Lecture 8 - calibrated convex surrogate losses

Thursday, February 9, 2023 1:26 PM

today: consistency for convex surrogate losses

non-parametric viewpoint on scores

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

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

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

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

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

$\otimes$: Kronecker product

$v \otimes w \to vw^T$

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

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

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

e.g. $k_2(x,x') = \exp\left(\frac{-\|x-x'\|^2}{2\sigma^2}\right)$ RBF kernel (universal)

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

back to consistency & surrogate losses

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

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] [McAllester 2007] showed that multiclass hinge loss

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

is not consistent for 0-1 loss when have no "majority" label (i.e. $p(\tilde{y}|x) < \frac{1}{2}\ \forall \tilde{y}$, p = true p)

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

↳ exponential sum, → could be intractable for structured prediction

which is consistent for 0-1 loss

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

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

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

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

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

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

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

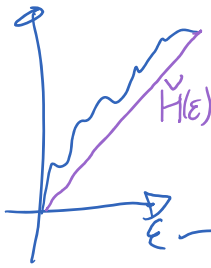
$$L_q(w) \triangleq \mathbb{E}_{q(\tilde{y})}[\ell(\tilde{y}, h_w(x))]$$

"conditional (Vapnik) risk"

[conditional on x version]

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

$\Rightarrow L_q(w) \le L_q^* + \varepsilon$



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

$\mathbb{E}_{(x,y)\sim p}[\ell(x,y;w)]$

→ convex lower envelope of $H(\varepsilon)$

(I basically shown using Jensen's ineq.)

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

if $\check{H}$ is invertible

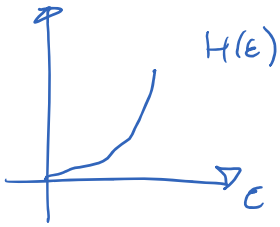
$$L(w) - L^* \le \check{H}^{-1}(\mathcal{L}(w) - \mathcal{L}^*)$$

$\mathcal{L}$ is consistent

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

min regret I can guarantee

standard H:



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

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

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

note: scale of H is arbitrary

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

14h34

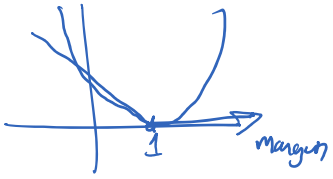
Concrete example: simplest surrogate loss: square loss?

$s(\cdot) \in \mathbb{R}^k$ (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 generalization of squared loss for binary class. to multiclass

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



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

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

does not depend on s

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

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

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

$\arg\max_{\tilde{y}} s^*(\tilde{y}) = \arg\min_{\tilde{y}} \ell q_x(\tilde{y})$

ie. you predict optimally pointwise on x

so here $\mathcal{L}$ is consistent ie. $S^* \in \operatorname{argmin}_{\forall s;\ \mathcal{X} \to \mathbb{R}^K} \mathcal{L}(s)$

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

$$\mathcal{L}_q(s) - \overbrace{\min_{s' \in \mathbb{R}^K} \mathcal{L}_q(s')}^{\mathcal{L}_q^*} = \frac{1}{2K} \| s - (-\ell q_x) \|_2^2$$

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

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

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

(*) 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 $\overleftrightarrow{L}$

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

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

Thm. 7 - if $\operatorname{span}(F) \supseteq \operatorname{span}(\overleftrightarrow{L})$ $\quad H_{\mathcal{L}_{\text{squared}}, \ell, F}(\varepsilon) \geq \frac{\varepsilon^2}{2K \max_{i \neq j} \| P_F \Delta_{ij} \|_2^2} \geq \frac{\varepsilon^2}{4K}$ (lower bound $\Rightarrow$ easiness result; this is bad)

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

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

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

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

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

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

[correct → all these bounds are worst case]

→ Without structure on F, you're doomed in worst case over $p(x,y)$

but for Hamming loss, if add constraints that $s(\tilde{y}) = \sum_{\text{repts}} s(\tilde{y}_p)$ over T binary variables, $H(\epsilon) = \frac{\epsilon^2}{8T}$ } not too big ⇒ we can learn?

(polynomial sample complexity) ☺

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

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