Rademacher Complexity and VC Dimension

Hanxiao Liu

January 13, 2015

Rademacher Complexity

Notations

- Data $z_i = (x_i, y_i) \sim D$, $S = \{z_1, z_2, \dots z_m\} \sim D^m$
- Mapping from data to loss: $g(z_i) = L(h(x_i), y_i) \in [0, 1]$
- Rademacher RVs: $\sigma_i \overset{Unif}{\sim} \{-1, +1\}$

Empirical Rademacher Complexity

$$\hat{\mathfrak{R}}_S(G) = \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i)\right]$$

Rademacher Complexity

$$\mathfrak{R}_m(G) = \mathbb{E}_{S \sim D^m}\left[\hat{\mathfrak{R}}_S(G)\right]$$

Rademacher Generalization Bound

With probability $> 1 - \delta$

$$\underbrace{\sup_{g \in G} \left(\mathbb{E}\left[g\left(z\right)\right] - \frac{1}{m}\sum_{i=1}^{m} g\left(z_i\right) \right)}_{\Phi(z_1,\ldots z_m)} \leq 2\mathfrak{R}_m\left(G\right) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Theorem (McDiarmid's Inequality)

If $\left|\Phi\left(z_1, \ldots, z_i, \ldots, z_m\right) - \Phi\left(z_1, \ldots z_i', \ldots, z_m\right)\right| \leq \frac{1}{m}$

$$\Phi\left(z_1, \ldots z_m\right) \leq \mathbb{E}\left[\Phi\left(z_1, \ldots z_m\right)\right] + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Therefore, it suffices to show $\mathbb{E}\left[\Phi\left(z_1, \ldots z_m\right)\right] = 2\mathfrak{R}_m\left(G\right)$

$$
\begin{aligned}
&\mathbb{E}_S\left[\Phi\left(z_1,\ldots z_m\right)\right] \\
&= \mathbb{E}_S\left[\sup_{g\in G}\mathbb{E}\left(g\right)-\hat{\mathbb{E}}_S\left(g\right)\right] = \mathbb{E}_S\left[\sup_{g\in G}\mathbb{E}_{S'}\left[\hat{\mathbb{E}}_{S'}\left(g\right)-\hat{\mathbb{E}}_S\left(g\right)\right]\right] \\
&\leq \mathbb{E}_{S,S'}\left[\sup_{g\in G}\hat{\mathbb{E}}_{S'}\left(g\right)-\hat{\mathbb{E}}_S\left(g\right)\right] = \mathbb{E}_{S,S'}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\left(g\left(z_i'\right)-g\left(z_i\right)\right)\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma},S,S'}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\left(g\left(z_i'\right)-g\left(z_i\right)\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma},S'}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}\sigma_i g\left(z_i'\right)\right] + \mathbb{E}_{\boldsymbol{\sigma},S}\left[\sup_{g\in G}\frac{1}{m}\sum_{i=1}^{m}-\sigma_i g\left(z_i\right)\right] \\
&= 2\mathfrak{R}_m\left(G\right)
\end{aligned}
$$

Data-dependent Bound

From McDiarmid's

$$\mathfrak{R}_m(G) \leq \hat{\mathfrak{R}}_S(G) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

$$\Longrightarrow \sup_{g \in G} \left(\mathbb{E}[g(z)] - \frac{1}{m} \sum_{i=1}^{m} g(z_i) \right) \leq 2\hat{\mathfrak{R}}_S(G) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

When $h \in H$ is binary, we can get bound w.r.t. H instead of G

$$\sup_{h \in H} \left(R(h) - \hat{R}(h) \right) \leq \mathfrak{R}_m(H) + \sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

$$\sup_{h \in H} \left(R(h) - \hat{R}(h) \right) \leq \hat{\mathfrak{R}}_S(H) + 3\sqrt{\frac{\log \frac{2}{\delta}}{2m}}$$

Growth function

In Rademacher complexity, $\sup_{g \in G} \frac{1}{m} \sum_{i=1}^{m} \sigma_i g(x_i)$ can be hard to compute

R complexity is bounded by another quantity called the growth function (a.k.a. shattering number), which is easier to deal with

Definition (Growth function)

$$\forall m \in \mathbb{N},\ \Pi_H(m) = \max_S |\{h(x_1), \ldots h(x_m)\} : h \in H| \triangleq \max_S |H_{|S}|$$

$\Pi_H(m)$: maximum number of distinct ways in which m points can be classified. Hence $\Pi_H(m) \leq 2^m$.

From Massart's lemma

$$\mathfrak{R}_m(H) \leq \sqrt{\frac{2 \log \Pi_H(m)}{m}}$$

VC Dimension

What if we want to further get rid of "m" in growth function $\Pi_H(m)$? —VC-dimension

Given H, as m grows, it becomes more and more unlikely that the data points can be classified in 2^m ways by $h \in H$

Definition (VC Dimension)

$$VCdim(H) = \max\{m : \Pi_H(m) = 2^m\}$$

E.g.: $VCdim$(intervals) $= 2$, $VCdim$(hyperplanes) in $\mathbb{R}^2 = 3$, ...

Why VC-dimension? $\Pi_H(m) = O\left(m^{VCdim(H)}\right)$

- can be derived from Sauer's lemma

Sauer's lemma

Theorem (Sauer's lemma)

Let $VCdim\,(H) = d$, $\forall m \in \mathbb{N}$

$$\Pi_H\,(m) \leq \sum_{i=0}^{d} \binom{m}{i} \overset{def}{=} \kappa\,(m, d)$$

Assume the lemma holds for $(m-1, d-1)$ and $(m-1, d)$. Let $S = \{x_1, \ldots, x_m\}$, $S' = \{x_1, \ldots x_{m-1}\}$.
We can close the proof if $\forall H_{|S}$, $\exists H_1$, H_2 s.t.

- $|H_{|S}| = |H_{1_{|S'}}| + |H_{2_{|S'}}|$
- $VCdim\,(H_1) \leq d$, $VCdim\,(H_2) \leq d-1$.

Why? Because in this case

$$\begin{aligned} |H_{|S}| = |H_{1_{|S'}}| + |H_{2_{|S'}}| &\leq \Pi_{H_1}\,(m-1) + \Pi_{H_2}\,(m-1) \\ &\leq \kappa\,(m-1, d) + \kappa\,(m-1, d-1) \\ &\equiv \kappa\,(m, d) \end{aligned}$$

Sauer's Lemma

$$
\begin{aligned}
&\kappa(m-1,d)+\kappa(m-1,d-1)\\
=&\sum_{i=0}^{d}\binom{m-1}{i}+\sum_{i=0}^{d-1}\binom{m-1}{i}\\
=&\binom{m-1}{0}+\sum_{i=1}^{d}\binom{m-1}{i}+\sum_{i=1}^{d}\binom{m-1}{i-1}\\
=&1+\sum_{i=1}^{d}\left[\binom{m-1}{i}+\binom{m-1}{i-1}\right]\\
=&1+\sum_{i=1}^{d}\binom{m}{i}=\sum_{i=0}^{d}\binom{m}{i}=\kappa(m,d)
\end{aligned}
$$

	H						H_1					H_2			
	x_1	x_2	x_3	x_4	x_5		x_1	x_2	x_3	x_4		x_1	x_2	x_3	x_4
h_1	0	1	1	0	0	$\rightarrow$	0	1	1	0					
h_2	0	1	1	0	1						$\rightarrow$	0	1	1	0
h_3	0	1	1	1	0	$\rightarrow$	0	1	1	1					
h_4	1	0	0	1	0	$\rightarrow$	1	0	0	1					
h_5	1	0	0	1	1						$\rightarrow$	1	0	0	1
h_6	1	1	0	0	1	$\rightarrow$	1	1	0	0					

Construction procedure [1]

- H_1: ignore the behavior on x_5
- H_2: dichotomies that "collapsed" in H_1

Check

- $|H_{|S}| = |H_{1_{|S'}}| + |H_{2_{|S'}}|$
- $VCdim\left(H_1\right) \leq VCdim\left(H\right) = d$
- Notice if S' is shattered by H_2, then $S' \cup \{x_5\}$ can always be shattered by $H \implies VCdim\left(H_2\right) \leq d - 1$

[1] thanks to http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0220.pdf

VC Generalization Bound

Sauer's lemma implies[2]

$$\Pi_H(m) \leq \left(\frac{em}{d}\right)^d$$

Further recall that

$$\mathfrak{R}_m(H) \leq \sqrt{\frac{2\log \Pi_H(m)}{m}}$$

Therefore, from Rademacher generalization bound

Theorem (VC-dimension Generalization Bound)

With probability $> 1-\delta$,

$$R(h) \leq \hat{R}(h) + \sqrt{\frac{2d\log\frac{em}{d}}{m}} + \sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

[2] see also http://www.svms.org/vc-dimension/ for a visualization

VC Generalization Bound

We can directly achieve a similar VC bound (of the same order) without using Rademacher complexity

Theorem (Vapnik and Chervonenkis)

$$\mathbb{P}\left(|R(h) - \hat{R}(h)| > \epsilon\right) \leq 4\Pi_H(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)$$

The proof relies on the following lemma [3]

Lemma (Symmetrization)

$\forall \epsilon > \sqrt{\frac{2}{m}}$, *let* $S' = \{x'_1, x'_2, \ldots x'_m\}$ *be a ghost sample*

$$\mathbb{P}\left(\sup_{h \in H} |R(h) - \hat{R}_S(h)| > \epsilon\right) \leq 2\mathbb{P}\left(\sup_{h \in H} |\hat{R}_{S'}(h) - \hat{R}_S(h)| > \frac{\epsilon}{2}\right)$$

i.e. if samples are concentrated, then they are all close to the mean.

[3] thanks to http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf

VC Generalization Bound

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{h\in H}\left|R(h)-\hat{R}_S(h)\right|>\epsilon\right)\\
\le&2\mathbb{P}\left(\sup_{h\in H}\left|\hat{R}_{S'}(h)-\hat{R}_S(h)\right|>\frac{\epsilon}{2}\right)\\
=&2\mathbb{P}\left(\max_{v\in\left\{H_{|S}\cup H_{|S'}\right\}}\left|\hat{R}_{S'}(v)-\hat{R}_S(v)\right|>\frac{\epsilon}{2}\right)\\
\le&2\sum_{v\in\left\{H_{|S}\cup H_{|S'}\right\}}\mathbb{P}\left(\left|\hat{R}_{S'}(v)-\hat{R}_S(v)\right|>\frac{\epsilon}{2}\right)\\
\le&2\sum_{v\in\left\{H_{|S}\cup H_{|S'}\right\}}2\exp\left(-\frac{m\epsilon^2}{8}\right)\\
\le&4\Pi_H(2m)\exp\left(-\frac{m\epsilon^2}{8}\right)
\end{aligned}
$$

2-sample Hoeffding's: $\mathbb{P}\left(\hat{R}_{S'}(v)-\hat{R}_S(v)>\epsilon\right)\le\exp\left(-\frac{n\epsilon^2}{2}\right)$

Lower Bound

Theorem (Lower bound, realizable case)

For $d > 1$, $\exists$ a "bad" distribution D and target function f, s.t.

$$\mathbb{P}_{S\sim D^m}\left[R_D\left(h_S, f\right) > \frac{d-1}{32m}\right] \geq \frac{1}{100}$$

Theorem (Lower bound, non-realizable case)

For $d > 1$, $\exists$ a "bad" distribution D, s.t.

$$\mathbb{P}_{S\sim D^m}\left(R_D\left(h_S\right) > \inf_{h\in H} R_D\left(h\right) + \sqrt{\frac{d}{320m}}\right) \geq \frac{1}{64}$$

- realizable: $x \sim D$, $\exists f : y = f\left(x\right)$; non-realizable: $\left(x, y\right) \sim D$.
- h_S: hypothesis learned based on S using any algorithm
- $R_D\left(h_S, f\right)$ and $R_D\left(h_S\right)$: the best we can do
- $\inf_{h\in H} R_D\left(h\right)$: the true optimal

Reference I

Bartlett, P. L. and Mendelson, S. (2003).
Rademacher and gaussian complexities: Risk bounds and structural results.
The Journal of Machine Learning Research, 3:463–482.

Feng, Y. and Schapire, R. (2008).
Theoretical Machine Learning.
http://www.cs.princeton.edu/courses/archive/spr08/cos511/scribe_notes/0220.pdf.
[Online; accessed 27-Dec-2014].

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012).
Foundations of machine learning.
MIT press.

Wasserman, L. (2008).
Concentration of Measure.
http://www.stat.cmu.edu/~larry/=sml/Concentration.pdf.
[Online; accessed 27-Dec-2014].