Lecture 6 - generalization error bounds

Tuesday, February 2, 2021 14:32

today: generalization error bounds
& structured SVM

generalization error bounds:

for binary classification

a classical PAC bound is:

$$\boxed{\begin{array}{l}\text{for any fixed dist } p \text{ on data} \\ \quad \text{with prob.} \geq 1-\delta \text{ on } D_n \\ \forall w \in \mathcal{W} \quad L_{01}(w) \leq \hat{L}_n(w) + \frac{1}{\sqrt{n}}\sqrt{d \log \frac{d}{n} + \log \frac{1}{\delta}} \\ \quad \text{where } d \text{ is VC-dimension of } \mathcal{H} = \{h_w : w \in \mathcal{W}\}\end{array}}$$

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

"shattering the set of points"

# of prediction functions on m pts is 2^m

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

* one issue for this bound is that it's true for all distributions => too loose bound

=> motivates going to data distribution dependent measure of complexity

example: empirical Rademacher complexity

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

"correlation with random noise"

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

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

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

complexity depends on D_n (implicitly on P) complexity depends on D_n (implicitly on P)

<u>structured prediction generalization bounds</u> [Cortes et al. NIPS 2016]

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

suppose $s(x,y) = \sum_{c \in \mathcal{C}_n} s_c(x, y_c)$ → set of cliques of a graph. model G / factor graph

<u>thm. 7</u> with prob. $\geq 1-\delta$

depends on $\ell(y,y')$ — $\max_{y,y'} \ell(y,y')$

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

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

actually only depends on $(x^{(i)})_{i=1}^n$

"empirical factor graph complexity"

indep Rademacher R.V.

<u>thm 2</u>: if $s_c(x, y_c; w) = \langle w, \varphi_c(x, y_c) \rangle$

and consider $W_\Lambda \triangleq \{ w : \|w\|_2 \leq \Lambda \}$; let $R = \max_{i,c,\tilde{y}} \|\varphi_c(x_i, \tilde{y})\|_2$

$$\text{then} \quad \boxed{\hat{R}^{G}_{D_n}(\mathcal{H}_{W_\Lambda}) \leq \frac{R\Lambda}{\sqrt{n}} |\mathcal{C}| \sqrt{\max_c |\mathcal{Y}_c|}}$$

so want <u>small cliques</u>!

* plug thm. 2 back in thm. 7

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

min of RHS <u>suggests</u>

SVM struct alg. $\hat{w}_n = \arg\min_w \hat{\mathcal{L}}_{\text{hinge}}(w) + \frac{\lambda_n}{2} \|w\|_2^2$

<u>missing link</u>: ① $\min f(w)$ s.t. $\|w\| \leq \Lambda$ (if f is convex) use Lagrangian duality $\exists$ a $\lambda(\Lambda)$ s.t.

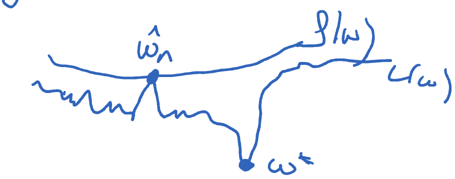
② $\min f(w) + \frac{\lambda}{2} \|w\|^2$ sol'n to (2) gives same solution as ①

② $\min f(w) + \frac{\lambda}{2}\|w\|^2$ ... gives same solution as 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]

can think of SVMstruct as minimizing an upper bound on gen. error

properties: • minimize upper bound, hope that $\min L(w)$



but no general guarantees

caveat: also note here: no consistency guarantees

• can evaluate bound to get guarantees

next: consistency + convex surrogate

15h40

consistency & calibration

need to relate $S(w)$ to $L(w)$: tool 'calibration fct. [Steinwart]

relationship is usually very complicated

$\Rightarrow$ current results look mainly at non-parametric setting (∞ # of parameters)

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

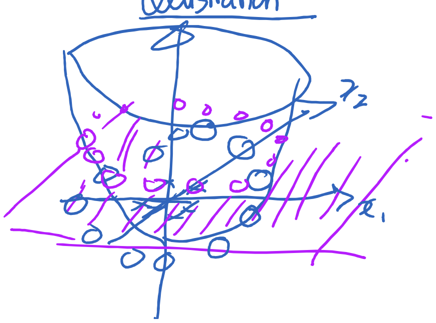
i.e. we suppose that $s(x, y; w)$ can be arbitrary for any x
(i.e. w is ∞-dim)
$\to$ can do this using a universal kernel

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

RKHS:

motivation: generalize linear structure

illustration

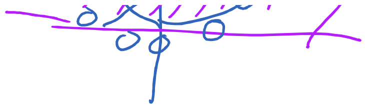


$\langle w, \varphi(x) \rangle$ to higher dim. space

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

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

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



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

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

polynomial kernel e.g. $(\langle x,x'\rangle + 1)^p = k(x,x')$

equivalent to mapping data to a space of dimension <u>exponential in p</u>, $\langle \Phi(x), \Phi(x')\rangle$

even have ∞-dim, e.g. $k(x,x') = \exp\left(-\frac{\|x-x'\|_2^2}{2}\right)$

"RBF kernel"

<u>RKHS</u> (reproducing kernel Hilbert space)

$\Phi : \mathcal{X} \to \mathcal{H}$ s.t. $\langle \Phi(x), \Phi(x') \rangle_{\mathcal{H}} = k(x,x')$ (important property of RKHS)

$\mathcal{H}$: RKHS

$\mathcal{H}$ is a space of fct. $\mathcal{X} \to \mathbb{R}$

Let $\tilde{\mathcal{H}} = \text{span}\{k(x,\cdot) : x \in \mathcal{X}\}$

e.g. $f \in \tilde{\mathcal{H}} \Rightarrow f = \sum_i \alpha_i^f k(x_i^f, \cdot)$ for some finite $\{x_i\}_{i=1}^{n(f)}$, $\alpha_i \in \mathbb{R}$

$\boxed{\text{"pre-Hilbert" space [inner product space] with } \langle f, g\rangle_{\tilde{\mathcal{H}}} \triangleq \sum_{i,j} \alpha_i^f \alpha_j^g \underbrace{k(x_i^f, x_j^g)}_{\langle k(x_i^f,\cdot), k(x_j^g,\cdot)\rangle_{\tilde{\mathcal{H}}}}}$

α_i^f means a_i in $f = \sum_i a_i k(a_i, \cdot)$

$\|f\|_{\tilde{\mathcal{H}}} \triangleq \sqrt{\langle f, f\rangle_{\tilde{\mathcal{H}}}}$

then RKHS $\mathcal{H}$ is $=$ completion$(\tilde{\mathcal{H}})$ using $\|\cdot\|_{\tilde{\mathcal{H}}}$ as your norm

i.e. add all limit points of $\tilde{\mathcal{H}}$-Cauchy sequences to get $\mathcal{H}$

you could think $f = \sum_{i=1}^{\infty} a_i k(x_i, \cdot)$

$\boxed{\text{"reproducing" property of } \mathcal{H}: \text{ for } f \in \mathcal{H} \quad \langle f, k(x,\cdot)\rangle_{\mathcal{H}} = f(x)}$ (*)

$k(x,\cdot) = \Phi(x)$

nice property of RKHS, fct. evaluation is a <u>cts.</u> operation

mapping $E_x : \mathcal{H} \to \mathbb{R}$

$$E_x(f) = f(x)$$

$$|f(x) - g(x)| \leq |\langle f - g, k(x,\cdot)\rangle_{\mathcal{H}}|$$

$$\leq \|f-g\|_{\mathcal{H}} \, \|k(x,\cdot)\|_{\mathcal{H}}$$

i.e. E_x is Lipschitz cts with $L = \|k(x,\cdot)\|_{\mathcal{H}}$

(*) this property is important to do statistics