Lecture 4 - scribbles - PAC Bayes

Friday, February 2, 2018
14:31

today: concentration inequality
PAC-Bayes

Concentration inequality

Markov inequality: R.V. $X \geq 0$ (i.e. $X(\omega) \geq 0 \ \forall \omega \in \Omega$), $a > 0$

$$\boxed{P\{X \geq a\} \leq \frac{\mathbb{E}X}{a}}$$

proof:

$$a \mathbb{1}\{X \geq a\} \leq X$$

$$\mathbb{E}[\ \ ] \leq \mathbb{E}X$$

$$a P\{X \geq a\} \leq \mathbb{E}X \quad //$$

Corollary: if $\varphi: \mathbb{R} \to \mathbb{R}^+$ is an increasing fct. $x \geq a \Rightarrow \varphi(x) \geq \varphi(a)$

then $\varphi(x) \geq \varphi(a) \Leftarrow x \geq a$

$$P\{X \geq a\} \leq P\{\varphi(X) \geq \varphi(a)\} \leq \frac{\mathbb{E}\varphi(X)}{\varphi(a)}$$

(≤ true even if X is negative)

Chernoff trick: use $\varphi(x) \triangleq e^{tx}$ for $t > 0$

thus $P\{X \geq a\} \leq \frac{\mathbb{E}e^{tX}}{e^{ta}} \ \forall t > 0$ by Markov

ie. $P\{X \geq a\} \leq \inf_{t>0} \frac{\mathbb{E}e^{tX}}{e^{ta}}$

Next: Chernoff bound

$$\hat{L}_n(\omega) = \frac{1}{n} \sum_{i=1}^{n} \underbrace{\mathbb{1}\{y_i \neq h_\omega(x_i)\}}_{\triangleq B_i \sim \text{Bernoulli}(L(\omega))}$$

$$L(\omega) = \mathbb{E}\mathbb{1}\{y_i \neq h_\omega(x_i)\} = P\{y_i \neq h_\omega(x_i)\} = p$$

want to bound the prob that: $L - \hat{L}_n \geq \varepsilon$

to get ... bound,

Let $X \triangleq L - \hat{L}_n = \frac{1}{n}\sum_{i=1}^{n}(p - B_i)$ $\quad \mathbb{E}e^{tX} = \mathbb{E}\exp\left(\frac{t}{n}\sum_i (p - B_i)\right)$

Chernoff bound on wikipedia

$\overset{\text{independence of } B_i\text{'s}}{=} \prod_i \mathbb{E}\exp\left(\frac{t}{n}(p - B_i)\right)$

Hoeffding's Lemma:

for R.V. Y s.t. $a \le Y \le b$ and $\mathbb{E}Y = 0$

then $\mathbb{E}e^{tY} \le \exp\left(\frac{t^2}{8}(b-a)^2\right)$

(use Taylor's expansion of $\exp$.)

Let $Y = \frac{p - B_i}{n} \in \left[\frac{p-1}{n}, \frac{p}{n}\right] \Rightarrow (b-a) = \frac{1}{n}$

$\mathbb{E}e^{tY} \le \exp\left(\frac{t^2}{8n^2}\right)$

(wikipedia proof)

$$P\{L - \hat{L}_n \ge \varepsilon\} \le \frac{\left(\mathbb{E}\exp(tY)\right)^n}{e^{t\varepsilon}} \le \exp\left(\frac{t^2}{8n} - t\varepsilon\right) \quad \forall t > 0$$

min with respect to t $\Rightarrow$ $t^* = 4n\varepsilon$

i.e. $P\{L - \hat{L}_n \ge \varepsilon\} \le \exp(-2n\varepsilon^2)$

PAC-Bayes

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

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

using complexity $|w|_{2^{-t}}$ "prior"

PAC-Bayes : generalize this to

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

by using a randomized predictor

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

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

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

↳ structured prediction

this will yield probit surrogate loss (see soon)

PAC-Bayes thm. [McAllester 1999, 2003]

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

and any distribution p on $X \times Y$

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

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

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

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

probit loss for structured prediction: (NIPS 2011 McAllester & Keshet)

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

then $\mathbb{E}_{w \sim q_w}[L(w)] = \mathbb{E}_{w' \sim q} \mathbb{E}_{(x,y)\sim p}[\ell(y, h_{w'}(x))] = \mathbb{E}_{(x,y)\sim p}\left[\underbrace{\mathbb{E}_{\varepsilon \sim N(0,I)} \ell(y, h_{w+\varepsilon}(x))}_{\triangleq \ell_{\text{probit}}(x,y;w)}\right]$

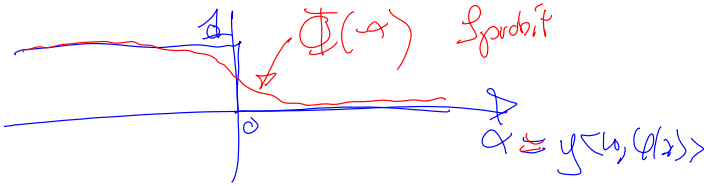
name probit: binary classification $y \in \{-1, +1\}$, 0-1 loss

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

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

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

$= P_{\varepsilon \sim N(0,I)}\{\varepsilon \geq \alpha\} = \Phi(-\alpha)$



McAllester 2011 uses Catoni's PAC-Bayes version:

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

$\hat{\ell}_{\text{probit}}(w)$; $\frac{1}{2}\|w\|^2$

if we use $\pi = N(0, I)$, $q_w = N(w, I)$ $\Rightarrow$ define $\hat{w}_n^{(\text{probit})} = \underset{w \in \mathbb{W}}{\text{argmin}}\ \hat{\ell}_{\text{probit}}(w) + \frac{\lambda_n}{2n}\|w\|^2$

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

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

McAllester calls this "consistency"

$w \in W$

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

(Lacoste-Julien unpublished fix: 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

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

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

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

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

$\Rightarrow \lim_{n \to \infty} \ell_{probit}(\hat{w}_n) = L(w^*)$ //

(see paper for details)