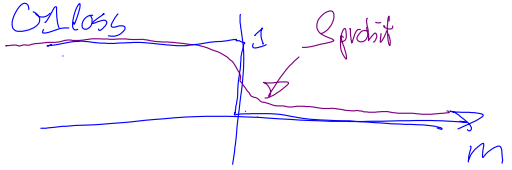
name probit: binary classification $Y = \{-1,+1\}$ 0-1 loss

$\varphi(x,y) = \frac{1}{2} y\varphi(x)$

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

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

then $\ell_{probit}(x,y;w) = P_{\varepsilon\sim N(0,I)}\{\varepsilon \ge m\} = \Phi(-m)$ (cdf of a normal)




McAllester 2011 uses Catoni's PAC-Bayes version

for fixed $\lambda > \frac{1}{2}$ (...)

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

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

$$\ell_{probit}(w) \le \left(\frac{1}{1-\frac{1}{2\lambda_n}}\right)\left(\hat{\ell}_{probit}(w) + \frac{\lambda_n}{n}\left(\frac{1}{2}\|w\|^2 + \ln\frac{1}{\delta}\right)\right) \quad \forall w$$

$$\left(\ell_{probit}(w) \triangleq \mathbb{E}_{(x,y)\sim P}\left[\ell_{probit}(x,y;w)\right]\right)$$

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

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

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

McAllester calls this "consistency"

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

then $L(\hat{w}_n) \xrightarrow{a.s.} L^* = \min_{w\in W} L(w)$

↑ consistency even in the misspecified setting

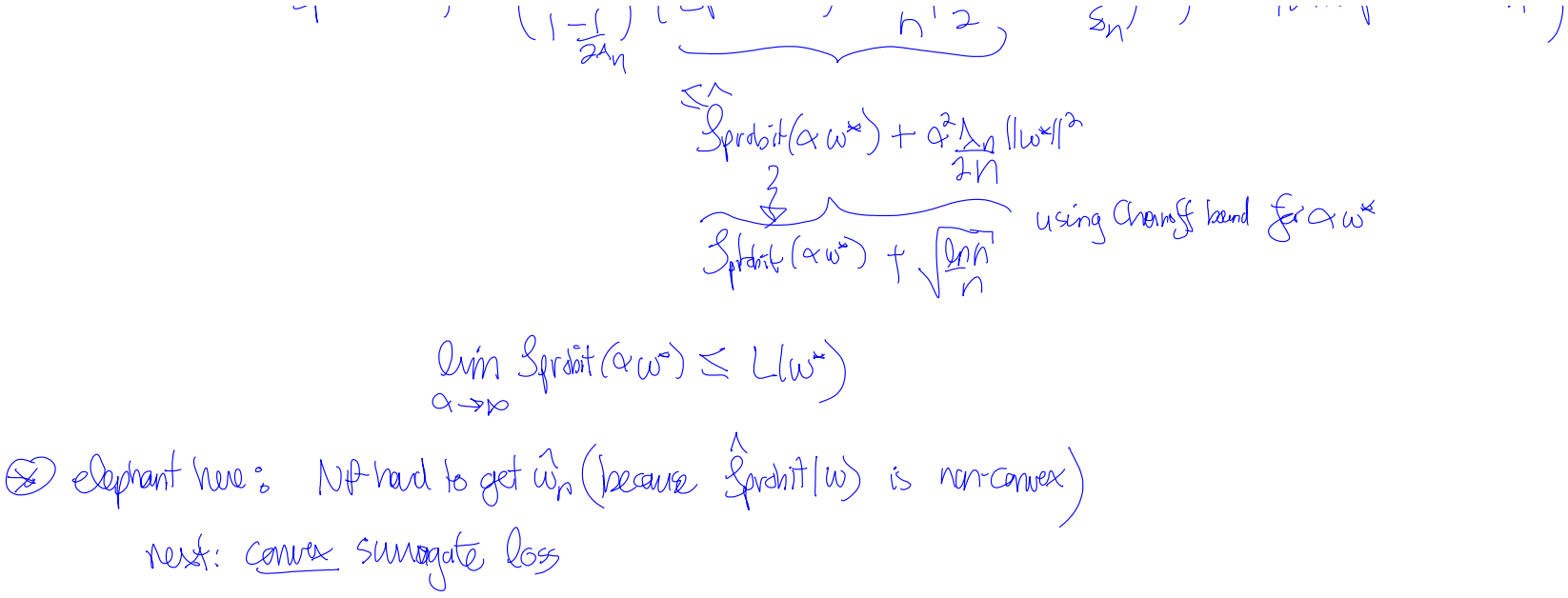
well-specified → let $h_P^* = \underset{h\in\mathcal{X}}{\operatorname{arg\,min}}\ L(h)$;

↳ $\exists\, \tilde{w}\in W$ s.t. $L(h_{\tilde{w}}) = L(h_P^*)$

$$L(\hat{w}_n) - L(h_P^*) = \underbrace{L(\hat{w}_n) - L(w^*)}_{\text{"estimation regret"}} + \underbrace{L(w^*) - L(h_P^*)}_{\text{"approximation regret"}}$$

proof idea: use Catoni's PAC-Bayes bound

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




**Pointers:**

- PAC-Bayes bound from McAllester 2003 was taken Lemma 4 in:
  - Generalization Bounds and Consistency for Structured Labeling in Predicting Structured Data, edited by G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. V. N. Vishwanathan. MIT Press, 2007
    http://nagoya.uchicago.edu/~dmcallester/colbounds.pdf
  - see also proof in his lecture notes that I had linked to last lecture:
    http://nagoya.uchicago.edu/~dmcallester/ttic101-07/lectures/generalization/generalization.pdf
- probit loss and its consistency for structured prediction:
  David McAllester, Joseph Keshet, Generalization Bounds and Consistency for Latent Structural Probit and Ramp Loss, (oral), NIPS 2011
  https://papers.nips.cc/paper/4268-generalization-bounds-and-consistency-for-latent-structural-probit-and-ramp-loss.pdf