Lecture 8 - calibrated convex surrogate losses

Thursday, February 1, 2024 9:35 AM

today: consistency for convex surrogate losses

non-parametric viewpoint on scores

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

$\Rightarrow \langle w, \varphi(x,y)\rangle = \sum_{i,\tilde{y}} \alpha_{i|\tilde{y}} \langle \varphi(x_i,\tilde{y}), \varphi(x,y)\rangle$

$k((x_i,\tilde{y}),(x,y))$

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

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

$\otimes$: Kronecker product

$v \otimes w \quad vw^T$

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

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

$= \langle v,v'\rangle \mathrm{tr}(w^Tw') = \langle v,v'\rangle\langle w,w'\rangle$ //

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

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

$k \sim 10^{23}$

$s(\tilde{y}) = \langle w_{\tilde{y}}, \varphi(\tilde{y})\rangle$

back to consistency & surrogate losses

$\hat{w}_n = \underset{w}{\mathrm{argmin}}\ \hat{\mathcal{L}}_n(w) + \lambda_n \frac{\|w\|^2}{2}$

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

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

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 we $\sum_{\tilde{y}}$ 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 gives a richer theory than binary class. for consistency

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

(*) [Ozkan & al. 2017] → we looked at effect of $\ell(y,\tilde{y})$

for a easy to analyze convex surrogate loss & consistent in the simplest possible setting

and we were careful about exponential constants (e.g. $|\mathcal{Y}| \geq k$)

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

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

($\min_{w' \in W} \mathcal{L}_q(w') = \mathcal{L}_q^*$)

s.t. $L_q(w) - \min_{w' \in W} L_q(w') \geq \varepsilon$ ("for ε-bad w's")

($\min_{w' \in W} L_q(w') = L_q^*$)

(*) x is fixed outside

q is a potential $p(\tilde{y}|x)$

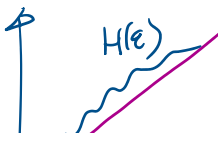
$\mathcal{L}_q(w) \triangleq \mathbb{E}_{q(\tilde{y})}[\mathcal{L}(x,\tilde{y},w)]$

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

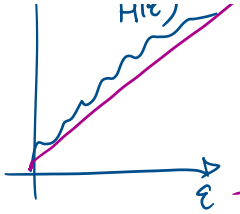
"conditional (Vapnik) risk" [conditional on x version]

L_q → smallest (over all dist. q) "surrogate optimization regret" 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) $\forall p:\ \mathcal{L}(w) < \mathcal{L}^* + \breve{H}(\varepsilon)$



(thm. 2) $\forall p: \underbrace{\ell(w)}_{\mathbb{E}_{(x,y)\sim p}[\ell(x,y;w)]} < \ell^* + \check{H}(\varepsilon) \Rightarrow L(w) \leq L^* + \varepsilon$

$\check{H}(\varepsilon) \triangleq$ convex lower envelope of $H(\varepsilon)$ (basically shown using Jensen's inequality)

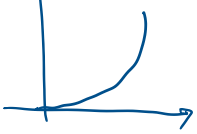
$\check{H}(\varepsilon) \triangleq H^{**}(\varepsilon)$ $\quad \ell^*(z) \triangleq \sup_x x^\top z - \ell(x)$ "Fenchel-Legendre" conjugate

If $\check{H}$ is invertible

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

ℓ is consistent iff $H(\varepsilon) > 0 \ \forall \varepsilon > 0$ (and $H(\varepsilon)$ is finite for some $\varepsilon > 0$)

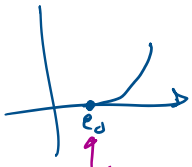
long standard H:



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

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

You want small C; for structured pred. $C \approx |\mathcal{Y}|$ often? (bad)



0 in regret I can guarantee

note: scale of H is arbitrary (scale of ℓ is arbitrary)

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

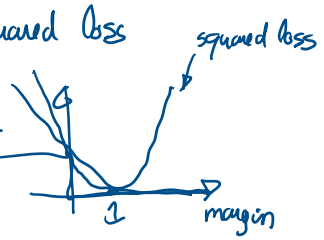
Concrete example: simplest surrogate loss: square loss?

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

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

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

$[1 - y_i \underbrace{\langle w, \varphi(x_i)\rangle}_{s(x_i,+1)}]^2$



$y_i^2(1 - y_i s)^2$

$(y_i - \underbrace{y_i^2}_{+1} s)^2$

does not depend on s

$$\ell_q(s) \triangleq \mathbb{E}_{q(y)}\ell(x,y;s) = \frac{1}{2k}\sum_{\tilde{y}} \mathbb{E}_{q(y)}[s(\tilde{y})^2 + 2s(\tilde{y})l(y,\tilde{y})] + \text{cst.}$$

$$= \frac{1}{2k}\|s + l_{q_x}\|_2^2 + \text{cst.} \qquad l_{q_x}(\tilde{y}) \triangleq \mathbb{E}_{q(y)} l(y,\tilde{y})$$

Suppose s is unconstrained $\min_s \ell_q(s) \Rightarrow s^*(\tilde{y}) = -l_{q_x}(\tilde{y})$

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

ie. you predict optimally pointwise on x

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

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

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

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

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

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

(*) to get consistency for ℓ, it is sufficient to consider $s \in \operatorname{span}(\overleftrightarrow{L})^Y$

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 $\overleftrightarrow{L}$

(row of F as $\varphi(\tilde{y}) \in \mathbb{R}^r$)

$s = F\theta$ $\theta \in \mathbb{R}^r$ (ie. $s(\tilde{y}) = \langle \varphi(\tilde{y}), \theta \rangle$)

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

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

lower bound $\Rightarrow$ easiness result

this is hard

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

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

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

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

hardness result

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

then $H(\varepsilon) \leq \frac{\varepsilon^2}{4k}$ for any loss? ← hardness result

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

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

⇒ we can learn?

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

→ efficient to compute for Hamming loss & (additive) separable score functions