# Lecture 5 - PAC-Bayes

Tuesday, January 23, 2024 9:32 AM

today: • PAC-Bayes
• probit loss
• review of 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
i.e. instead of having $\hat{w}$, predicting $y = h_{\hat{w}}(x)$
consider $\hat{q}$ distribution over $W$
predict: first $w \sim \hat{q}(w)$; $y = h_w(x)$

⇝ we 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
$\mathbb{E}_q[\hat{L}_n(w)]$ ⇝ our structured prediction will yield probit surrogate loss (see soon)
↑ optimize over $q$ to get $\hat{q}$

### PAC-Bayes theorem [McAllester 1999, 2003]

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

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

(new complexity term)

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

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

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

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

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

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

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

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

### 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 $\alpha \triangleq y\langle w, \varphi(x)\rangle$

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

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

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

$\alpha$

$-\alpha > \langle \underbrace{\varepsilon \cdot y}_{\text{Gaussian}}, \varphi(x)\rangle$

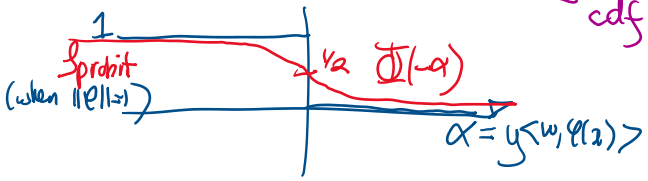
RHS is a Gaussian with mean $\mathbb{E}\langle \varepsilon y, \varphi(x)\rangle = \langle (\mathbb{E}\varepsilon) y, \varphi(x)\rangle = 0$ $\to N(0, \|\varphi(x)\|^2)$

$\mathbb{E}[y^2 \varphi^T \varepsilon \varepsilon^T \varphi]$

$\varphi^T \underbrace{\mathbb{E}[\varepsilon\varepsilon^T]}_{I} \varphi = \|\varphi\|^2$

$$\ell_{probit} = P\{\underbrace{\varepsilon_1}_{N(0,1)} \|\varphi\| < -\alpha\} = \Phi\left(\frac{-\alpha}{\|\varphi(x)\|}\right)$$

$\Phi$: cdf of standard Gaussian




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

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

McAllester 2011 use Catoni's PAC-Bayes thm. version

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

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

$$\hat{f}_{\text{probit}}(w) + \frac{\lambda_n}{n}\frac{\|w\|^2}{2} \quad (*)$$

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

10h24

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

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

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 cte., then $f_{\text{probit}}(\hat{w}_n) \xrightarrow{a.s.} L^* \Rightarrow L(\hat{w}) \xrightarrow{a.s.} L^* = L(w^*)$ ]

proof idea: use Catoni's PAC Bayes bound

with prob $\ge 1-\delta_n$

$$f_{\text{probit}}(\hat{w}_n) \le \left(\frac{1}{1-\frac{1}{2\lambda n}}\right)\left(\hat{f}_{\text{probit}}(\hat{w}_n) + \frac{\lambda_n}{2n}\|\hat{w}_n\|^2 + \frac{\lambda_n}{n}\ln\frac{1}{\delta_n}\right)$$

$$\le f_{\text{probit}}(\alpha w^*) + \frac{\lambda_n}{2n}\alpha^2\|w^*\|^2 \qquad (\text{set } \delta_n = \tfrac{1}{n^2})$$

$$\le f_{\text{probit}}(\alpha w^*) + \sqrt{\frac{\ln n}{n}}$$

using Chernoff bound for $\alpha w^*$ with prob $\ge 1-\frac{1}{n^2}$

$\Rightarrow$ get $\lim_{n\to\infty} f_{\text{probit}}(\hat{w}_n) \le f_{\text{probit}}(\alpha w^*)$ with prob 1

$\Leftarrow$ also use $\lim_{\alpha\to\infty} f_{\text{probit}}(\alpha w^*) \le L(w^*)$

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

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

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

Recall: all convex surrogates mentioned so far

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

## Review of convex surrogates mentioned so far

Let $m(\tilde{y}) = s(y) - s(\tilde{y})$

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

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

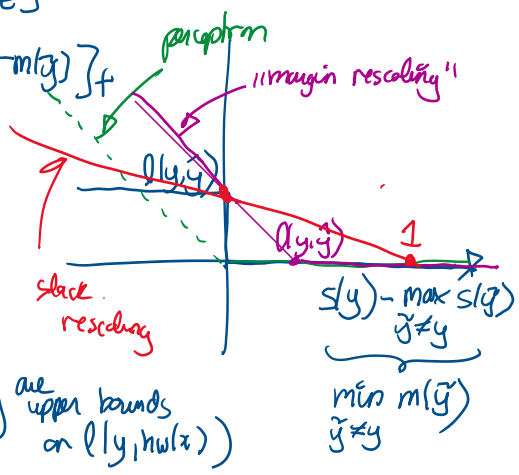
$$\ell_{\text{hinge}}(\text{structured SVM}) = \max_{\tilde{y}} [s(\tilde{y}) + \ell(y, \tilde{y})] - s(y)$$

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

"margin rescaling" vs. "slack rescaling" →

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

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




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

$$\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(-\beta m(\tilde{y}))\right)$$

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 $\ell(y, y')$ [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. [next$^2$ class]

↳ relationship between $L(w)$ & $\ell(w)$

## why structured score functions?

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

motivation similar to graphical models

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

⇒ easier to learn [See Cortes & al. NeurIPS 2016]

(generalization guarantees) next class

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