Lecture 6 - scribbles - consistency

Friday, February 9, 2018
14:34

today: consistency & tractible methods

consistency & calibration function

need to relate $g(w)$ to $L(w)$: tool "calibration function" [Steinwart]

relationship is usually very complicated

⇒ current results focus mainly at non-parametric setting (∞ # of parameters)

all functions $h: \mathcal{X} \to \mathcal{Y}$ are considered ; → this evacuates the dependence on x in the analysis

ie. we suppose that $S(x,y;w)$ can be arbitrary for any x (ie. w is ∞-dim) ; → could use a universal kernel $S(\cdot,\cdot;w) \in \mathcal{H}_{\mathcal{X}\times\mathcal{Y}}$

Kernels :

$\langle w, \varphi(x,y)\rangle = S(x,y;w)$ 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 $\varphi(x,y) = \varphi_x(x) \otimes \varphi_y(y)$ ⇕]

Kronecker product

$v \otimes w$

vw^T

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

$= \mathrm{tr}\left(w(v^Tv')w'^T\right)$

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

eg: $\varphi: \mathcal{Y} \to \mathbb{R}^d$ $d = k$ if $|\mathcal{Y}| = k$

$k_y(y,y') = \langle \varphi(y), \varphi(y')\rangle$

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

RBF kernel "universal"

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

$\mathcal{H} = \text{RKHS} = \text{closure}\left(\text{span}\{k(x,\cdot) : x \in \mathcal{X}\}\right)$

"reproducing property"

$\langle k(x,\cdot), h \rangle_{\mathcal{H}} = h(x)$ for $h \in \mathcal{H}$

$h \in \mathcal{H}$ is s.t. $\sum_{i=0}^{\infty} \alpha_i v_i$ (v_i: basis for $\mathcal{H}$)

$\mathcal{H}$ is ℓ_2-dense in $\ell_2 = \{$space of square integrable functions$\}$

$\|f\|_{\ell_2}^2 = \int_x f(x)^2 dx$

$\langle f, g \rangle = \int_x f(x) g(x) dx$

ie. for any $f \in \ell_2$, $\exists$ sequence $h_n \in \mathcal{H}$ s.t. $\|h_n - f\|_{\ell_2} \xrightarrow{n \to \infty} 0$

$$\langle h, h \rangle_{\mathcal{H}} = \left\langle \sum_i \alpha_i k(x_i,\cdot), \sum_i \alpha_i k(x_i,\cdot) \right\rangle$$

$$= \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j)$$

$L(h) = \mathbb{E}_{x,y}\, \ell(y, h(x)) = \mathbb{E}_x \underbrace{\left[\mathbb{E}_{y|x}\, \ell(y, h(x))\right]}_{\ell_x(h(x))}$

$q_x \triangleq p(\cdot|x)$

define $h^*(x) = \underset{\hat{y} \in \mathcal{Y}}{\text{argmin}}\ \ell_{q_x}(\hat{y})$

then $L(h^*) = \min_{\text{all functions } h: \mathcal{X} \to \mathcal{Y}} L(h)$

$\mathbb{E}_{y|x}\, \ell(y, y') \geq \mathbb{E}_{y|x}\, \ell(y, h^*(x)) \quad \forall y'$

$\mathbb{E}_x \mathbb{E}_{y|x}\, \ell(y, y'(x)) \geq \mathbb{E}_x \mathbb{E}_{y|x}\, \ell(y, h^*(x))$

ie. $L(y'(\cdot)) \geq L(h^*) \quad \forall y'(\cdot)$

L for SGD setup; $L(w) = \mathbb{E}_\xi \tilde{L}(w, \xi)$

$\nabla L(w) = \mathbb{E}_\xi \nabla_w \tilde{L}(w, \xi)$

in SGD, you have $\nabla_w \tilde{L}(w, \xi) \neq \nabla_w L(w)$

$\mathbb{E}\left[\nabla_w \tilde{L}(w_t, \xi) \,\middle|\, w_t\right] = \nabla_w L(w_t)$

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

$\nabla \hat{\ell}(w) = \frac{1}{|\text{mini-batch}|} \sum \nabla_w \ell(x, y; w) \rightarrow$ SGD $(x_i, y_i) \sim p$

$\hat{w}_n = \arg\min_w \hat{\ell}(w) + \lambda_n \|w\|^2$ $\quad w_{t+1} = w_t - \gamma_t \nabla_w \ell(x_i, y_i; w_t)$

back to 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 consistent, logistic regression

for multiclass classification [Lee & al. 2004, Mc Allester 2007] showed that multiclass hinge loss $\ell_{\text{hinge}}(x, y; w) = \max_{\tilde{y}} s(x, \tilde{y}; w) + \ell(y, \tilde{y}) - s(x, y)$ is not consistent for 0-1 loss when no majority class (i.e. $p(y|x) < \frac{1}{2} \ \forall y$)

↝ they propose a different surrogate loss that has $\sum_{\tilde{y}}$ instead of $\max_{\tilde{y}}$ which is consistent 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) $p(y|x)$ "noise model" is much richer
2) $\ell(y, y')$ much richer

* [Osokin & al. 2017] → we looked at effect of $\ell(y, y')$ for a convex consistent surrogate loss in the simplest possible setting, and were careful about exponential constants

Calibration function for structured cost ℓ, surrogate ℓ_q^* and set W

$$H_{q,\ell,W}(\varepsilon) \triangleq \inf_{w \in W} \left[\ell_q(w) - \min_{w' \in W} \ell_q(w') \right]$$

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

$q \in \Delta_{|\mathcal{Y}|}$ s.t. $\left[L_q(w) - \min_{w' \in \mathcal{W}} L_q(w')\right] \geq \varepsilon$

$\mathcal{L}_q(w) \triangleq \mathbb{E}_{q(y)}[\ell(x,y,w)]$ — true conditional risks

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

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

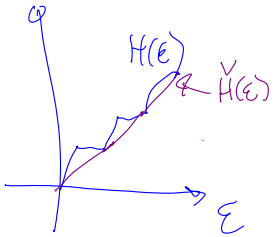
↳ smallest optimization surrogate regret (over all dist. q) s.t. true regret $\geq \varepsilon$

Consequence (conditional on x versions)

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

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

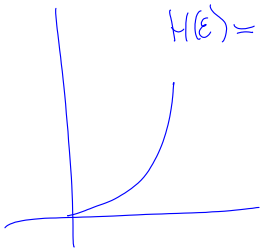
$\mathcal{L}(w) = \mathbb{E}_{x,y}[\ell(x,y,w)]$



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

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

$H(\varepsilon) = \frac{\varepsilon^2}{C}$



if $\check{H}$ is invertible

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

e.g. here $0 \leq L(w) - L^* \leq \sqrt{\mathcal{L}(w) - \mathcal{L}^*}$

[next time: we complete argument for sample complexity using SGD convergence result]