# Lecture 6 - scribbles - generalization error bounds

Thursday, January 24, 2019 13:36

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. $1-\delta$ over $D_n$

$$\boxed{\forall w \in W \quad L_{01}(w) \leq \hat{L}_n(w) + \frac{1}{\sqrt{n}} \sqrt{d \log \frac{d}{n} + \log \frac{2}{\delta}}}$$

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

VC-dimension of $\mathcal{H} \triangleq \max \{ m :$ $\exists$ a set of $m$ pts, s.t. $\forall$ labelings of those points, $\exists w$ 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 linear classifiers of $p$ parameters, VC-dim $= p+1$

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

$\Rightarrow$ motivates going to data distribution dependent

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

example: empirical Rademacher complexity

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

"correlations with random noise"

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

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

$$\boxed{\forall w \quad L_{0,1}(w) \leq \hat{L}_n(w) + \hat{R}_{D_n}(\mathcal{H}) + \frac{1}{\sqrt{n}} 3 \sqrt{\frac{\log 2/\delta}{2}}}$$

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

high level idea to prove bound:

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

+

"symmetrization trick" $\longrightarrow$ bound sup of differences between $L(w)$ & $\hat{L}_n(w)$

+

union bound as usual + concentration inequality

Structured prediction generalization bounds [Cortes & al. NIPS 2016] general loss $\ell(y,y')$ s.t. $\ell(y,y') \neq 0$ if $y \neq y'$

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

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

depends on $\ell(y,y')$

$\max_{y,y'} \ell(y,y')$

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

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

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

"empirical factor graph complexity"

set of labels for $y_c$

Rademacher r.v.

**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 $R = \max_{i,c,\tilde{y}} \|\varphi_c(x_i, \tilde{y})\|_2$

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

so want small cliques ?

back to bound: $L(w) \leq \hat{\ell}_{hinge}(w) + \underbrace{\left( \frac{R|\mathcal{G}|}{\sqrt{n}} \sqrt{\max_c |\mathcal{Y}_c|} \right)}\, \underbrace{\Lambda}$

back to bound: $L(w) \leq S_{hinge}(w) + \underbrace{\left( \frac{R|G|}{\sqrt{n}} \frac{\sqrt{\max_c |\mathcal{Y}_c|}}{C} \right)}_{\lambda_n} \underbrace{\Omega}_{\|w\|_2}$

min of RHS suggests

SVMstruct alg. $\hat{w}_n = \arg\min \left\{ S_{hinge}(w) + \frac{\lambda_n \|w\|_2^2}{2} \right.$

missing link: (1) $\min f(w)$ s.t. $\|w\| \leq \Omega$ (if $f$ is convex) $\xrightarrow{\text{(use Lagrangian duality)}}$ $\exists$ a $\lambda(\Omega)$ st.

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

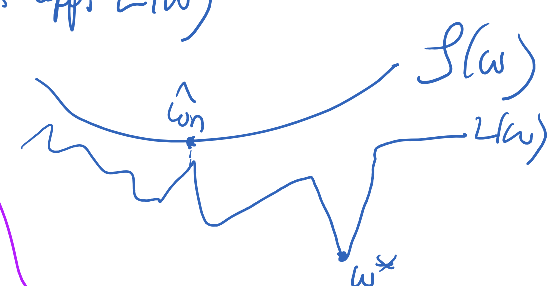
gives same solution as (1)

sidenote: penalized formulation less sensitive to choice of $\lambda$ vs. constrained formulation

can think of SVMstruct as minimizing upper bound on generalization error

properties:

- minimize upper bound, hope that minimizes approx. $L(w)$
  but no general guarantees? → caveat: here, no consistency guarantee




- can evaluate bound to get guarantees

next: consistency

guarantee

next: consistency
1h34

## Consistency & calibration fct.:

need to relate $\mathcal{L}(w)$ to $L(w)$: tool "calibration function" [Steinwart]

relationship is usually very complicated

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

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

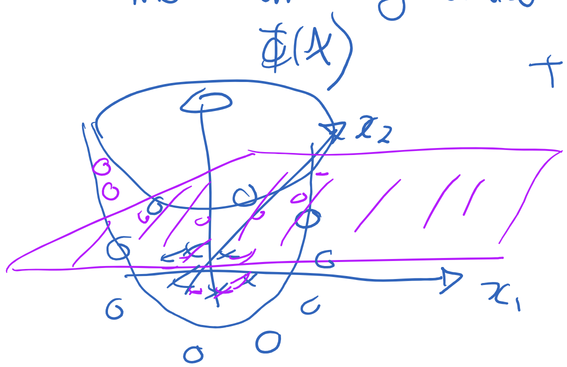
i.e. we suppose that $S(x,y;w)$ can be arbitrary for any $x$ (i.e. $w$ is $\infty$-dim) $\rightarrow$ could use a universal kernel

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

## RKHS

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

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




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

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

$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$

$$\Phi\begin{pmatrix}x_1\\x_2\end{pmatrix} \quad \begin{pmatrix}\sqrt{2}\,x_1x_2\\ x_2^2\end{pmatrix}$$

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

polynomial kernel e.g. $\left(\langle x, x' \rangle + 1\right)^d = \tilde{k}(x,x')$

equivalent to mapping data to space of dimension exponential in $d$

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

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

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

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

$\hookrightarrow$ "pre-Hilbert" space [inner product space]

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

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

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

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

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

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

nice property of RKHS, fct. evaluation is cts.

mapping $E_x : \mathcal{H} \to \mathbb{R}$
$E_x(f) = f(x)$

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

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

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

☆ this property important to do statistics