Lecture 17 - scribbles - surrogate

Friday, March 17, 2017
15:00

today: · surrogate loss
· generalization bounds
· consistency

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

convex surrogates so far

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

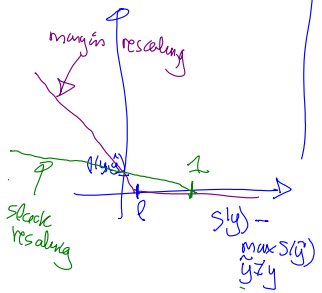
let $m(\tilde{y}) \triangleq s(y) - s(\tilde{y})$

Structural hinge loss (SVMstruct)

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

$\max_{\tilde{y} \in \mathcal{Y}} [s(\tilde{y}) + \ell(y,\tilde{y})] - s(y)$ "margin rescaling"
$\max_{\tilde{y}} [\ell(y,\tilde{y}) - m(\tilde{y})]$

$\max_{\tilde{y} \in \mathcal{Y}} \ell(y,\tilde{y}) [1 + s(\tilde{y}) - s(y)]$ "slack rescaling"
$\max_{\tilde{y}} \ell(y,\tilde{y}) [1 - m(\tilde{y})]$




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

"smoothed hinge"

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

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

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

[e.g. Fletcher et al. 2010]

what are (theoretical) properties?

a) generalization error bounds
b) consistency properties & calibration function

(*) 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 (see later today) [Cortes & al. 2016]
2) computational || to compute the $\operatorname{argmax}_{\tilde{y} \in \mathcal{Y}} s(\tilde{y})$

but compare to what happens for Hamming loss:

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

expected error of prediction $\tilde{y}$: $\mathbb{E}_{q_x(y)}[\ell(y,\tilde{y})] \triangleq \boxed{\ell_{q_x}(\tilde{y})}$

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

$\Rightarrow$ best decision is for each $p$, $\hat{y}_p = \underset{\tilde{y}_p}{\operatorname{argmax}}\, p(\tilde{y}_p|x)$ (marginal of $p(y|x)$)

$\rightarrow$ (if no constraints), no need of "consistency" between parts "marginal decoding"

- could just train independent models for each part marginal $p(y_p|x)$

but this function might be too complicated

$$\left(\begin{array}{l} \min f(x) \\ \text{s.t. } \|x\|^2 \leq 1 \end{array} \rightarrow \begin{array}{l} f(x) + \lambda(\|x\|^2 - 1) \\ x^*_\lambda = \underset{x}{\operatorname{argmin}} \end{array}\right)$$

claim: $x^*_\lambda \in \underset{x \text{ s.t. } \|x\|^2 \leq \|x_k^*\|^2}{\operatorname{argmin}} f(x)$

$$\min_x \max_\lambda f(x,\lambda) = \min_{x \text{ s.t. constraints}} f(x) = p^*$$

$$\max_\lambda \min_x f(x,\lambda) \leq p^*$$

(weak duality)

generalization error bounds:

for binary classification, a classical bound is (with prob $\geq 1-\delta$ over $D_n$)

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

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

$\Pi_{\mathcal{H}}(m) \triangleq \max\{m : \#\text{ of different prediction functions on } m \text{ examples}\} = 2^m$

for linear functions of $p$ parameters: VC-dim $= p+1$

bound above is true for any distribution

$\Rightarrow$ too loose

motivates going to data distribution dependent measure of complexity

example: empirical Rademacher complexity

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

"correlations with random noise"

$\sigma_i \in \{+1, -1}$ uniformly "Rademacher" R.V.

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

complexity depends on $D_n$ (implicitly on $P$)

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

+

"symmetrization trick" $\rightarrow$ bound the step of differences between $L(w)$ & $\hat{L}_n(w)$

+

union bound as usual

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

graphical model / factor graph

general loss $\ell(y,y')$ s.t. $\ell(y,y') \neq 0$ if $y \neq y'$ ; suppose $S(x,y) = \sum_{c\in\mathcal{C}} S(x, y_c)$

$\mathcal{C}$ a set of cliques

Thm. 7: with prob. $\ge 1-\delta$

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

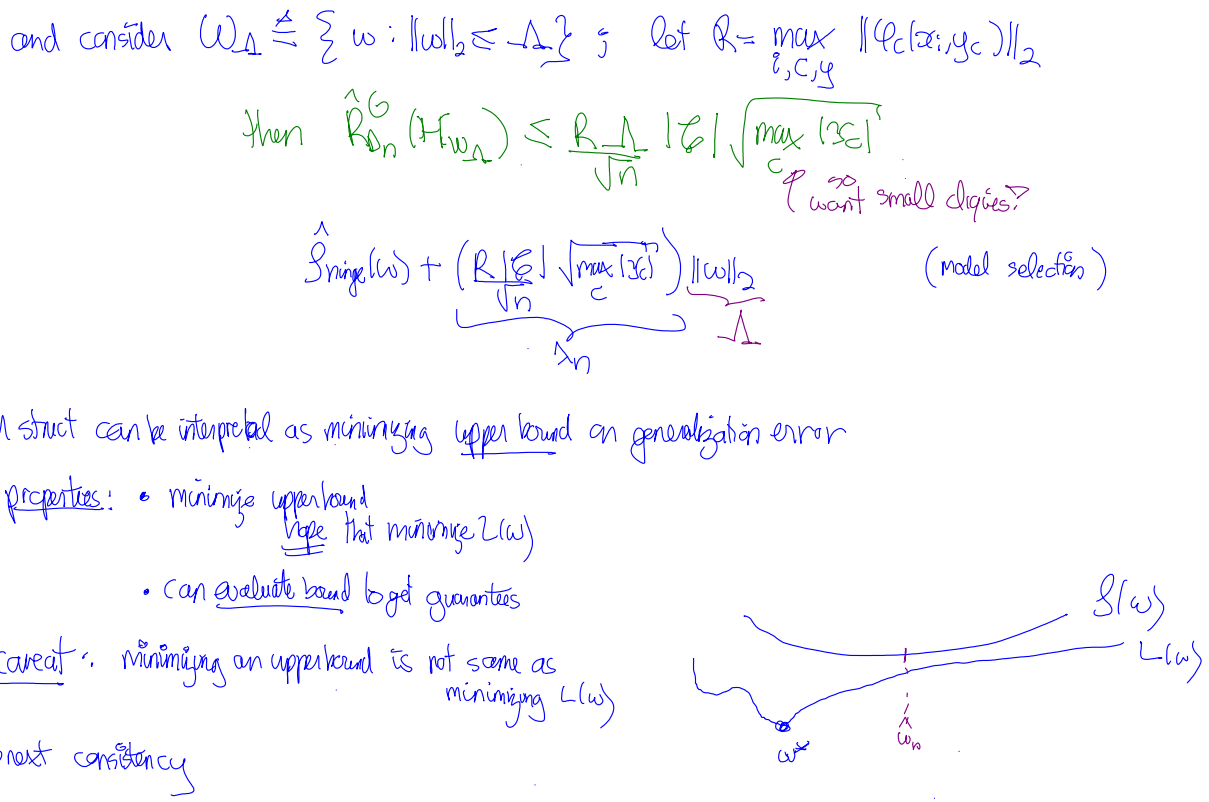
$\ell_{\max} \triangleq \max_{y\neq y'} \ell(y,y')$

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

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

"empirical factor graph complexity"

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




**pointers**

- sidenote: relationship between **constrained** and **regularized / penalized** formulations:
  - see section 4.7.3 (Pareto) and 4.7.4 (scalarization) of Boyd's book for the formal relationships
  - (in French): see exercise 2 from this homework from my ENS class: http://www.di.ens.fr/~slacoste/teaching/apprentissage-fall2015/TP/TP_5.pdf
- VC dimension / Rademacher complexity for binary case:
  - see slides from presentation by John Shawe-Taylor at MLSS 2009: http://mlg.eng.cam.ac.uk/mlss09/mlss_slides/ShawTaylor_1.pdf
    - VC dimension definition: slide 38
    - generalization error bound for binary classification with VC dimension: slide 46
    - Rademacher complexity: slide 85
    - generalization error bound with Rademacher complexity: slide 87
- structured prediction generalization bound:
  - Structured Prediction Theory Based on Factor Graph Complexity
    Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, Scott Yang
    NIPS 2016