Lecture 8 - scribbles - Calibrated convex surrogate losses

Thursday, January 31, 2019 13:06

today: consistency for convex surrogate losses

Non-parametric viewpoint on scores

$s(x,y;w) = \langle w, \varphi(x,y) \rangle$ $\quad$ $k(x_i,\cdot\,;\tilde{y},\cdot)$

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

"product kernel" ⇑

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

$\otimes$ Kronecker product

$v \otimes w \qquad v w^T$

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

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

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

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

RBF kernel (universal)

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

Back to consistency & surrogate losses

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

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

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

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

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

is not consistent for 0-1 loss when have no majority class (i.e. $p(y|x) < \frac{1}{2} \; \forall y$)

→ they propose a different surrogate loss that $\sum_{\tilde{y}}$ instead of $\max_{\tilde{y}}$ which is consistent for 0-1 loss

exponential sum → could be intractable

for 0-1 loss

exponential sum → could be intractable

2 aspects of structured prediction which give a much richer theory than binary classification 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 in the simplest possible setting and we were careful about exponential constants

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

$$H_{\mathcal{L},\ell,\mathcal{W}}(\varepsilon) \triangleq \inf_{\substack{w \in \mathcal{W} \\ q \in \Delta_{|\mathcal{Y}|-1}}} \left[\mathcal{L}_q(w) - \min_{w' \in \mathcal{W}} \mathcal{L}_q(w') \right]$$

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

(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")

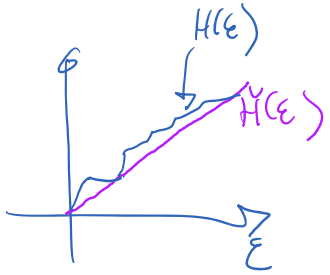
↳ smallest optimization surrogate regret (over all dist. q) s.t. true regret $\geq \varepsilon$

(conditional on x version)

$\forall q: \quad g_q(w) < g_q^* + H(\epsilon) \Rightarrow L_q(w) \leq L_q^* + \epsilon$

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

$\forall_{x,y} g(x,y,w)$



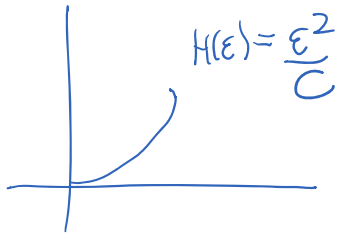
$\check{H}$ convex lower envelope of H

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

g is consistent iff $H(\epsilon) > 0 \ \forall \epsilon > 0$
(and $H(\epsilon)$ is finite for some $\epsilon > 0$)

if $\check{H}$ is invertible

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



$\Rightarrow \quad L(w) - L^* \leq \sqrt{C(g(w) - g^*)}$

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

14h24

note: scale of H is arbitrary ; normalize it using optimization perspective (e.g. SGD)

[see soon]

simplest surrogate loss: square loss?

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

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

[can be seen as generalization of squared loss for binary classification to multiclass $(1 - y_i \langle w, \phi(x_i) \rangle)^2$]

$$\ell_q(s) \triangleq \mathbb{E}_{q(y)} \ell(x, y; s)$$

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

(constant does not depend on s)

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

suppose s is unconstrained

$$\min_s \ell_{q_x}(s) \qquad s(\tilde{y}) + \ell_{q_x}(\tilde{y}) = 0 \quad \forall \tilde{y}$$

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

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

i.e. you predict optimally (pt. wise)

so here ℓ is consistent

i.e. $s^* \in \operatorname{argmin}_{\text{all } s} \ell(s) \Rightarrow L(h_{s^*}) = \min_{\text{all } h} L(h)$

$h_S^*(x) = \underset{\tilde{y}}{\operatorname{argmax}}\ S(x,\tilde{y})$

$$g_q(s) = \|s - (-\ell q)\|^2 + \underline{cst}$$

$$g_q(s) - \min_{s' \in \mathbb{R}^k} g_q(s') = \frac{1}{2k}\|s - (-\ell q)\|_2^2$$

$$\ell q_x = \sum_y q(y|x)\,\ell(y,\cdot)$$

Let $\overleftrightarrow{L}$ be a $k \times k$ matrix where $\overleftrightarrow{L}_{\tilde{y},y} = \ell(y,\tilde{y})$

$$\ell q_x = \overleftrightarrow{L}\, q_x$$

$$\Downarrow$$

$$s^* = -\ell q_x = -\overleftrightarrow{L}\, q_x \in \operatorname{span}(\overleftrightarrow{L}) \quad \text{i.e.} \quad \sum_{\tilde{y}} q_y\, \overleftrightarrow{L}(\cdot,\tilde{y})$$

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 $\overleftrightarrow{L}$

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

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

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

Cover bound $\Rightarrow$ easiness result

thm. 7

if $\text{span}(F) \supseteq \text{span}(L)$

$$H_{\ell_{\text{squared}},\ell,F}(\varepsilon) \geq \frac{\varepsilon^2}{2k \max_{i \neq j} \|P_F \Delta_{ij}\|_2^2} \geq \frac{\varepsilon^2}{4k}$$

Cover bound ⇒ easiness result

this bad

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

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

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

thm. 8: if $\text{span}(F) = \mathbb{R}^k$ (ie. no constraints) hardness result

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

ie. for any loss, we need an exponential # of samples (in the worst case) to learn "well" { caveat → all those are bounds and worst case }

⊛ but for hamming loss, if add constraints that $s(\tilde{y}) = \sum_{p \in \text{parts}} s_p(\tilde{y})$

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

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

→ efficient for Hamming loss & separable score ...

optimization formulation

Setup Let $s(x,\tilde{y})$ be of the form $F\Theta(x)$ $\Theta(x) \in \mathbb{R}^r$

$\Theta_j(\cdot) \in \mathcal{H}$ ← RKHS

Θ_j: optimization variables

$$g(\Theta) = \mathbb{E}_{(x,y)\sim p}\, g(x,y,\Theta)$$

$(x^{(t)}, y^{(t)}) \overset{iid}{\sim} p$

run projected SGD (kernelized) on $g(\Theta)$ i.e. $\Theta^{(t+1)} = P_{\mathcal{H}}\{\Theta^{(t)} - \gamma \nabla_\Theta g(x^{(t)}, y^{(t)}, \Theta^{(t)})\}$

$P_{\mathcal{H}}$: a ball of radius D

$\nabla_\Theta g(x^{(t)}, y^{(t)}, \Theta^{(t)}) = F^T \nabla_S g(x^{(t)}, y^{(t)}, S)\, \Phi(x^{(t)})^T$

Φ: feature map on $\mathcal{H}$

$K(x^{(t)}, \cdot)$

Convergence result (thm. 5): if $\|\Theta^*\|_{HS} \le D$

if $\mathbb{E}_{(x,y)\sim p} \|\nabla_\Theta g(x,y,\Theta)\|_{HS} \le M^2$

then averaged SGD with step-size $\gamma = \frac{2D}{M\sqrt{n}}$

$$\mathbb{E}[g(\bar{\Theta}^{[n]})] - g(\Theta^*) \le \frac{2DM}{\sqrt{n}}$$

$\bar{\Theta}^{[n]} = \frac{1}{n}\sum_{t=1}^{n} \Theta^{(t)}$

(convergence result)

$\mathbb{E}[J(\theta^{(n)})] - J(\theta^*) \leq \frac{2LM}{\sqrt{n}}$ (convergence result)

$\theta^{(n)} = \frac{1}{n}\sum_{t=1}^{n} \theta^{(t)}$

thm. 6 learning complexity (to be cted)?