Lecture 6 - generalization error bounds

Thursday, January 23, 2020 13:32

Today: generalization error bounds
↳ Structured SVM

~~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
>
> $$\forall w \in \mathcal{W} \quad L_{01}(w) \leq \hat{L}_n(w) + \frac{1}{\sqrt{n}} \sqrt{d \log \frac{d}{n} + \log \frac{2}{\delta}}$$
>
> where d is VC-dimension of $\mathcal{H} = \{h_w : w \in \mathcal{W}\}$

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

↓

# 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 true for all distributions ⇒ too loose bound

⇒ motivates going to data distribution dependent measure of complexity

example: empirical Rademacher complexity

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

"correlations with random noise"

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

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

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

$$\forall w \quad L_{0,1}(w) \le L_n(w) + R_{D_n}(H) + \frac{1}{\sqrt{n}} 3\sqrt{\frac{\log \frac{2}{\delta}}{2}}$$

complexity depends D_n (implicitly on P)

high level idea to prove bound:

"double sample trick" → use a second sample D_n' for gen. error — $L(w) = \mathbb{E}_{D_n'}[\hat{L}_n'(w)]$

+

"symmetrization trick" → bound sup of differences between $\mathbb{E}_{D_n'}[\hat{L}_n'(w)]$ and $\hat{L}_n(w)$

+

union bound as usual + concentration inequality

structured prediction generalization bounds [Cortes & al. NIPS 2016]

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

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

↳ set of cliques of a graph. model CO / factor graph

thm. 7 with prob. $\ge 1-\delta$ (depends on $\ell(y,y')$)

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

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

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

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

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

then $$\hat{R}^{\mathcal{G}}_{D_n}(H_{\mathcal{W}_\Lambda}) \le \frac{R\Lambda}{\sqrt{n}} |\mathcal{C}| \sqrt{\max_c |\mathcal{Y}_c|}$$

so want small cliques!

* plug thm. 2 back in thm. 7:

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

min of RHS suggests

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

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

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

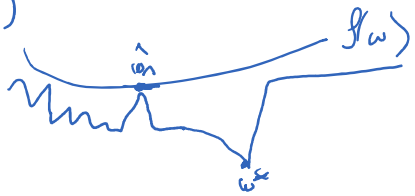
(If f is convex) → use Lagrangian duality $\exists a\ \lambda(\Lambda)$ s.t. s'ln of ② gives same solution as ①

[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 SVM struct as minimizing an upper bound on gen. error

Properties:
- minimize upper bound, hope that min. $L(w)$ but no general guarantees



- can evaluate bound to get guarantees

caveat: also note here: no consistency guarantee

next: consistency

HHH

Consistency & calibration

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

relationship is usually very complicated

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

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

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

→ can do this using a universal kernel

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

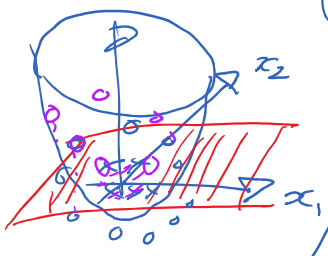
RKHS

motivation:

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

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

illustration: (not to scale :))



formal example of kernel trick

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

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

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

polynomial kernel eg $(\langle x,x'\rangle + 1)^d \triangleq k(x,x')$

equivalent to mapping data to a space of dimension exponential in d $\langle \Phi(x), \Phi(x')\rangle$

even have ∞-dim, e.g. $k(x,x') = \exp\left(-\frac{1}{2}\|x-x'\|^2\right)$ (RBF kernel)

RKHS (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)

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

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

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

→ "pre-Hilbert" space [inner product space]

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

with $\langle f, g\rangle_{\tilde{\mathcal{H}}} \triangleq \sum_{i,j} \alpha_i^f \alpha_j^g \underbrace{k(x_i^f, x_j^g)}_{\langle k(x_i^f,\cdot), k(x_j^g,\cdot)\rangle_{\mathcal{H}}}$

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

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

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

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

* "reproducing" property of $\mathcal{H}$: for $f \in \mathcal{H}$

$$\langle f, k(x, \cdot) \rangle_{\mathcal{H}} = f(x)$$

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

nice property of RKHS, fct. evaluation is a cts. operation

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

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

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

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

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

(*) this property is important to do statistics

$f : \mathcal{X} \to \mathbb{R}$

$S : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$

or

$\mathcal{X} \to \mathbb{R}^k$