Generalisation Bounds (3): Rademacher average and bounds

Qinfeng (Javen) Shi

The Australian Centre for Visual Technologies,
The University of Adelaide, Australia

17 Aug. 2012

Course Outline

Generalisation Bounds:

1. Basics
2. VC dimensions and bounds
3. Rademacher complexity and bounds (Today)
4. PAC Bayesian Bounds
5. Regret bounds for online learning
6. . . .

Recap: VC bound

Denote h as the VC dimension. For all $n \geq h$ (since the growth function $S_{\mathcal{G}}(n) \leq (\frac{en}{h})^h$), we have

Theorem (VC bound)

For any $\delta \in (0,1)$, with probability at least $1-\delta$, $\forall g \in \mathcal{G}$

$$R(g) \leq R_n(g) + 2\sqrt{2\frac{h \log \frac{2en}{h} + \log(\frac{2}{\delta})}{n}}.$$

Problems:

- data dependency only come through training error
- very loose

Recap: VC dimension

Assume $x \in \mathbb{R}^d, \Phi(x) \in \mathbb{R}^D$ (Note D can be $+\infty$).

- linear $\langle x, w \rangle$, $h = d + 1$
- polynomial $(\langle x, w \rangle + 1)^p$, $h = \binom{d+p-1}{p} + 1$
- Gaussian RBF $\exp\left(-\frac{\|x-x'\|^2}{\sigma^2}\right)$, $h = +\infty$.
- Margin γ, $h \leq \min\{D, \lceil \frac{4R^2}{\gamma^2} \rceil\}$, where the radius $R^2 = \max_{i=1}^n \langle \Phi(x_i), \Phi(x_i) \rangle$ (assuming data are already centered)

Rademacher complexity (1)

Definition (Rademacher complexity)

Given $S = \{z_1, \cdots, z_n\}$ from a distribution P and a set of real-valued functions $\mathcal{G}$, the empirical Rademacher complexity of $\mathcal{G}$ is the random variable

$$\hat{\mathcal{R}}_n(\mathcal{G}, S) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right|\right],$$

where $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_n\}$ are independent uniform $\{\pm 1\}$-valued (Rademacher) random variables. The Rademacher complexity of $\mathcal{G}$ is

$$\mathcal{R}_n(\mathcal{G}) = \mathbb{E}_S[\hat{\mathcal{R}}_n(\mathcal{G}, S)] = \mathbb{E}_{S\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right|\right]$$

Rademacher complexity (2)

$\sup_{g\in\mathcal{G}} \left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right|$

- measures the best correlation between $g \in \mathcal{G}$ and random label (*i.e.* noise) $\sigma_i \sim U(\{-1, +1\})$.
- ability of $\mathcal{G}$ to fit noise.
- the smaller, the less chance of detected pattern being spurious
- if $|\,\mathcal{G}\,| = 1$, $\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g\in\mathcal{G}} \left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i g(z_i)\right|\right] = 0$.

Rademacher bound

Theorem (Rademacher)

Fix $\delta \in (0,1)$ and let $\mathcal{G}$ be a set of functions mapping from Z to $[a, a+1]$. Let $S = \{z_i\}_{i=1}^n$ be drawn i.i.d. from P. Then with probability at least $1-\delta$, $\forall g \in \mathcal{G}$,

$$\begin{aligned} \mathbb{E}_P[g(z)] &\le \hat{\mathbb{E}}[g(z)] + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n}} \\ &\le \hat{\mathbb{E}}[g(z)] + \hat{\mathcal{R}}_n(\mathcal{G}, S) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}, \end{aligned}$$

where $\hat{\mathbb{E}}[g(z)] = \frac{1}{n}\sum_{i=1}^n g(z_i)$

Note: $\hat{\mathcal{R}}_n(\mathcal{G}, S)$ is computable whereas $\mathcal{R}_n(\mathcal{G})$ is not.

Properties of empirical Rademacher complexity

Let $\mathcal{F}, \mathcal{F}_1, \cdots, \mathcal{F}_m$ and $\mathcal{G}$ be classes of real functions. Let $S = \{z_i\}_{i=1}^n$ i.i.d. from any unknown but fixed P.Then

1. If $\mathcal{F} \subseteq \mathcal{G}$, then $\hat{\mathcal{R}}_n(\mathcal{F}, S) \leq \hat{\mathcal{R}}_n(\mathcal{G}, S)$
2. For every $c \in \mathbb{R}$, $\hat{\mathcal{R}}_n(c\mathcal{F}, S) = |c|\hat{\mathcal{R}}_n(\mathcal{F}, S)$
3. $\hat{\mathcal{R}}_n(\sum_{i=1}^m \mathcal{F}_i, S) \leq \sum_{i=1}^m \hat{\mathcal{R}}_n(\mathcal{F}_i, S)$
4. For any function h,
   $\hat{\mathcal{R}}_n(\mathcal{F}+h, S) \leq \hat{\mathcal{R}}_n(\mathcal{F}, S) + 2\sqrt{\hat{\mathbb{E}}[h^2]/n}$
5. $\hat{\mathcal{R}}_n(\mathcal{F}, S) = \hat{\mathcal{R}}_n(\text{conv}(\mathcal{F}), S)$
6. If $\mathcal{A} : \mathbb{R} \to \mathbb{R}$ is Lpschitz with constant $L > 0$ (*i.e.* $|\mathcal{A}(a) - \mathcal{A}(a')| \leq L|a - a'|$ for all $a, a' \in \mathbb{R}$), and $\mathcal{A}(0) = 0$, then $\hat{\mathcal{R}}_n(\mathcal{A} \circ \mathcal{F}, S) \leq 2L\hat{\mathcal{R}}_n(\mathcal{F}, S)$

An example

Let $S = \{(x_i, y_i)\}_{i=1}^{n} \sim P^n$. $y_i \in \{-1, +1\}$
One form of soft margin binary SVMs is

$$\min_{w,\gamma,\xi} -\gamma + C \sum_{i=1}^{n} \xi_i \tag{1}$$

$$\text{s.t. } y_i \langle \phi(x_i), w \rangle \geq \gamma - \xi_i, \xi_i \geq 0, \|w\|^2 = 1$$

- The Rademacher Margin bound (next slide) applies.
- $\hat{\mathcal{R}}_n(\mathcal{G}, S)$ is essential, where $\mathcal{G} = \{-yf(x; w), f(x; w) = \langle \phi(x_i), w \rangle, \|w\|^2 = 1\}$.

Rademacher Margin bound

Theorem (Margin)

Fix $\gamma > 0, \delta \in (0,1)$, let $\mathcal{G}$ be the class of functions mapping from $\mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ given by $g(x,y) = -yf(x)$, where f is a linear function in a kernel-defined feature space with norm at most 1. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be drawn i.i.d. from $P(X,Y)$ and let $\xi_i = (\gamma - y_i f(x_i))_+$. Then with probability at least $1-\delta$ over sample of size n, we have

$$\mathbb{E}_P[\mathbf{1}_{y \neq \mathrm{sgn}(f(x))}] \leq \frac{1}{n\gamma}\sum_{i=1}^{n}\xi_i + \frac{4}{n\gamma}\sqrt{\mathrm{tr}(\mathbf{K})} + 3\sqrt{\frac{\ln(2/\delta)}{2n}},$$

- data dependency come through training error and margin
- tighter than VC bound $(\frac{4}{n\gamma}\sqrt{tr(\mathbf{K})} \leq \frac{4}{n\gamma}\sqrt{nR^2} \leq 4\sqrt{\frac{R^2}{n\gamma^2}})$

Proof of Margin bound (1)

Let $\mathcal{H}(a) = 1$ if $a > 0$, $\mathcal{H}(a) = 0$ otherwise. Thus
$\mathbf{1}_{y \neq \mathrm{sgn}(f(x))} = \mathcal{H}\left(-yf(x)\right)$
Let

$$\mathcal{A}(a) = \begin{cases} 1, & a > 0 \\ 1 + a/\gamma, & -\gamma \leq a \leq 0 \\ 0, & \text{otherwise.} \end{cases}$$

We can check that $\mathcal{H}(a) \leq \mathcal{A}(a)$ for all a.

$$\mathbb{E}_P[\mathbf{1}_{(y \neq f(x))} - 1] \leq \mathbb{E}_P[\mathcal{A}(-yf(x)) - 1]$$

$$\leq \hat{\mathbb{E}}[\mathcal{A}(-yf(x)) - 1] + \hat{\mathcal{R}}_n((\mathcal{A} - 1) \circ \mathcal{G}, S) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}$$

Proof of Margin bound (2)

Recall $\xi_i = (\gamma - y_i f(x_i))_+$. Thus

$$\mathcal{A}(-y_i f(x_i)) \leq 1 - y_i f(x_i)/\gamma \leq \frac{(\gamma - y_i f(x_i))_+}{\gamma} = \xi_i/\gamma$$

$$\mathbb{E}_P[\mathbf{1}_{(y \neq f(x))}] \leq \frac{1}{n}\sum_{i=1}^{n} \frac{\xi_i}{\gamma} + \hat{\mathcal{R}}_n((\mathcal{A} - 1) \circ \mathcal{G}, S) + 3\sqrt{\frac{\ln(2/\delta)}{2n}}$$

Apply property 6 (since $(\mathcal{A} - 1)(0) = 0$, $L = 1/\gamma$), we have

$$\hat{\mathcal{R}}_n((\mathcal{A} - 1) \circ \mathcal{G}, S) \leq 2\hat{\mathcal{R}}_n(\mathcal{G}, S)/\gamma$$

Proof of Margin bound (3)

$$\hat{\mathcal{R}}_n(\mathcal{G}, S) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i y_i f(x_i)\right|\right]$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left|\frac{2}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] (\text{if } \sigma_i \sim U(\{-1,+1\}), \text{then } \sigma_i y_i \sim U)$$

$$= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\|w\|^2=1}\left|\frac{2}{n}\left\langle w, \sum_{i=1}^{n}\sigma_i \phi(x_i)\right\rangle\right|\right]$$

$$\leq \frac{2}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i=1}^{n}\sigma_i \phi(x_i)\right\| \cdot \|w\|\right] (\text{Cauchy Schwarz ineq})$$

$$= \frac{2}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j k(x_i, x_j)\right]^{1/2}$$

Proof of Margin bound (4)

$$\frac{2}{n}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j k(x_i, x_j)\right]^{1/2}$$

$$\leq \frac{2}{n}\left\{\mathbb{E}_{\boldsymbol{\sigma}}\left[\sum_{i,j=1}^{n}\sigma_i\sigma_j k(x_i, x_j)\right]\right\}^{1/2} \text{(Jensen's ineq)}$$

$$= \frac{2}{n}\left\{\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i^2]k(x_i, x_i) + \sum_{i\neq j, i,j=1,}^{n}\mathbb{E}_{\boldsymbol{\sigma}}[\sigma_i\sigma_j]k(x_i, x_j)\right\}^{1/2}$$

$$= \frac{2}{n}\left(\sum_{i=1}^{n}k(x_i, x_i)\right)^{1/2} = \frac{2}{n}\sqrt{\mathrm{tr}(\mathbf{K})}$$

Proof of Rademacher bound (1)

$$\mathbb{E}_P[g(z)] - \hat{\mathbb{E}}_P[g(z)] \leq \sup_{f \in \mathcal{G}} \left(\mathbb{E}_P[f(z)] - \hat{\mathbb{E}}_P[f(z)] \right) \text{(by sup def)}$$

$$\mathbb{E}_P[g(z)] \leq \hat{\mathbb{E}}_P[g(z)] + \sup_{f \in \mathcal{G}} \left(\mathbb{E}_P[f(z)] - \hat{\mathbb{E}}_P[f(z)] \right)$$

$$= \hat{\mathbb{E}}_P[g(z)] + \underbrace{\sup_{f \in \mathcal{G}} \left(\mathbb{E}_P[f(z)] - \frac{1}{n} \sum_{i=1}^{n} [f(z_i)] \right)}_{:= f'(z_1, \cdots, z_n)}$$

$$f'(z_1, \cdots, z_n) \leq \mathbb{E}_S[f'(z_1, \cdots, z_n)] + \sqrt{\frac{\ln(2/\delta)}{2n}} \text{(McDiarmid's ineq)}$$

$$\Rightarrow$$

$$\mathbb{E}_P[g(z)] \leq \hat{\mathbb{E}}_P[g(z)] + \mathbb{E}_S[f'(z_1, \cdots, z_n)] + \sqrt{\frac{\ln(2/\delta)}{2n}} \quad (2)$$

Proof of Rademacher bound (2)

$$
\begin{aligned}
&\mathbb{E}_{S\sim P^n}[f'(x_1,\cdots,x_n)] \\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{G}}\left(\mathbb{E}_{z\sim P(z)}[f(z)] - \hat{\mathbb{E}}_P[f(z)]\right)\right] \\
&= \mathbb{E}_S\left[\sup_{f\in\mathcal{G}}\left(\mathbb{E}_{\tilde{S}\sim P^n}[\frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i)] - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right)\right] \\
&= \mathbb{E}_S\left\{\sup_{f\in\mathcal{G}}\mathbb{E}_{\tilde{S}}\left[\frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right]\right\} \\
&\leq \mathbb{E}_S\,\mathbb{E}_{\tilde{S}}\sup_{f\in\mathcal{G}}\left[\frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right]
\end{aligned}
$$

Proof of Rademacher bound (3)

$$\mathbb{E}_S \, \mathbb{E}_{\tilde{S}} \sup_{f\in\mathcal{G}} \left[\frac{1}{n}\sum_{i=1}^{n} f(\tilde{z}_i) - \frac{1}{n}\sum_{i=1}^{n} f(z_i)\right]$$

$$= \mathbb{E}_{\boldsymbol{\sigma} S\tilde{S}} \sup_{f\in\mathcal{G}} \left[\frac{1}{n}\sum_{i=1}^{n} \sigma_i (f(\tilde{z}_i) - f(z_i))\right]$$

$$\leq \mathbb{E}_{\boldsymbol{\sigma} S} \left[\sup_{f\in\mathcal{G}} \left|\frac{2}{n}\sum_{i=1}^{n} \sigma_i f(z_i)\right|\right]$$

$$= \mathcal{R}_n(\mathcal{G})$$

Via equation (2), we have

$$\mathbb{E}_P[g(z)] \leq \hat{\mathbb{E}}_P[g(z)] + \mathbb{E}_S[f'(x_1, \cdots, x_n)] + \sqrt{\frac{\ln(2/\delta)}{2n}}$$

$$\leq \hat{\mathbb{E}}_P[g(z)] + \mathcal{R}_n(\mathcal{G}) + \sqrt{\frac{\ln(2/\delta)}{2n}}$$

Related concepts

- PAC bayesian bounds (will be covered in the next talk)