Lecture 6 - generalization error bounds

Thursday, February 2, 2023 1:24 PM

today: • review surrogate losses
• generalization error bounds & structured SVM

Review of convex surrogate losses mentioned so far:

$$\ell_{\text{perceptron}}(x, y; w) = \max_{\tilde{y} \in \mathcal{Y}} s(\tilde{y}) - s(y) \qquad \text{let } m(\tilde{y}) \triangleq s(y) - s(\tilde{y})$$

$$= \max_{\tilde{y} \in \mathcal{Y}} [-m(\tilde{y})] \quad \left(= \left[\max_{\tilde{y} \neq y} -m(\tilde{y})\right]_+ \right) \text{ if } y \in \mathcal{Y}$$

$$\ell_{\text{hinge}}(\quad) = \max_{\tilde{y}} [s(\tilde{y}) + \ell(y, \tilde{y})] - s(y) \quad \text{(structured SVM)}$$

"margin rescaling" $= \max_{\tilde{y} \in \mathcal{Y}} [\ell(y, \tilde{y}) - m(\tilde{y})]$

vs.

"slack rescaling" $= \max_{\tilde{y}} \ell(y, \tilde{y})[1 - m(\tilde{y})]$

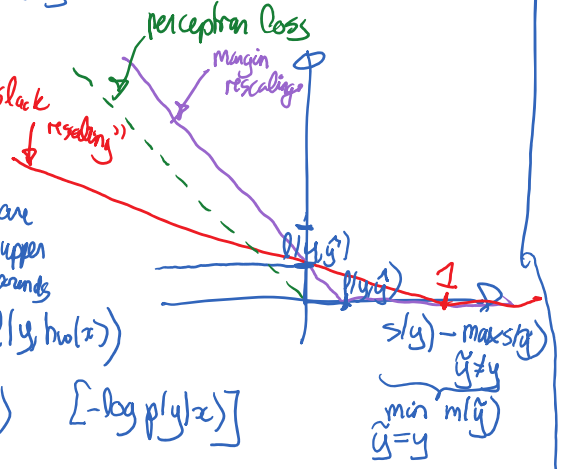
} are upper bounds $\ell(y, h_{w0}(x))$

$$\ell_{\text{CRF}}(\quad) = \frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta s(\tilde{y}))\right) - s(y) \qquad [-\log p(y|x)]$$

$\beta \to \infty \Rightarrow$ perceptron loss

$\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(-\beta m(\tilde{y}))\right)$ suggests "smoothed hinge loss"

$\frac{1}{\beta} \log\left(\sum_{\tilde{y}} \exp(\beta(\ell(y,\tilde{y}) - m(\tilde{y})))\right)$ [e.g. Pletscher et al. 2010]



note: slack rescaling more robust when have small $\ell(y,\tilde{y})$ [e.g. 0]

but more computationally costly

What are theoretical properties could we look at?

a) generalization error bounds [today]

b) consistency properties & calibration function [next(2) class]

↳ relationship between $L(w)$ & $\ell(w)$

Why structured score functions?

$$s(x, y) = \sum_{c \in \mathcal{C}} s_c(x, y_c)$$

motivation similar to graphical models

1) statistical efficiency : less # of parameters (simpler score fct. S)
$\Rightarrow$ easier to learn (generalization guarantees) [see Cortes & al. NeurIPS 2016] (today)

2) computational " : Compute $\operatorname{argmax}_{y \in \mathcal{Y}} S(y)$

generalization error bounds :

for binary classification

a classical PAC bound is :

for any fixed dist. p on data with prob. $\geq 1-\delta$ on D_n

$$\forall w \in W \quad L_{01}(w) \leq \hat{L}_n(w) + \frac{1}{\sqrt{n}} \sqrt{d \log \frac{d}{n} + \log \frac{1}{\delta}}$$

where d is VC-dimension of $\mathcal{H} \triangleq \{h_w : w \in W\}$

VC-dimension of $\mathcal{H} \triangleq \max \{ m : \exists$ a set of m points s.t. $\forall$ labelings of these points $\exists w$ s.t. h_w gives the correct label on these points $\}$
"shattering the set of points"

↳ # of binary functions on m points is 2^m

for $\mathcal{H} = \{$ linear classifiers of p parameters $\}$ VC-dim $(\mathcal{H}) = p+1$

* one issue for this is that it's true for all distributions $\Rightarrow$ too loose bound

$\Rightarrow$ motivates going to data distribution dependent measure of complexity

example : empirical Rademacher complexity

$$\hat{R}_{D_n}(\mathcal{H}) \triangleq \mathbb{E}_\sigma \left[\sup_{h_w \in \mathcal{H}} \left| \frac{2}{n} \sum_{i=1}^{n} \sigma_i \mathbb{1}\{y_i \neq h_w(x_i)\} \right| \right]$$

"correlation with random noise"

$\sigma_i = \begin{cases} +1 \\ -1 \end{cases}$ uniformly "Rademacher R.V."

bound : with prob $\geq 1-\delta$

$$\forall w \quad L_{01}(w) \leq \hat{L}_n(w) + \hat{R}_{D_n}(\mathcal{H}) + \frac{1}{\sqrt{n}} 3 \sqrt{\log \frac{2}{\delta}}$$

$$\forall w \quad L_{01}(w) \le \hat{L}_n(w) + \hat{R}_{D_n}(H) + \frac{1}{\sqrt{n}} 3\sqrt{\log \frac{2/\delta}{2}}$$

↳ complexity depends on D_n (implicitly on P)

14h30

Structured prediction generalization bounds [Cortes & al. NeurIPS 2016]

general loss fct. $\ell(y,y')$ s.t. $\ell(y,y') \neq 0 \ \forall y \neq y'$

suppose $s(x,y) = \sum_{c \in \mathcal{C}} s_c(x, y_c)$

↳ set of cliques of a graph model G / factor graph

thm. 7 with prob $\ge 1-\delta$

$$\forall w \in W \quad L(w) \le \hat{\ell}_{hinge}(w) + 4\sqrt{2}\, \hat{R}^{\mathcal{G}}_{D_n}(H_w) + 3 \frac{L_{\max}}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}}$$

($L_{\max} = \max_{y,y'} \ell(y,y')$)

where $\hat{R}^{\mathcal{G}}_{D_n} \triangleq \frac{1}{n} \mathbb{E}_\sigma \left[\sup_{w \in W} \sum_{i=1}^n \sqrt{|\mathcal{C}_i|} \sum_{c \in \mathcal{C}_i} \sum_{\tilde{y}_c \in Y_c} \sigma_{i,c,\tilde{y}_c} s_c(x_i, \tilde{y}_c; w) \right]$

- actually only depends on $(x^{(i)})_{i=1}^n$
- "empirical factor graph complexity"
- $\sigma_{i,c,\tilde{y}_c}$: indep Rademacher R.V.

thm. 2: if $s_c(x, y_c; w) = \langle w, \psi_c(x, y_c) \rangle$

and consider $W_\Lambda \triangleq \{ w : \|w\|_2 \le \Lambda \}$; let $R = \max_{i,c,y_c} \|\psi_c(x_i, y_c)\|_2$

then $$\hat{R}^{\mathcal{G}}_{D_n}(H_{W_\Lambda}) \le \frac{R \Lambda}{\sqrt{n}} \max_i |\mathcal{C}_i| \sqrt{\max_{c \in \mathcal{C}_i} |Y_c|}$$

So want small cliques?

* plug thm. 2 in thm. 7:

$$L(w) \le \hat{\ell}_{hinge}(w) + \underbrace{\left(\frac{R |\mathcal{C}| \sqrt{\max_c |Y_c|}}{\sqrt{n}} \right)}_{\frac{\lambda_n}{2}} \underbrace{\Lambda}_{\|w\|_2} + \text{cst.}$$

min of RHS suggests

SVM struct alg. $\hat{w} \in \operatorname{argmin}_w \hat{\ell}_{hinge}(w) + \frac{\lambda_n}{2} \|w\|_2^2$

missing link: ① $\min_{\text{s.t. } \|w\|_2 \le \Lambda} f(w)$ (if f is convex) } use Lagrangian

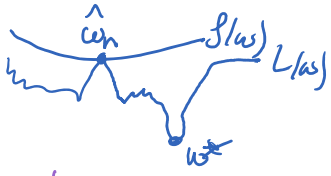
missing link: ① $\min f(w)$ s.t. $\|w\|_2 \leq \Lambda$

② $\min f(w) + \frac{\lambda}{2}\|w\|^2$

(if f is convex) use Lagrangian ⇒ $\exists$ a $\lambda(\Lambda)$ s.t. solution to ② same solution to ①

[sidenote: constrained formulation can have solutions not achievable for ② when f is non-convex but penalized/reg. formulation is less sensitive to choice of λ vs. constrained formulation]

properties: • minimize upper bound, hope that min $L(w)$ but no general guarantees



→ caveat: also note here: no consistency guarantee

• can evaluate bound to get guarantees

next: consistency + convex surrogate

Consistency & calibration

need to relate $f(w)$ to $L(w)$: bad "calibration fct." [Steinwart]

relationship is usually very complicated

⇒ usual results look mainly at non-parametric setting (∞ # of parameters)

all functions $h: \mathcal{X} \to \mathcal{Y}$ are considered ⇒ this evacuates the dependence on x of the analysis "pointwise analysis"

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

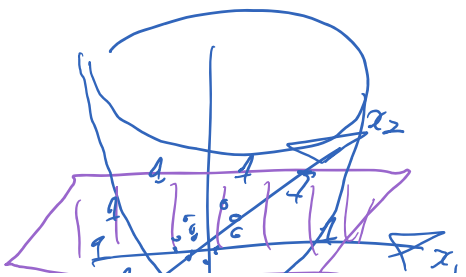
(i.e. w is ∞-dim.)

→ can do this using a universal kernel

$s(\cdot, \cdot, w) \in \mathcal{H}_{\mathcal{X} \times \mathcal{Y}}$

RKHS:

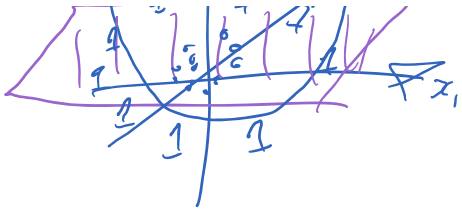
motivation:



generalize linear structure $\langle w, \varphi(x) \rangle$ to higher dim.

+ kernel trick $\langle \varphi(x), \varphi(x') \rangle = k(x, x')$

$\Phi : \mathcal{X} \to \mathbb{R}^3$



$\Phi : \mathcal{X} \to \mathbb{R}^3$

$\Phi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}$ ($x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$)

$\langle \Phi(x), \Phi(x') \rangle_{\mathbb{R}^3} = \left(\langle x, x' \rangle_{\mathbb{R}^2} \right)^2 = k(x, x')$

polynomial kernel e.g. $(\langle x, x' \rangle + 1)^p = k_{poly}(x, x')$

equivalent to mapping data to a space of dimension exponential in p

$\langle \Phi(x), \Phi(x') \rangle$

even have ∞-dim e.g. $k(x, x') = \exp\left(-\|x - x'\|^2 \right)$

"RBF kernel"