Lecture 8 - scribbles - Calibrated convex surrogate losses
Thursday, January 31, 2019 13:06

Today: Consistency for convex surrogate losses

Non-parametric viewpoint on scores

\[ s(x, y; w) = \langle w, \phi(x, y) \rangle \]

\[ k(x_i; \tilde{y}_j) \]

\[ w = \sum_{i,y} \alpha_i \phi(x_i, \tilde{y}_j) \]

\[ \Rightarrow \langle w, \phi(x, y) = \sum_{i,y} \alpha_i \phi(x_i, \tilde{y}_j) \rangle \]

\[ k(x, x'; y, y') \]

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

"product kernel"

\[ \text{is equivalent to having } \phi(x, y) = \phi_x(x) \otimes \phi_y(y) \]

\[ \text{Kroncker product} \]

\[ \forall w, v \in \mathbb{R}^d \]

\[ \langle v \otimes w, v' \otimes w' \rangle = \text{tr} \left( (v w^T) (v' w'^T) \right) \]

\[ = \text{tr} \left( w v^T v'^T w'^T \right) \]

\[ = \langle w, w' \rangle \langle v, v' \rangle \]

\[ \text{e.g. } k_x(x, x') = \exp \left( \frac{-||x - x'||^2}{2\sigma^2} \right) \]
Back to consistency & surrogate losses

$$\hat{w}_n \leq \arg \min_w \frac{\hat{L}_n(w) + \lambda_n \|w\|^2}{2}$$

consistency: $$L(\hat{w}) \rightarrow \min_w L(w)$$

---

**Binary classification** [Fanelli et al. 2004] characterized a whole family of consistent surrogate losses

$$L \rightarrow \text{binary sum logistic regression}$$

---

**Multiclass classification** [Lee et al. 2004] showed that multiclass hinge loss

$$L_{\text{multiclass hinge loss}} = \max_{y \neq \hat{y}} s(x, y; w) + 0(y = \hat{y})$$

is not consistent for 0-1 loss when there is no majority class (i.e., $$\frac{1}{Y}$$)

They propose a different surrogate loss that

$$\rightarrow \text{exponential sum}$$

which is consistent

$$\rightarrow \text{could be intractable}$$
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) \( p(y, x) \) much richer

\[ \text{Calibration function for a structured loss } \ell \text{, surrogate loss } \ellq \text{ and set } \hat{\ell} \]

\( H_{\ell, \ellq} \) \( \overset{\Delta_{151}}{=} \inf_{w \in W} \left[ \ellq(w) - \min_{w' \in W} \ellq(w') \right] \)

s.t. \( \ellq(w) - \min_{w' \in W} \ellq(w') \geq 3 \)

\( Lq(w) \approx E_{q(y|x)} [\ell(x, y; w)] \)

\( \ellq(w) \approx E_{q(y|x)} [\ellq(x, y; w)] \)

\( \text{(conditional risk)} \)

\( \text{smallest optimization surrogate regret (over all dist. q)} \)

s.t. true regret \( \geq \varepsilon \)

[conditional on \( x \) version]
\[ V_q : S_q(w) < S_q^* + H(\varepsilon) \Rightarrow L_q(w) < L_q^* + \varepsilon \]

\[ \text{(thm. 2) } \forall \varepsilon : S(w) < S^* + H(\varepsilon) \Rightarrow L(w) < L^* + \varepsilon \]

\[ \forall_{x,y} S(x,y,w) \]

\[ \text{(convex lower envelope of } H(\varepsilon) \text{)} \]

\[ H(\varepsilon) = H^{\varepsilon}(\varepsilon) \]

\[ \overline{S}(\varepsilon) = \sup_{x,y} x^T z - f(z) \quad \text{"Fenchel-Legendre" conjugate} \]

\[ \overline{S} \text{ is consistent if } H(\varepsilon) \geq 0 \quad \forall \varepsilon > 0 \]

\[ \text{and } H(\varepsilon) \text{ is finite for some } \varepsilon > 0 \]

\[ \text{If } H \text{ is univex} \]

\[ L(w) - L^* = H^{\varepsilon}(S(w) - S^*) \]

\[ H(\varepsilon) = \frac{\varepsilon^2}{C} \Rightarrow L(w) - L^* < \sqrt{C(S(w) - S^*)} \]

\[ \text{you want small } C \text{ for structured prediction } C \approx 127 \text{ often} \]

\[ \text{Note: scale of } H \text{ is arbitrary; normalize it using optimization perspective (e.g. SGD)} \]
The simplest surrogate loss: square loss.

\[ s(\cdot) \in \mathbb{R}^k \quad (s(x)) \]

\[ g(x, y; s) \triangleq \frac{1}{2k} \| s - (s(y)) \|^2 = \frac{1}{2k} \left[ s(x, y) + b(y) \right]^2 
\]

\[ g_q(s) = \mathbb{E}_{q(y)} g(x, y; s) \]

\[ = \frac{1}{2k} \mathbb{E}_{q(y)} \left[ s(y)^2 + a s(y) b(y) + \text{constant} \right] \]

\[ = \mathbb{E}_{q(y)} b(y) = l_q(x, y) \]

Suppose \( s \) is unconstrained.

\[ \min_s g_q(s) \quad s(y) + l_q(x, y) = 0 \quad \forall y \]

\[ \Rightarrow s^*(y) = -l_q(x, y) \]

\[ \text{argmax} \quad s^*(y) = \text{argmin} \quad l_q(x, y) \]

i.e. you predict optimally (pointwise)

So here \( g \) is consistent.

i.e. \( s^* \in \text{argmin}_s g(s) \Rightarrow L(h(s)) = \min_h L(h) \)
\[ s_q(s) = \| s - \langle -q \rangle \|_2^2 + \text{cost} \]

\[ s_q(s) - \min_{s' \in \mathbb{R}^k} s_q(s') = \frac{1}{2\kappa} \| s - \langle -q \rangle \|_2^2 \]

Let \( L \) be a \( k \times k \) matrix where \( \langle y, y \rangle = \langle y, \tilde{y} \rangle \)

\[ \ell_q = \langle L \tilde{q} \rangle \]

\[ s^* = -L \tilde{q} = -L q_x \in \text{span}(L) \quad \text{i.e.} \quad \exists \tilde{y} \quad \langle \tilde{y}, \tilde{y} \rangle \]

To get consistency for \( L \), it is sufficient to consider \( s \in \text{span}(L) \) or that \( s \in \text{span}(F) \supseteq \text{span}(L^*) \)

\( F \in \mathbb{R}^{k \times r} \) matrix

Can be chosen carefully depending on \( L^* \)

\[ s = F \Theta \quad \Theta \in \mathbb{R}^r \]

\[ s_q(\Theta) - \min_{\Theta \in \mathbb{R}^r} s_q(\Theta) = \frac{1}{2\kappa} \| F \Theta - \langle L \tilde{q} \rangle \|_2^2 \]

\[ \min_{\Theta} \quad \text{cost} \]

Cover bound \Rightarrow easiness result
\[ \text{thm. 7} \quad \text{if } \text{span}(\mathcal{E}) = \text{span}(\mathcal{L}) \]

\[ H(\mathcal{E}) = \frac{\varepsilon^2}{2K \max_{i \neq j} \| P = A_{ij} \|^2} \geq \frac{\varepsilon^2}{4K} \]

\[ \Delta_{ij} = \mathbf{e}_i - \mathbf{e}_j \in \mathbb{R}^K \]

\[ P = \text{orthogonal projection on span}(\mathcal{F}) \quad P = F (F^T F)^{-1} F^T \]

- In the paper, we show that for 0-1 loss, \[ H(\varepsilon) = \frac{\varepsilon^2}{4K} \]

\[ \text{thm. 8} : \text{if } \text{span}(\mathcal{F}) = 1 \mathbb{R}^K \text{ (i.e. no constraints)} \quad \text{hardness result} \]

- If span(\mathcal{F}) = \mathbb{R}^K, then \[ H(\varepsilon) \geq \frac{\varepsilon^2}{2K} \text{ for any loss} \]

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

- Cautious: all these are heuristics.

\[ \mathbb{O} \quad \text{but for hamming loss, if add constraints that } s(\mathbf{y}) = \mathbf{1} \quad \text{for any loss} \]

\[ H(\varepsilon) = \frac{\varepsilon^2}{8T} \quad \text{not too big} \quad \Rightarrow \text{we can learn?} \]
optimization formulation

setup  let \( s((x, y), \gamma) \) be of the form \( F(x) \Theta(z) \in \mathbb{R}^r \)

\[ \Theta(z) \in \mathcal{H} \quad \text{rk}(z) \]  

optimization variables

\[ s((x, y)) \sum_{(x, y, \gamma)} \]

\[ s((x, y)) \sum_{(x, y, \gamma)} \]

run projected SGD on \( s((x, y)) \) i.e. \[ \Theta^{(n)} = \Theta^{(0)} - \eta \nabla_{\Theta} s((x, y), \Theta^{(n)})) \]

a ball of radius \( D \)

\[ s((x, y)) \sum_{(x, y, \gamma)} \]

\[ s((x, y)) \sum_{(x, y, \gamma)} \]

Convergence result: if \[ \| \Theta^{(n)} \|_{\mathcal{H}} D \]

If \[ \| \Theta^{(n)} \|_{\mathcal{H}} D \]

then averaged SGD with stepsize \( \eta = \frac{2D}{M} \)

\[ \mathbb{E} \left[ s((x, y), \gamma)) - s((x, y)) \right] \leq 2D \sqrt{n} \]

(Convergence occurs)
thm. 6 Learning complexity (to be filled)?