Lecture 5 - scribbles - generalization error bounds

Tuesday, February 6, 2018
14:33

today: · surrogate losses
· generalization error bounds

last time: $\ell_{probit}(x,y;w) = \mathbb{E}_{\varepsilon \sim N(0,\Sigma)}[\ell(y, h_{w+\varepsilon}(x))]$ $\quad h_w(x) = \operatorname{argmax}_{\tilde{y}\in\mathcal{Y}} s(x,\tilde{y};w)$

issue: non-convex in w $\Rightarrow$ no optimization guarantee

today: convex surrogate $\quad s(\tilde{y}) \triangleq s(x,\tilde{y};w)$ i.e. x & w are implicit

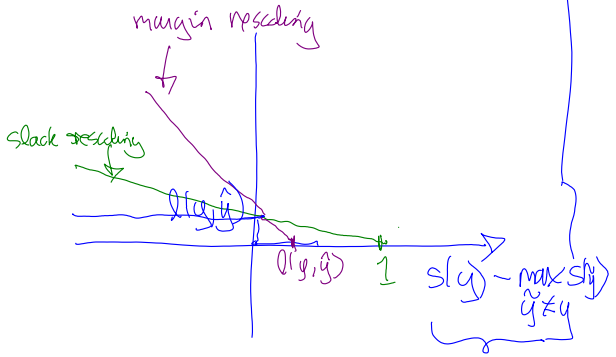
recap:

$$\ell_{perceptron}(x,y;w) = \max_{\tilde{y}\in\mathcal{Y}} s(\tilde{y}) - s(y) = \max_{\tilde{y}\in\mathcal{Y}} [-m(\tilde{y})] = [\max_{\tilde{y}\neq y} -m(\tilde{y})]_+$$

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

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

(structured SVM) "margin rescaling"



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

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

log-loss (CRF) (soft-max) $\quad \ell_{CRF}(x,y;w) = [\frac{1}{\beta}\log(\sum_{\tilde{y}} \exp(\beta s(\tilde{y})))] - s(y)$ $\quad [-\log p_w(y|x)]$

$\beta \to \infty \Rightarrow$ perceptron loss

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

"smoothed hinge loss"

$$\frac{1}{\beta}\log(\sum_{\tilde{y}} \exp(\beta[\ell(y,\tilde{y}) - m(\tilde{y})]))$$

[e.g. Pletscher & al. 2010]

what are theoretical properties?

what are theoretical properties?

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 later today, Cortes & al NIPS 2016]

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

BUT compare to what happens for Hamming loss:

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

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

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

$q_x(\tilde{y}_p) \triangleq \sum_{y : y_p = \tilde{y}_p} q_x(y)$

$\Rightarrow$ best decision $y^* = \operatorname{argmin}_{\tilde{y} \in \mathcal{Y}} \ell_x(\tilde{y})$ is just $\tilde{y}_p^* = \operatorname{argmax}_{\tilde{y}_p} p(\tilde{y}_p | x)$ (marginal of $p(y|x)$)

"marginal decoding"

* thus, if no constraint, i.e. no need "consistency" between parts,
could just train independently models for each part marginal $p(y_p|x)$
i.e. $Sp(y_p|x; w_p)$

model $p(y_p|x) \propto \exp(Sp(y_p; x, w))$

but a) this function might be too complicated
b) statistically, could be beneficial to share learning together "transfer learning"

generalization error bounds

for binary classification, a classical PAC bound is
for any fixed distribution on data
with prob $\geq 1-\delta$ over D_n

$$\forall w \in W.\ L_{01}(w) \leq \hat{L}_n(w) + \frac{1}{\sqrt{n}}\sqrt{d \log\frac{d}{n} + \log\frac{2}{\delta}}$$

where d is VC-dimension of $\mathcal{H} = \{h_w : w \in W\}$

VC-dimension of $\mathcal{H} \triangleq \max\{m :$ $\exists$ a set of m points s.t. for any labeling of m points, $\exists w$ s.t. h_w gives the correct labels on these points$\}$

"shattering the set of points"

also means that # of prediction functions on fixed m points is 2^m

for linear classifiers of p parameters, VC-dim $= p+1$

* one issue for this bound is that true for all distributions $\Rightarrow$ too loose bound

$\Rightarrow$ motivates going to data distribution dependent measure of complexity

example: empirical Rademacher complexity

$$\hat{R}_{D_n}(\mathcal{H}) \triangleq \mathbb{E}_\sigma\left[\sup_{h_w \in \mathcal{H}} \left|\frac{2}{n}\sum_{i=1}^{n} \sigma_i \mathbb{1}\{y_i \neq h_w(x_i)\}\right|\right]$$

"correlations with random noise"

$\sigma_i = \begin{cases} +1 \\ -1 \end{cases}$ uniformly "Rademacher" R.V.

$$\forall w \quad L_{D_n}(w) \leq \hat{L}_{D_n}(w) + \hat{R}_{D_n}(\mathcal{H}) + \frac{1}{\sqrt{n}} 3\sqrt{\frac{\log \frac{2}{\delta}}{2}}$$

complexity depends on D_n (implicitly on p)

high level idea to prove bound:

"double sample trick" → use a second sample D_n' for generalization error $L(w) = \mathbb{E}_{D_n'}[\hat{L}_{D_n'}(w)]$

+

"symmetrization trick" → bound the sup of differences between $L(w)$ & $\hat{L}_{D_n}(w)$

+

union bound as usual + concentration inequality

Structured prediction generalization bounds [Cortes & al. NIPS 2016]

general loss $\ell(y, y')$ s.t. $\ell(y, y') \neq 0$ if $y \neq y'$

suppose $s(x, y) = \sum_{c \in \mathcal{C}} s(x, y_c)$ → set of cliques of a graphical model / factor graph

thm. 7

with prob. $\geq 1-\delta$, depends on $\ell(y, y')$

$$\forall w \in W \quad L(w) \leq \underbrace{\ell_{\text{hinge}}(w)}_{\text{(truncated)}} + 4\sqrt{2}\,\hat{R}_{D_n}(H_w) + 3 L_{\max}\sqrt{\frac{\log \frac{1}{\delta}}{2n}}$$

$L_{\max} = \max_{y, y'} \ell(y, y')$

$\forall w \in \mathcal{W}$ $\quad L(w) \le \hat{f}_{\text{hinge}}(w) + 4\sqrt{2}\,\hat{R}^{G}_{D_n}(H_w) + 3L_{\max}\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$

where $\hat{R}^{G}_{D_n} \triangleq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{w\in\mathcal{W}} \sum_{i=1}^{n} \sqrt{|\mathcal{C}_i|} \sum_{c\in\mathcal{C}_i} \sum_{y_c \in \mathcal{Y}_c} \sigma_{i,c,y_c}\, s_c(x_i, y_c; w)\right]$

"empirical factor graph complexity"

only uses $(x^{(i)})_{i=1}^n$

$\mathcal{Y}_c$ = set of labels for y_c

σ_{i,c,y_c} — Rademacher R.V.

Thm. 2: if $s_c(x_i, y_c; w) = \langle w, \varphi_c(x_i, y_c)\rangle$

and consider $\mathcal{W}_\Lambda \triangleq \{w : \|w\|_2 \le \Lambda\}$; let $R = \max_{i,c,y} \|\varphi_c(x_i, y_c)\|_2$

then $\hat{R}^{G}_{D_n}(H_{\mathcal{W}_\Lambda}) \le \frac{R\Lambda}{\sqrt{n}}|\mathcal{C}|\sqrt{\max_c |\mathcal{Y}_c|}$

so want small cliques?

$L(w) \le \hat{f}_{\text{hinge}}(w) + \underbrace{\left(R\frac{|\mathcal{C}|}{\sqrt{n}}\sqrt{\max_c|\mathcal{Y}_c|}\right)}_{\lambda_n}\Lambda$ $\quad \Lambda \le \frac{\|w\|_2^2}{2}$

min of RHS gives SVMstruct $\hat{w}_n = \operatorname{argmin} \hat{f}_{\text{hinge}}(w) + \lambda_n\frac{\|w\|_2^2}{2}$

(if f is convex)

missing links: ① $\min f(w)$ s.t. $\|w\|_2 \le \Lambda$ $\rightsquigarrow$ $\exists\, \lambda(\Lambda)$ s.t. $\min f(w) + \lambda\frac{\|w\|^2}{2}$ ②

② gives same solution as ①

SVMstruct can be interpreted as minimizing upper bound on generalization error

properties: • minimize upper bound, hope that minimize $L(w)$

- can evaluate bound to get guarantees

caveat: minimizing an upper bound is not same thing as minimizing $L(w)$ (i.e. here, no consistency guarantees)

next: consistency