Rademacher Complexities

Peng Lingwei

August 7, 2019

Contents

26 Rademacher Complexities

1. **Uniform convergence** is a sufficient condition for learnability.
2. **Rademacher complexities** measures the rate of uniform convergence.

26.1 THE RADEMACHER COMPLEXITY

Definition 1. ***(ϵ-Representative Sample).*** *(w.r.t. domain $Z = (\mathcal{X}, \mathcal{Y}) \sim \mathcal{D}$, hypothesis class $\mathcal{H}$, loss function l). A training set S is called ϵ-representative if*

$$\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \epsilon$$

We have $m_{\mathcal{H}}(\epsilon, \delta) \le m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$.

Definition 2. ***(The representativeness of S with respect to $\mathcal{F}$).***

$$Rep_{\mathcal{D}}(\mathcal{F}, S) := \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) \tag{1}$$

where,

$$\mathcal{F} := l \circ \mathcal{H} := \{z \mapsto l(h, z) : z \in Z, h \in \mathcal{H}\}$$

$$f \in \mathcal{F}, \quad L_{\mathcal{D}}(f) = \mathbb{E}_{z \sim \mathcal{D}}[f(z)], \quad L_S = \frac{1}{m} \sum_{i=1}^{m} f(z_i).$$

Analogizing the concept of validation set which used to estimate the representativeness of S, we define **Rademacher complexity**.

Definition 3. ***(The rademacher complexity of $\mathcal{F}$ w.r.t. S).***

$$R(\mathcal{F} \circ S) := \frac{1}{m} \mathbb{E}_{\sigma \sim \{\pm 1\}^m} \left[\sup_{f \in \mathcal{F}} \sum_{i=1}^{m} \sigma_i f(z_i) \right] \tag{2}$$

where,

$$\mathcal{F} \circ S = \{(f(z_1), \ldots, f(z_m)) : f \in \mathcal{F}\}$$

$$\sigma = \{\sigma_i : \mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = 0.5\}$$

More generally, given a set of vectors, $A \subset \mathbb{R}^m$, we define

$$R(A) := \frac{1}{m} \mathbb{E}_{\sigma} \left[\sup_{\mathbf{a} \in A} \sum_{i=1}^{m} \sigma_i \mathbf{a}_i \right]$$

Lemma 1.

$$\mathbb{E}_{S \sim \mathcal{D}^m}[Rep_{\mathcal{D}}(\mathcal{F}, S)] \le 2 \mathbb{E}_{S \sim \mathcal{D}^m} R(\mathcal{F} \circ S) \tag{3}$$

Proof. Let $S' = \{z'_1, \ldots, z'_m\}$ be another i.i.d. sample.Then,

$$L_{\mathcal{D}}(f) - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f)] - L_S(f) = \mathbb{E}_{S'}[L_{S'}(f) - L_S(f)]$$

$$\begin{aligned} Rep_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} (L_{\mathcal{D}}(f) - L_S(f)) &= \sup_{f \in \mathcal{F}} (\mathbb{E}_{S'}[L_{S'}(f) - L_S(f)]) \\ &\le \mathbb{E}_{S'} \left[\sup_{f \in \mathcal{F}} (L_{S'}(f) - L_S(f)) \right] \end{aligned}$$

$$\mathbb{E}_{S\sim\mathcal{D}^m}[Rep_{\mathcal{D}}(\mathcal{F},S)] \le \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}(L_{S'}(f)-L_S(f))\right] \le \frac{1}{m}\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}(f(z_i')-f(z_i))\right]$$

In some techniques, we can get:

$$\begin{aligned}
\mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}(f(z_i')-f(z_i))\right] =&\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i')-f(z_i))\right]\\
\le&\mathbb{E}_{S,S',\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(z_i'))+\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}(-\sigma_i)f(z_i)\right]\\
=&m\mathbb{E}_{S'}[R(\mathcal{F}\circ S')]+m\mathbb{E}_S[R(\mathcal{F}\circ S)]=2m\mathbb{E}_S[R(\mathcal{F}\circ S)].
\end{aligned}$$

$\square$

Theorem 1.

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(ERM_{\mathcal{H}}(S))-L_S(ERM_{\mathcal{H}}(S))] \le 2\mathbb{E}_{S\sim\mathcal{D}^m}R(l\circ\mathcal{H}\circ S)$$

$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(ERM_{\mathcal{H}}(S))-L_S(h^*)] \le 2\mathbb{E}_{S\sim\mathcal{D}^m}R(l\circ\mathcal{H}\circ S)$, *where* $h^*=\arg\min_h L_{\mathcal{D}}(h)$

Because $L_{\mathcal{D}}(ERM_{\mathcal{H}}(S))-L_{\mathcal{D}}(h^*)\ge 0$, *then*

$$\mathbb{P}\left\{L_{\mathcal{D}}(ERM_{\mathcal{H}}(S))-L_{\mathcal{D}}(h^*)\ge\frac{2\mathbb{E}_{S'\sim D^m}R(l\circ\mathcal{H}\circ S')}{\delta}\right\}\le\delta$$

Lemma 2. ***(McDiarmid's Inequality).***
If

$$f(x_1,\dots,x_m)-f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_m)\in[a_i,b_i].$$

then,

$$\mathbb{P}\{f-\mathbb{E}f\ge\epsilon\}\le\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(a_i-b_i)^2}\right)$$

$$\mathbb{P}\{f-\mathbb{E}f\le-\epsilon\}\le\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{m}(a_i-b_i)^2}\right)$$

which also means

$$\mathbb{P}\left\{|f-\mathbb{E}f|\ge\sqrt{\frac{\sum_{i=1}^{m}(a_i-b_i)^2}{2}\log(2/\delta)}\right\}\le\delta$$

$$\mathbb{P}\left\{|f-\mathbb{E}f|\ge(b-a)\sqrt{\frac{m\log(2/\delta)}{2}}\right\}\le\delta$$

Theorem 2. ***(Data-dependent bound).*** *Assume that for all* z *and* $h\in\mathcal{H}$, *we have that* $l(h,z)\in[a,b]$. *Then,*

1. $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\forall h\in\mathcal{H}, L_{\mathcal{D}}(h)-L_S(h)\le 2\mathbb{E}_{S'\sim\mathcal{D}^m}R(l\circ\mathcal{H}\circ S')+(b-a)\sqrt{2\ln(1/\delta)/m}\right\}\ge 1-\delta$$

Proof. $Rep_{\mathcal{D}}(\mathcal{F}, S)$ satisfies the preceeding condition with a constant $[(a-b)/m, (b-a)/m]$, $\square$

2. $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) - L_S(h) \le 2R(l\circ\mathcal{H}\circ S) + 3(b-a)\sqrt{2\ln(2/\delta)/m}\right\} \ge 1-\delta$$

   Proof.
   $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{Rep_{\mathcal{D}}(F, S) \le \mathbb{E}_{S'} Rep_{\mathcal{D}}(l\circ\mathcal{H}\circ S') + (b-a)\sqrt{2\ln(2/\delta)/m}\right\} \ge 1-\delta/2$$
   $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\mathbb{E} Rep_{\mathcal{D}}(F, S) \le 2\mathbb{E}R(l\circ\mathcal{H}\circ S')\right\} = 1$$
   $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{\mathbb{E}_{S'} R(l\circ\mathcal{H}\circ S') \le R(l\circ\mathcal{H}\circ S) + (b-a)\sqrt{2\ln(2/\delta)/m}\right\} \ge 1-\delta/2$$
   $\square$

3. $$\forall h \in \mathcal{H}, \mathbb{P}_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(ERM_{\mathcal{H}}(S)) - L_{\mathcal{D}}(h) \le 2R(l\circ\mathcal{H}\circ S) + 4(b-a)\sqrt{2\ln(3/\delta)/m}\right\} \ge 1-\delta$$

   Proof.
   $$\begin{aligned} L_{\mathcal{D}}(h_S) - L_{\mathcal{D}}(h) =& L_{\mathcal{D}}(h_S) - L_S(h_S) + L_S(h_S) - L_S(h) + L_S(h) - L_{\mathcal{D}}(h) \\ \le& (L_{\mathcal{D}}(h_S) - L_S(h_S)) + (L_S(h) - L_{\mathcal{D}}(h)) \end{aligned}$$
   $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{L_{\mathcal{D}}(h_S) - L_S(h_S) \le 2R(l\circ\mathcal{H}\circ S) + 3(b-a)\sqrt{2\ln(3/\delta)/m}\right\} \ge 1-2\delta/3$$
   Because $L_{\mathcal{D}}(h)$ does not depend on S, so we can use hoeffding's inequality to get
   $$\mathbb{P}_{S\sim\mathcal{D}^m}\left\{L_S(h) - L_{\mathcal{D}}(h) \le (b-a)\sqrt{\ln(3/\delta)/(2m)}\right\} \ge 1-\delta/3$$
   $\square$

26.1.1 Rademacher Calculus

Lemma 3. *$\forall A \subset \mathbb{R}^m, c \in \mathbb{R}, \mathbf{a}_0 \in \mathbb{R}^m$, we have*
$$R(\{c\mathbf{a} + \mathbf{a}_0 : \mathbf{a} \in A\}) = |c|R(A) \tag{4}$$

Lemma 4. *$\forall A \subset \mathbb{R}^m$, if $A' = \left\{\sum_{j=1}^N \alpha_j \mathbf{a}^{(j)} : N \in \mathbb{N}, \forall j, \mathbf{a}^{(j)} \in A, \alpha_j \ge 0, \|\vec{\alpha}\|_1 = 1\right\}$, then $R(A') = R(A)$.*

Proof.
$$\begin{aligned} mR(A') =& \mathbb{E}_\sigma \sup_{\vec{\alpha}\succeq\vec{0}:\|\vec{\alpha}\|_1=1} \sup_{\mathbf{a}^{(1)},\dots,\mathbf{a}^{(N)}} \sum_{i=1}^m \sigma_i \sum_{j=1}^N \alpha_j a_i^{(j)} \quad (\vec{\sigma}^T\mathbf{A}\vec{\alpha}) \\ =& \mathbb{E}_\sigma \sup_{\vec{\alpha}\succeq\vec{0}:\|\vec{\alpha}\|_1=1} \sum_{j=1}^N \alpha_j \sup_{\mathbf{a}^{(j)}} \sum_{i=1}^m \sigma_i a_i^{(j)} \quad (\vec{\alpha}^T\mathbf{A}^T\vec{\sigma}) \\ =& \mathbb{E}_\sigma \sup_{\mathbf{a}\in A} \sum_{i=1}^m \sigma_i a_i \end{aligned}$$
$\square$

Lemma 5. ***(Massart Lemma)****. Let $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_N\}$ be a finite set of vectors in $\mathbf{R}^m$. Define $\bar{\mathbf{a}} = \frac{1}{N}\sum_{i=1}^N \mathbf{a}_i$. Then,*

$$R(A) \leq \max_{\mathbf{a}\in A} \|\mathbf{a} - \bar{\mathbf{a}}\|_2 \frac{\sqrt{2\log(N)}}{m} \tag{5}$$

Proof.

$$\begin{aligned}
\forall A, \quad mR(A) =& \mathbb{E}_{\vec{\sigma}}\left[\max_{\mathbf{a}\in A}\langle\vec{\sigma}, \mathbf{a}\rangle\right] = \mathbb{E}_{\vec{\sigma}}\left[\log\left(\max_{\mathbf{a}\in A} e^{\langle\vec{\sigma},\mathbf{a}\rangle}\right)\right] \\
=& \mathbb{E}_{\vec{\sigma}}\left[\log\left(\sum_{\mathbf{a}\in A} e^{\langle\vec{\sigma},\mathbf{a}\rangle}\right)\right] \leq \log\left[\mathbb{E}_{\vec{\sigma}}\left(\sum_{\mathbf{a}\in A} e^{\langle\vec{\sigma},\mathbf{a}\rangle}\right)\right] \\
\leq& \log\left(\sum_{\mathbf{a}\in A}\prod_{i=1}^m \mathbb{E}_{\sigma_i}[e^{\sigma_i a_i}]\right) = \log\left(\sum_{\mathbf{a}\in A}\prod_{i=1}^m [e^{a_i} + e^{-a_i}]/2\right) \\
\leq& \log\left(\sum_{\mathbf{a}\in A}\prod_{i=1}^m e^{a_i^2/2}\right) = \log\left(\sum_{a\in A}\exp\left(\|\mathbf{a}\|_2^2/2\right)\right) \\
\leq& \log\left(|A|\max_{\mathbf{a}\in A}\exp\left(\|\mathbf{a}\|_2^2/2\right)\right) = \log\left(|A|\right) + \max_{\mathbf{a}\in A}(\|\mathbf{a}\|_2^2/2)
\end{aligned}$$

We let $\lambda > 0, A' = \lambda A$,then $R(A) = R(A')/\lambda$ we obtain that

$$R(A) \leq \frac{\log(|A'|) + \max_{\lambda\mathbf{a}\in A'}(\|\lambda\mathbf{a}\|_2^2/2)}{m} = \frac{\log(|A|) + \lambda^2\max_{\mathbf{a}\in A}(\|\mathbf{a}\|_2^2/2)}{\lambda m}$$

□

Lemma 6. ***(Contraction Lemma).*** *$\forall i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a $\rho -$ Lipschitz function. For $\mathbf{a} \in \mathbb{R}^m$ let $\phi(a) = (\phi_1(a1), \ldots, \phi_m(y_m))$. Let $\phi \circ A = \{\phi(\vec{a}) : a \in A\}$. Then,*

$$R(\phi \circ A) \leq \rho R(A).$$

Proof. First, $\rho = 1$. Let $A_i = \{(a_1, \ldots, a_{i-1}, \phi_i(a_i), a_{i+1}, \ldots, a_m) : \mathbf{a} \in A\}$.

$$\begin{aligned}
mR(A_1) =& \mathbb{E}_\sigma\left[\sup_{\mathbf{a}\in A_1}\sum_{i=1}^m \sigma_i a_i\right] \\
=& \mathbb{E}_\sigma\left[\sup_{\mathbf{a}\in A}\sigma_1\phi(a_1) + \sum_{i=2}^m \sigma_i a_i\right] \\
=& \frac{1}{2}\mathbb{E}_{\sigma_2,\ldots,\sigma_m}\left[\sup_{\mathbf{a},\mathbf{a}'\in A}\left(\phi(a_1) - \phi(a_1') + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a_i'\right)\right] \\
\leq& \frac{1}{2}\mathbb{E}_{\sigma_2,\ldots,\sigma_m}\left[\sup_{\mathbf{a},\mathbf{a}'\in A}\left(|a_1 - a_1'| + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a_i'\right)\right] \\
=& \frac{1}{2}\mathbb{E}_{\sigma_2,\ldots,\sigma_m}\left[\sup_{\mathbf{a},\mathbf{a}'\in A}\left(a_1 - a_1' + \sum_{i=2}^m \sigma_i a_i + \sum_{i=2}^m \sigma_i a_i'\right)\right] = mR(A) \\
mR(A_1) \leq& mR(A)
\end{aligned}$$

□

26.2 RADEMACHER COMPLEXITY OF LINEAR CLASSES

1. $\mathcal{H}_1 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_1 \leq 1\}$
2. $\mathcal{H}_2 = \{\mathbf{x} \mapsto \langle \mathbf{w}, \mathbf{x} \rangle : \|\mathbf{w}\|_2 \leq 1\}$

Lemma 7.

$$R(\mathcal{H}_2 \circ S) \leq \frac{\max_i \|\mathbf{x}_i\|_2}{\sqrt{m}} \tag{6}$$

Proof.

$$\begin{aligned}
mR(\mathcal{H}_2 \circ S) =& \mathbb{E}_\sigma \left[\sup_{\mathbf{a} \in \mathcal{H}_2 \circ S} \sum_{i=1}^m \sigma_i a_i \right] = \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq 1} \sum_{i=1}^m \sigma_i \langle \mathbf{w}, \mathbf{x}_i \rangle \right] \\
=& \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_2 \leq 1} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i \mathbf{x}_i \rangle \right] \leq \mathbb{E}_\sigma \left[\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \|_2 \right] \\
\leq& \left(\mathbb{E}_\sigma \left[\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \|_2^2 \right] \right)^{1/2}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}_\sigma \left[\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \|_2^2 \right] &= \sum_{i \neq j} \langle \mathbf{x_i}, \mathbf{x_j} \rangle \mathbb{E}_\sigma[\sigma_i \sigma_j] + \sum_{i=1}^m \langle \mathbf{x}_i, \mathbf{x}_i \rangle \mathbb{E}_\sigma[\sigma_i^2] \\
&= \sum_{i=1}^m \|\mathbf{x}_i\|_2^2 \leq m \max_i \|\mathbf{x}_i\|_2^2
\end{aligned}$$

□

Lemma 8. *Let $S = (\mathbf{x}_1, \ldots, \mathbf{x}_m)$ be the vectors in $\mathbb{R}^n$, then,*

$$R(\mathcal{H}_1 \circ S) \leq \max_i \|\mathbf{x}_i\|_\infty \sqrt{\frac{2\log(2n)}{m}} \tag{7}$$

Proof. Using Holder's inequality, we have $\langle \mathbf{w}, \mathbf{v} \rangle \leq \|\mathbf{w}\|_1 \|\mathbf{v}\|_\infty$. Therefore,

$$mR(\mathcal{H}_1 \circ S) = \mathbb{E}_\sigma \left[\sup_{\mathbf{w}: \|\mathbf{w}\|_1 \leq 1} \langle \mathbf{w}, \sum_{i=1}^m \sigma_i \mathbf{x}_i \rangle \right] \leq \mathbb{E}_\sigma \left[\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \|_\infty \right].$$

Let $j \in [n]$ and $\mathbf{v}_j = (x_{1,j}, \ldots, x_{m,j}) \in \mathbb{R}^m$, and $V = \{\mathbf{v}_1, \ldots, \mathbf{v}_n, -\mathbf{v}_1, \ldots, -\mathbf{v}_n\}$. Note that $\|\mathbf{v}_j\|_2 \leq \sqrt{m} \max_i \|\mathbf{x}_i\|_\infty$.

$$\begin{aligned}
\mathbb{E}_\sigma \left[\| \sum_{i=1}^m \sigma_i \mathbf{x}_i \|_\infty \right] =& \mathbb{E}_\sigma \left[\max_j |\langle \mathbf{v}_j, \sigma \rangle| \right] = mR(V) \\
\leq& m \cdot \max_j \|\mathbf{v}_j\|_2 \frac{\sqrt{2\log(2n)}}{m} \\
\leq& m \cdot \max_i \|\mathbf{x}_i\|_\infty \sqrt{2\log(2n)/m}
\end{aligned}$$

□

In chapter6, we defined “Restriction of $\mathcal{H}$ to S ”:

$$\mathcal{H}_S = \{(h(s_1), h(s_2), \ldots, h(s_m)) : h \in \mathcal{H}, s_i \in S\},$$

Here is the growth function:

$$\forall m \in \mathbb{N}, \Pi_{\mathcal{H}}(m) = \max_{S \sim \mathcal{D}^m} |\mathcal{H}_S| \leq \left(\frac{em}{d}\right)^d$$

Then, the VC-dimension is:

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

Using Massart Lemma:

$$R_m(\mathcal{H} \circ S) = \mathbb{E}_{\vec{\sigma}}\left[\sup_{h_S \in \mathcal{H}_S} \frac{1}{m}\langle \vec{\sigma}, h_S \rangle\right] \leq \sqrt{m}\frac{\sqrt{2\log(\Pi_{\mathcal{H}})}}{m} \leq \sqrt{\frac{2\log(\Pi_{\mathcal{H}})}{m}} \leq \sqrt{\frac{2d\log(em/d)}{m}}$$

26.3 GENERALIZATION BOUNDS FOR SVM