Lecture 8 - calibrated convex surrogate losses

Thursday, January 30, 2020 13:35

today: consistency for convex surrogate losses

non-parametric viewpoint on scores

$s(x,y;w) = \langle w, \varphi(x,y)\rangle$ if $w = \sum_{i,\tilde{y}} \alpha_i(\tilde{y})\, \varphi(x_i,\tilde{y})$

$$\Rightarrow \langle w, \varphi(x,y)\rangle = \sum_{i,\tilde{y}} \alpha_i(\tilde{y}) \underbrace{\langle \varphi(x,y), \varphi(x_i,\tilde{y})\rangle}_{k(x,x_i;\, y,\tilde{y})}$$

often for simplicity: $k(x,x';\, y,y') = k_x(x,x')\, k_y(y,y')$

[is equivalent to having $\varphi(x,y) \triangleq \varphi_x(x) \otimes \varphi_y(y)$ "product kernel"]

$\otimes$ Kronecker product

$v \otimes w \quad vw^T$

$\langle v\otimes w, v'\otimes w'\rangle = \mathrm{tr}\left((vw^T)^T(v'w'^T)\right)$

$= \mathrm{tr}\left(w \underbrace{v^T v'}_{\langle v,v'\rangle} w'^T\right)$

$= \langle w, w'\rangle \langle v, v'\rangle$ //

e.g. $k_x(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right)$ RBF kernel (universal)

$\varphi_y : \mathcal{Y} \to \mathbb{R}^d \quad d \ll |\mathcal{Y}| \triangleq k \quad k_y(y,y') = \langle \varphi_y(y), \varphi_y(y')\rangle$

Back to consistency & surrogate losses

$$\hat{w}_n \triangleq \operatorname*{argmin}_w \hat{L}_n(w) + \frac{\lambda_n \|w\|^2}{2}$$

consistency: $L(\hat{w}_n) \xrightarrow{n\to\infty} \min_w L(w)$

(*) binary classification [Bartlett & al. 2004] characterized a whole family of consistent surrogate losses

↳ binary SVM
logistic regression

for multiclass classification [Lee & al. 2004] showed that multiclass hinge loss
McAllester 2007
is not consistent for 0-1 loss
when have no "majority" class
(i.e. $p(\tilde{y}|x) < \frac{1}{2}\ \forall \tilde{y}$)

$$S_{hinge}(s,y,w) = \max_{\tilde{y}} \left(s(\tilde{y}) + \ell(y,\tilde{y})\right) - s(y)$$

true p

↳ they propose a different surrogate loss that uses $\underset{\tilde{y}}{\sum}$ instead of $\underset{\tilde{y}}{\max}$
which is constant for 0-1 loss
exponential sum → could be intractable for structured prediction

2 aspects of structured pred. which give a much richer theory than binary classif. for Consistency

1) "noise model" $p(y|x)$ is much richer
2) $\ell(y,y')$ much richer

⊛ [Osokin & al. 2017] → we looked at effect of $\ell(y,y')$
for a easy to analyze convex surrogate loss & consistent
in the simplest possible setting
and we were careful about exponential constants (e.g. $\gamma = |\mathcal{Y}|$)

calibration function for a structured loss ℓ, surrogate loss $\mathcal{L}$ and set W

$$H_{\mathcal{L},\ell,W}(\varepsilon) \triangleq \inf_{\substack{w \in W \\ q \in \Delta_{|\mathcal{Y}|}}} \left[\mathcal{L}_q(w) - \min_{w' \in W} \mathcal{L}_q(w')\right] \quad \text{s.t.} \quad L_q(w) - \min_{w' \in W} L_q(w') \geq \varepsilon$$

(x is fixed outside and q is a potential $p(y|x)$)

$\mathcal{L}_q(w) \triangleq \mathbb{E}_{q(\tilde{y})}[\mathcal{L}(x,\tilde{y},w)]$

$L_q(w) \triangleq \mathbb{E}_{q(\tilde{y})}[\ell(\tilde{y}, h_w(x))]$ "conditional risk"

↳ smallest "optimization regret" (surrogate) (over all dist-q) s.t. true regret $\geq \varepsilon$

ie. $\forall q : \mathcal{L}_q(w) < \mathcal{L}_q^* + H(\varepsilon) \Rightarrow L_q(w) \leq L_q^* + \varepsilon$

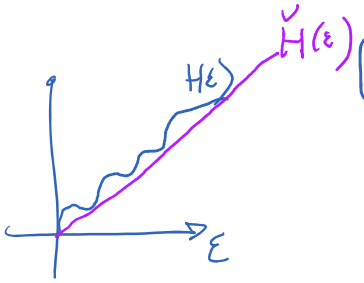
[Conditional on x version]

(thm. 2) $\forall p : \mathcal{L}(w) < \mathcal{L}^* + \overset{\vee}{H}(\varepsilon) \Rightarrow L(w) \leq L^* + \varepsilon$

(thm. 2) $\forall p$: $\mathcal{L}(w) < \mathcal{L}^* + \check{H}(\varepsilon) \Rightarrow L(w) \leq L^* + \varepsilon$

($\mathcal{L}$ basically uses Jensen's ineq.)

$\check{H}$: convex lower envelope of $H(\varepsilon)$



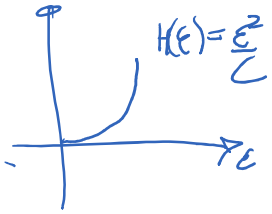
$\check{H}(\varepsilon) \triangleq H^{**}(\varepsilon)$ $\quad f^*(z) \triangleq \sup_x x^T z - f(x)$ "Fenchel-Legendre" conjugate

if $\check{H}$ is invertible

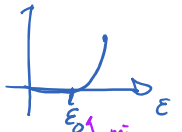
$$L(w) - L^* \leq \check{H}^{-1}(\mathcal{L}(w) - \mathcal{L}^*)$$

$\mathcal{L}$ is consistent iff $H(\varepsilon) > 0 \ \forall \varepsilon > 0$ (and $H(\varepsilon)$ is finite for some ε)

standard H:



$H^{-1}(z) = \sqrt{Cz}$ $\Rightarrow$ $L(w) - L^* \leq \sqrt{C(\mathcal{L}(w) - \mathcal{L}^*)}$



you want small C; for structured prediction $C = |\mathcal{Y}|$ often?

14h48

note: scale of H is arbitrary;

normalize it using optimization perspective (e.g. SGD) [see next class]

Simplest surrogate loss: square loss?

$s(\cdot) \in \mathbb{R}^K$ (for x)

$$\mathcal{L}(x, y; s) \triangleq \frac{1}{2K} \| s - (-\ell(y, \cdot)) \|_2^2 = \frac{1}{2K} \sum_{\tilde{y}} (s(x, \tilde{y}) + \ell(y, \tilde{y}))^2$$

[can be seen as a generalization of squared loss for binary class. to multiclass $(1 - y_i \langle a, \ell(x_i) \rangle)^2$]

$$\mathcal{L}_q(s) \triangleq \mathbb{E}_{q(y)} \mathcal{L}(x, y; s) = \frac{1}{2K} \sum_{\tilde{y}} \mathbb{E}_{q(y)} [s(\tilde{y})^2 + 2 s(\tilde{y}) \ell(y, \tilde{y}) + \text{cst.}] \overset{\text{does not depend on } s}{=}$$

$\mathbb{E}_{q(y)} \ell(y, \tilde{y}) \triangleq \ell_{q_x}(\tilde{y})$

$$= \frac{1}{2K} \| s + \ell_{q_x} \|^2 + \text{cst.}$$

suppose s is unconstrained $\min_s \mathcal{L}_{q_x}(s) \Rightarrow s^*(\tilde{y}) = -\ell_{q_x}(\tilde{y})$

$\operatorname{argmax}_{\tilde{y}} s^*(\tilde{y}) = \operatorname{argmin}_{\tilde{y}} \ell_{q_x}(\tilde{y})$

ie. you predict optimally pointwise

so here ℓ is <u>consistent</u>

ie. $s^* \in \operatorname{argmin}_{\text{all } s:\mathcal{X}\to\mathbb{R}^K} \ell(s) \Rightarrow L(h_{s^*}) = \min_{\text{all } h} L(h)$

$$\ell_q(s) - \min_{s' \in \mathbb{R}^K} \ell_q(s') = \frac{1}{2K} \|s - (-\ell_q)\|_2^2$$

(ℓ_q^* = $\min_{s'} \ell_q(s')$)

Let $\overleftrightarrow{L}$ be a $k \times k$ matrix where $\overleftrightarrow{L}_{\tilde{y},y} = \ell(y,\tilde{y})$ $\quad q_x = \sum_y q(y|x)\, \ell(y,\cdot)$

$$\ell_{q_x} = \overleftrightarrow{L} q_x$$

recall: $s^* = -\ell_{q_x} = -\overleftrightarrow{L} q_x \in \operatorname{span}(\overleftrightarrow{L})$ ie. $\sum_y \alpha_y \overleftrightarrow{L}(\cdot, y)$

to get <u>consistency</u> for ℓ, it is sufficient to consider $S \in \operatorname{span}(\overleftrightarrow{L})$

or that $S \in \operatorname{span}(F) \supseteq \operatorname{span}(\overleftrightarrow{L})$

restriction on scores

$F \in \mathbb{R}^{K \times r}$ matrix

can be chosen cleverly depending on $\overleftrightarrow{L}$

$S = F\theta \quad \theta \in \mathbb{R}^r$

true if $\operatorname{span}(F) \supseteq \operatorname{span}(\overleftrightarrow{L})$

$$\ell_q(\theta) - \min_{\theta \in \mathbb{R}^r} \ell_q(\theta) = \frac{1}{2K} \|F\theta - (\overleftrightarrow{L} q)\|_2^2$$

lower bound $\Rightarrow$ easiness result

Thm. 7 if $\operatorname{span}(F) \supseteq \operatorname{span}(\overleftrightarrow{L})$ $\quad H_{\ell_{\text{squared}},\ell,F}(\varepsilon) \geq \frac{\varepsilon^2}{2K \max_{i \neq j} \|P_F \Delta_{ij}\|_2^2} \geq \frac{\varepsilon^2}{4K}$

this is bad

$\Delta_{ij} \triangleq e_i - e_j \in \mathbb{R}^K$

P_F is orthogonal projection on $\operatorname{span}(F)$ $\quad P_F = F(F^TF)^+F^T$

- in paper, we show that for 0-1 loss, $H(\varepsilon) = \frac{\varepsilon^2}{4K}$

thm. 8: if $\operatorname{span}(F) = \mathbb{R}^K$ (i.e. no constraints) $\quad$ hardness result

thm. 8: if $\text{span}(\mathcal{F}) = \mathbb{R}^K$ (i.e. no constraints)

hardness result

then $H(\epsilon) \leq \frac{\epsilon^2}{2K}$ for any loss?

i.e. for any loss, we need an exp # of samples (in the worst case) to learn "well"

[caveat → all these are bounds and worst case]

but for Hamming loss, if add constraints that $s(\tilde{y}) = \sum_p s(\tilde{y}_p)$

P parts

over T binary variables, $H(\epsilon) = \frac{\epsilon^2}{8T}$ } not too big

⇒ we can recurse

note: computation how to compute $\sum_{\tilde{y}} \ell(y,\tilde{y}) s(\tilde{y})$

→ efficient to compute for Hamming loss & separable score e.g.