Lecture 6 - generalization error bounds

Thursday, January 25, 2024 9:30 AM

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

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

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

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") $\}$

$h : \mathbb{R}^m \to \{-1, 1\}$

note: # of prediction functions on n points is 2^n

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

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

"correlation with random noise"

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

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

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

complexity depends on D_n (implicitly depends on P)

structured prediction generalization bounds [Cortes & al NeurIPS 2016]

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

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

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

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

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

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

"empirical factor graph complexity" — σ_{i,c,y_c} indep Rademacher R.V.

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

thm 2: 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 $R = \max_{i,c,y} \|\varphi_c(x_i, y_c)\|_2$

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

so want small cliques !

* plug thm. 2 back in thm. 7

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

min of R.H.S suggests

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

missing link: ① $\min f(w)$ s.t. $\|w\|_2 \leq \Lambda$, $\|w\|_2^2 \leq \Lambda^2$

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

(if f is convex) use Lagrangian duality $\exists\, a\ \lambda(\Lambda)$ s.t. soln to ② gives 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]

10h36 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 + convex surrogate

Consistency & calibration

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

relationship is usually very complicated

⇒ classical 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"

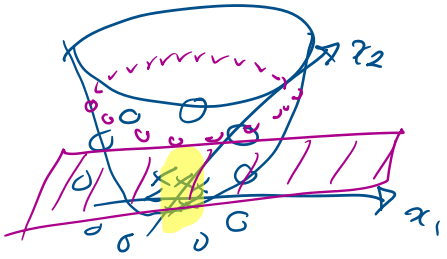
i.e. we suppose that $s(x,y;w)$ can be arbitrary for any x (i.e. w is ∞-dim)

→ can do this using a universal kernel

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

RKHS:

motivation for "kernel trick"



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

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

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

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

$$\langle \Phi(x), \Phi(x')\rangle_{\mathbb{R}^3} = \left(\langle x, x'\rangle_{\mathbb{R}^2}\right)^2 = (x_1x_1' + x_2x_2')^2$$

$$= k(x,x')$$

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

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

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

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

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

"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)

RKHS

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

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

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 $\{x_i\}_{i=1}^n$, $\alpha_i \in \mathbb{R}$

"pre-Hilbert" space [inner product space]

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

$\langle k(x_i^f, \cdot), k(x_j^g, \cdot) \rangle_{\tilde{\mathcal{H}}}$

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

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

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

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

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

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

$\Phi(x)$ ↙ $\langle w, \phi(x) \rangle$

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

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

$E_x(f) = f(x)$

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

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

⊛ this property is important to do statistics