Lecture 18 - scribbles - consistency

Wednesday, March 22, 2017
10:40

today:
- consistency
- CRF objective and variance reduced SGD

Consistency & calibration function

need to relate $\delta(w)$ to $L(w)$ → "calibration function" [Steinwart]

relationship is usually very complicated;
⇒ current results look mainly at non-parametric setting (where dependence on x is evacuated (?))

→ we suppose that $s(x,y;w)$ can be arbitrary for any x (i.e. w is ∞-dim)

[can use a universal kernel to formalize this and re-include x]

* binary classification, Bartlett et al. characterized a whole family of consistent surrogate losses

for multiclass: [Lee & al. 2004] [McAllester 2007] showed that multiclass hinge loss is not consistent for 0-1 loss when no majority class (i.e. $p(y|x) \leq \frac{1}{2}\ \forall y$)

2 aspects of structured prediction which give a much richer theory than for binary classification for consistency:

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

→ proposed a fix with a surrogate loss that has $\sum_{\tilde{y}}$ instead of $\max_{\tilde{y}}$

exponential # ⇒ might cause problem?

[Osokin & al. 2017] → looked at this exponential aspect more carefully, in the simplest setup

Calibration function for structured cost ℓ, surrogate loss ψ, and set W

$$H_{\psi,\ell,W}(\varepsilon) \triangleq \inf_{w\in W,\ q\in\Delta_{|\mathcal{Y}|}} \psi_q(w) - \min_{w'\in W}\psi_q(w') \quad \text{s.t.} \quad L_q(w) - \min_{w'\in W} L_q(w') \geq \varepsilon$$

(ψ_q^* / L_q^*)

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

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

Smallest optimization surrogate regret possible (over all dist. q) s.t. true regret is $\geq\varepsilon$

consequence (Thm. 2)

$$\forall q,\ \psi_q(w) < \psi_q^* + \check{H}(\varepsilon) \Rightarrow L_q(w) \leq L_q^* + \varepsilon$$

$\check{H}$: convex envelope of $H(\varepsilon)$ in $\mathbb{R}^{+*}$

If $\check{H}$ is invertible (note: it is increasing as Gauthier pointed out)

$$L_q(w) - L_q^* \leq \check{H}^{-1}(\psi_q(w) - \psi_q^*)$$

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

Sample complexity:

can link learning with optimization using SGD: while running SGD with $y^{(i)} \sim q$ we are optimizing population ψ_q (instead of ψ)

thus from SGD convergence on ψ_q can see how many iterations needed

thus from SGD convergence on ℓ_q, can see how many iterations needed (= # of samples) to get surrogate-regret of $H(\varepsilon)$

$\Rightarrow$ translates to ε-true regret

e.g. constant step-size SGD projected on a ball of diameter D

gives $\mathbb{E}[\ell_q(w_t)] - \ell_q^* \le \frac{DM}{\sqrt{t}}$ after t iterations $\quad M^2 \ge \mathbb{E}_{qy}[\|\nabla_w \ell(x,y;w)\|_2^2]$

(thm 4): to get $\mathbb{E}[L_q(w_t)] - L_q^* \le \varepsilon$

need at most $t \ge \left(\frac{DM}{H(\varepsilon)}\right)^2$

gives (sample complexity)

in our paper, consider the consistent convex surrogate

$$\ell(y;w) \triangleq \frac{1}{|\mathcal{Y}|} \sum_{\tilde{y}} \frac{1}{2}\left(s_w(\tilde{y}) + l(y,\tilde{y})\right)^2$$

$$\ell_q(w) = \mathbb{E}_q[\ell(y;w)] = \frac{1}{|\mathcal{Y}|} \sum_{\tilde{y}} \frac{1}{2}\left(s_w(\tilde{y}) + \underbrace{\mathbb{E}_q[l(y,\tilde{y})]}_{l_q(\tilde{y})}\right)^2 + \text{const.}$$

so w^* is s.t. $s_{w^*}(\tilde{y}) = -l_q(\tilde{y})$

max score $\Rightarrow$ min conditional risk and thus consistent

ℓ is consistent for any l and q as long as $S_W \triangleq \{s_w(\cdot) : w \in W\} \supseteq \text{span}\{l(y,\cdot) : y \in \mathcal{Y}\}$

$\Rightarrow$ in paper, we show that if no constraint on $s_w(\cdot)$

then $H(\varepsilon) \asymp \varepsilon^2$ for any l

$2^{|\mathcal{Y}|}$ } need exponential accuracy ⇒ exponential sample complexity?

also, for 0-1 loss, to be consistent, we basically need no constraint on $S_w(\cdot)$

but for Hamming loss; if add constraint that $S(\tilde{y}) = \sum_p s_p(\tilde{y}_p)$

over T binary variables then $H(\varepsilon) = \varepsilon^2$

$\varepsilon \to 0$ } not too big ⇒ can learn?

(*) Moral here:

- some losses are harder to learn than others (0-1 difficult in general)
- have linked computation to statistical performance in consistency framework
  ↳ (convex surrogate loss)

  but lost dependence on x (need to use SGD with RKHS (kernel stuff) to fill in details)

  → different approach than gen. error bound; but gives insights
- still need more theory?

- main pointer covered today:
  - Anton Osokin, Francis Bach, Simon Lacoste-Julien
    On Structured Prediction Theory with Calibrated Convex Surrogate Losses
    https://arxiv.org/abs/1703.02403
- other pointers:
  - canonical paper which presented consistency analysis for binary classification:
    Bartlett, Peter L., Jordan, Michael I., and McAuliffe, Jon D
    Convexity, classification, and risk bound
    Journal of the American Statistical Association, 101(473):138–156, 2006.
  - paper which showed that multiclass SVM is not consistent (for the 0-1 loss) and proposed a consistent alternative:
    Lee, Yoonkyung, Lin, Yi, and Wahba, Grace.
    Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data.
    Journal of the American Statistical Association, 99(465):67–81, 2004.
    - see also the McAllester 2007 paper:

Generalization Bounds and Consistency for Structured Labeling in Predicting Structured Data, edited by G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. V. N. Vishwanathan. MIT Press, 2007
http://nagoya.uchicago.edu/~dmcallester/colbounds.pdf

- and interestingly, this recent paper shows that the multiclass SVM *is* consistent for a loss on 3 classes with an "abstain" notion:
  Ramaswamy, Harish G. and Agarwal, Shivani.
  Convex calibration dimension for multiclass loss matrices.
  JMLR, 17(14):1–45, 2016.
- see also the extensive related work section of the arxiv 2017 paper Osokin et al.