Lecture 8 - calibrated convex surrogate losses

Tuesday, February 9, 2021 14:41

today: consistency for convex surrogate losses

non-parametric viewpoint on scores

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

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

often for simplicity: $K(x,x';y,y') = K_x(x,x')\, K_y(y,y')$

[is equivalent to have $\varphi(x,y) \triangleq \varphi_x(x) \otimes \varphi_y(y)$]

$\otimes$ Kronecker product "product kernel"

$v \otimes w \quad vw^T$

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

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

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

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

$\varphi_y : \mathcal{Y} \to \mathbb{R}^d \quad d << |\mathcal{Y}| \triangleq k \quad$ e.g. 10^{23} $\quad k_y(y,y') = \langle \varphi_y(y), \varphi_y(y')\rangle$

back to consistency & surrogate losses

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

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

* binary classification [Bartlett & al. 2004] characterized a whole family of consistent (convex) surrogate losses ?
  - binary SVM
  - logistic regression

for multiclass classification [Lee & al. 2004] showed that multiclass hinge loss
McAllester 2007

$$\ell_{hinge}(x,y;w) = \max_{\tilde{y}} (s(\tilde{y}) + \ell(y,\tilde{y})) - s(y)$$

is not consistent for 0-1 loss when have no "majority" class

(i.e. $p(\tilde{y}|x) < \frac{1}{2}\ \forall \tilde{y}$)

p true p

→ they propose a different surrogate loss that use $\sum_{\tilde{y}}$ instead of $\max_{\tilde{y}}$

which is consistent for 0-1 loss

exponential sum → could be intractable for structured prediction

2 aspects of structured prediction which give a richer theory than binary class. 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. $|\mathcal{Y}| = k$)

15h18

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

$$H_{\mathcal{L},\ell,W}(\varepsilon) \triangleq \inf_{w \in W,\ q \in \Delta_{|\mathcal{Y}|}} \left[\mathcal{L}_q(w) - \min_{w' \in W} \mathcal{L}_q(w') \right]$$

$\mathcal{L}_q^*$

$$\text{s.t. } L_q(w) - \min_{w' \in W} L_q(w') \geq \varepsilon$$

L_q^*

x is fixed outside

q is a potential $p(y|x)$

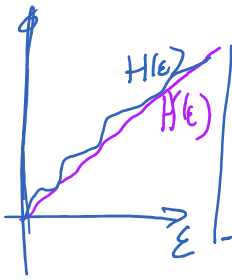
$$\mathcal{L}_q(w) \triangleq \mathbb{E}_{q(\tilde{y})}[\mathcal{L}(x,\tilde{y};w)]$$

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

"conditional (logit) risk"
[conditional on x version]

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

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



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

$\mathbb{E}_{(x,y)\sim p}[\ell(x,y,w)] \Rightarrow L(w) \leq L^* + \varepsilon$

(↑ basically shown using Jensen's ineq.)

convex lower envelope of $H(\varepsilon)$

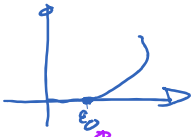
$\breve{H}(\varepsilon) \triangleq H^{**}(\varepsilon)$ $\quad f^*(z) \triangleq \sup_x x^T z - f(x)$ ↔ "Fenchel-Legendre" conjugate

If $\breve{H}$ is invertible

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

ℓ is consistent iff $H(\varepsilon) > 0 \ \forall \varepsilon > 0$

(and $H(\varepsilon)$ is single for some $\varepsilon = 0$)



min regret I can guarantee

standard H :

$H(\varepsilon) = \frac{\varepsilon^2}{C}$ $\quad H^{-1}(z) = \sqrt{Cz}$ $\Rightarrow$

$$L(w) - L^* \leq \sqrt{C(\mathcal{L}(w) - \mathcal{L}^*)}$$

you want small C ; for structured prediction $C = |\mathcal{Y}|$ often! (bad)

note : scale of H is arbitrary

normalize it using stochastic optimization properties (e.g. SGD) [next class]

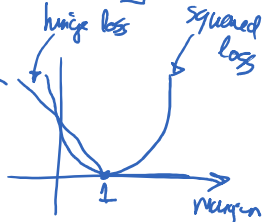
Concrete example : Simplest surrogate loss : square loss ?

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

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

[can be seen as a generalization of squared loss for binary class. to multiclass]

$$[1 - y_i \langle w, \varphi(x_i) \rangle]^2$$



$$\mathcal{L}_q(s) \triangleq \mathbb{E}_{q(y)} \ell(x, y, s)$$

$$= \frac{1}{2K} \sum_{\tilde{y}} \mathbb{E}_{q(y)} \left[s(\tilde{y})^2 + 2 s(\tilde{y}) \ell(y, \tilde{y}) \right] + \text{cst.}$$

does not depend on s

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

$$= \frac{1}{2k} \|s + \ell_{q_x}\|_2^2 + \text{cst.}$$

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

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

i.e. you predict optimally pointwise on x

so here ℓ is consistent i.e. $s^* \in \operatorname{argmin}_{\forall s: \mathcal{X} \to \mathbb{R}^k} \ell(s)$

$$\Rightarrow L(h_{s^*}) = \min_{\text{all } h} L(h)$$

$$\underbrace{\ell_q(s) - \min_{s' \in \mathbb{R}^k} \ell_q(s')}_{\ell_q^*} = \frac{1}{2k} \|s - (-\ell_{q_x})\|_2^2$$

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

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

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

(*) to get consistency 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 $\overleftrightarrow{L}$

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

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

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 loose

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

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

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

Thm. 8: if $\text{span}(F) = \mathbb{R}^K$ (i.e. no constraints) — hardness result

then $H(\varepsilon) \leq \frac{\varepsilon^2}{4K}$ for any loss?

i.e. for any loss, we need an exp. # of samples (in worst case) to learn well [correct → all those bounds are worst case]

(*) but for Hamming loss, if add constraints that $s(\tilde{y}) = \sum_{\text{parts}} s_p(\tilde{y}_p)$

over T binary variables, $H(\varepsilon) = \frac{\varepsilon^2}{8T}$ } not too big ⇒ we can learn?

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

→ efficient to compute for Hamming loss & separable score fct. eg