Modern Topics in Statistical Learning Theory
Lecture 4 Notes

Joseph Cappadona, Joseph Stansil, Boyuan Zhang

February 17, 2023

1 Introduction

Recap Last week we showed that uniform convergence implies generalization. That is, over the data distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$,

$$\sup_{\theta\in\Theta}\left[L(\theta)-\hat{L}_n(\theta)\right]\leq\epsilon \text{ w.h.p } \Rightarrow \text{Ł}(\hat{\theta})-L(\theta^*)\leq 2\epsilon$$

where

$$\hat{\theta}\leftarrow \text{argmin}_{\theta\in\Theta}\hat{L}_n(\theta)$$
$$\theta^*\leftarrow \text{argmin}_{\theta\in\Theta}L(\theta)$$

Additionally, we showed convergence for (1) a finite hypothesis class $\mathcal{H}$

$$|L(h)-\hat{L}_n(h)|\leq\tilde{O}\left(\sqrt{\frac{\log|\mathcal{H}|}{n}}\right)$$

and (2) a bounded p-dimensional l_2 ball defined by $\mathcal{H}=\{h\mid h_0,\|\theta\|_2\leq B\}$

$$|L(h)-\hat{L}_n(h)|\leq\tilde{O}\left(\sqrt{\frac{p}{n}}\right)$$

where $\tilde{O}(g(n))$ is equivalent to $O(g(n)\log^k n)$ for some k.

This week For a more general hypothesis class, we first seek a weaker result

$$\mathop{\mathbb{E}}_{S}\left[\sup_{h\in\mathcal{H}}L(h)-\hat{L}_n(h)\right]\leq \text{upper bound}$$

where $S\coloneqq\{(x_i,y_i)\}_{i=1}^n$. This result is weaker because we are proving a bound for an expectation over samplings of our data rather than guaranteeing a bound for all possible samplings.

2 Rademacher complexity

Definition 1 ((average) Rademacher complexity)**.** *Let $\mathcal{F}$ be a family of functions mapping $Z \to \mathbb{R}$, and let $\mathcal{P}$ be a distribution over Z.*

The (average) Rademacher complexity of $\mathcal{F}$ is

$$\Re_n(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{z_1,\ldots,z_n \overset{iid}{\sim} \mathcal{P}} \left[\mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right] \right]$$

where σ_i are Rademacher random variables

$$\sigma_i = \begin{cases} 1 & w.p. \ 1/2 \\ -1 & w.p. \ 1/2 \end{cases}$$

Notice that when $|\mathcal{F}| = 1$, the inner supremum will be close to 0, but when $|\mathcal{F}|$ is large, we expect $\Re_n(\mathcal{F})$ to also be large.

Theorem 1.

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n \overset{iid}{\sim} \mathcal{P}_{\mathcal{X},\mathcal{Y}}} \left[\sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} f(z_i) - \mathop{\mathbb{E}}_{z \sim \mathcal{P}_{\mathcal{X},\mathcal{Y}}} f(z) \right] \right] \leq 2\Re_n(\mathcal{F})$$

Applied to our setting, we have

$$\mathcal{F} = \{ z = (x, y) \mapsto \ell(h(x), y) \in \mathbb{R}, h \in \mathcal{H} \} \subseteq \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$$
$$\frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i) =: \hat{L}_n(h)$$

Corollary 1.

$$\mathop{\mathbb{E}}_{z_1,\ldots,z_n \overset{iid}{\sim} \mathcal{P}_{\mathcal{X},\mathcal{Y}}} \left[\sup_{h \in \mathcal{H}} \hat{L}_n(h) - L(h) \right] \leq 2\Re_n(\mathcal{F})$$

However, in practice we don't know the true distribution $\mathcal{P}_{\mathcal{X},\mathcal{Y}}$, therefore, we formulate the empirical Rademacher complexity:

Definition 2 (empirical Rademacher complexity)**.**

$$\Re_s(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{\sigma_1,\ldots,\sigma_n} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right]$$

for $S := \{(x_i, y_i)\}_{i=1}^{n}$

Accordingly, the relationship between the average and empirical Rademacher complexities is

$$\Re_n(\mathcal{F}) = \mathop{\mathbb{E}}_{S} \Re_s(\mathcal{F})$$

Theorem 2. *Suppose $\forall f \in \mathcal{F}$, $\forall z \in Z$, $0 \le f(z) \le 1$, then with probability $\ge 1 - \delta$,*

$$\sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} f(z_i) - \mathbb{E}\, f(z) \right] \le 2\Re_s(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}$$

This states that the difference between the empirical and population risks is bounded by the empirical Rademacher complexity, and we can derive a corresponding generalization bound.

Proof. (1) Recall the bounded difference condition and its implication, McDiarmid's inequality:

$$|g(x_1, ..., x_n) - g(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_n)| \le c_i \quad \forall x_1, ..., x_n, x_i'$$
$$\Rightarrow P(g(x_1, ..., x_n) - \mathbb{E}[g(x_1, ..., x_n)] \ge t) \le \exp(-\frac{-2t^2}{\sum_{i=1}^{n} c_i^2})$$

Now, define $g(z_1, ..., z_n) \triangleq \sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1}^{n} f(z_i) - \mathbb{E}\, f(z) \right]$. Notice that this definition of g satisfies the bounded difference condition with $c_i = \frac{1}{n}$. This is because $0 \le f(\cdot) \le 1$, so changing z to z_i' will change $f(\cdot)$ by at most 1, and this value is then scaled by a factor of $\frac{1}{n}$.

(2) Applying McDiarmid's inequality to g, we find

$$\begin{aligned} P(g(z_1, ..., z_n) \ge \mathop{\mathbb{E}}_{z}[g] + \epsilon) &\le \exp(-\frac{2\epsilon^2}{\sum_{i=1}^{n} c_i^2}) \\ &\le \exp(-2n\epsilon^2) \end{aligned}$$

(3) Now, by Theorem 1,

$$\mathop{\mathbb{E}}_{z_1, ..., z_n \overset{\text{iid}}{\sim} \mathcal{P}} g = \mathop{\mathbb{E}}_{z_1, ..., z_n \overset{\text{iid}}{\sim} \mathcal{P}} \left[\sup_{f \in \mathcal{F}} \left[\frac{1}{n} \sum_{i=1} n f(z_i) - \mathbb{E}\, f(z) \right] \right] \le 2\Re_n(\mathcal{F})$$

(4) Next, we will connect $\Re_n(\mathcal{F})$ to $\Re_s(\mathcal{F})$. Define

$$\tilde{g}(z_1, ..., z_n) = \Re_s(\mathcal{F}) \triangleq \mathop{\mathbb{E}}_{\sigma_i} \left[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(z_i) \right]$$

$\tilde{g}$ also satisfies the bounded condition with $c_i = \frac{1}{n}$ for the same reason as g. Thus,

$$P(\tilde{g}(z_1, ..., z_n) - \mathbb{E}[\tilde{g}] \ge \epsilon) \le \exp(-2n\epsilon^2)$$

or, equivalently,

$$P(\tilde{g}(z_1, ..., z_n) - \mathbb{E}[\Re_n(\mathcal{F})] \ge \epsilon) \le \exp(-2n\epsilon^2)$$

This shows that the difference between the average and empirical Rademacher complexities is tightly bounded and follows from the boundedness of f.

(5) Finally, if we set $\exp(-2n\epsilon^2) = \delta/2$, we get

$$\begin{aligned}
g &\coloneqq \sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f(z)]\right] \\
&\leq \mathbb{E}[g] + \epsilon && \text{(from (2), w.p. } 1-\tfrac{\delta}{2}) \\
&\leq 2\mathfrak{R}_n(\mathcal{F}) + \epsilon && \text{(from Theorem 1)} \\
&\leq 2(\mathfrak{R}_s(\mathcal{F}) + \epsilon) + \epsilon && \text{(from (4), w.p. } 1-\tfrac{\delta}{2}) \\
&= 2\mathfrak{R}_s(\mathcal{F}) + 3\epsilon
\end{aligned}$$

Thus, with probability $1-\delta$,

$$\begin{aligned}
\sup_{f\in\mathcal{F}} \left[\frac{1}{n}\sum_{i=1}^{n} f(z_i) - \mathbb{E}[f(z)]\right] &\leq 2\mathfrak{R}_s(\mathcal{F}) + 3\epsilon \\
&= 2\mathfrak{R}_s(\mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}
\end{aligned}$$

□

3 Properties of $\mathfrak{R}_n(\mathcal{F})$

Translation invariant

$$\mathfrak{R}_n(\mathcal{F}'_c) = \mathfrak{R}_n(\mathcal{F}) \quad \text{for all } c\text{, where } \mathcal{F}'_c \coloneqq \{f'(z) = f(z) + c \mid f \in \mathcal{F}\}$$

Reflection invariant

$$\mathfrak{R}_n(\mathcal{F}) = \mathfrak{R}_n(-\mathcal{F}) \quad \text{where } -\mathcal{F} \coloneqq \{-f \mid f \in \mathcal{F}\}$$

4 Examples of $\mathfrak{R}_n(\mathcal{F})$

4.1 Linear function

For some constant $B > 0$, let

$$\mathcal{H} \coloneqq \{x \to (w, x) \mid w \in \mathbb{R}^d, \|w\|_2 \in B\},$$

then

$$\mathfrak{R}_s(\mathcal{H}) \leq \frac{B}{n}\sqrt{\sum_{i=1}^{n} \|x_i\|^2}.$$

Moreover, if $\mathbb{E}_{x\sim P}[\|x\|_2^2] \leq C^2$, where P is some distribution and $C > 0$ is a constant, then

$$\mathfrak{R}_n(\mathcal{H}) \leq \frac{BC}{\sqrt{n}}$$

4.2 Two-layer neural network

Consider $f_\theta(x) \coloneqq \langle w, \phi(\mu_x) \rangle$ where $w \in \mathbb{R}^m$ and $\phi(\mu_x) \in \mathbb{R}^{m \times d}$.
For some constants $B_w > 0$ and $B_\mu > 0$, let

$$\mathcal{H} \coloneqq \{f_\theta \,|\, \|w\|_2 \leq B_w, \|\mu_i\|_2 \leq B_\mu, \forall i \in \{1, 2, \dots, m\}\},$$

and suppose $\mathbb{E}[\|x\|^2] \leq C^2$, then

$$\Re_n(\mathcal{H}) \leq 2B_w B_\mu C \sqrt{\frac{m}{n}}$$

For more details, please refer to (Du, Lee, 2017).

4.3 Deep neural network

Suppose that $\forall i, \|x^{(i)}\|_2 \leq 2$ and let

$$\tilde{\mathcal{F}} \coloneqq \{f_\theta \,:\, \|W_i\|_{op} \leq k_i, \|W_i^T\|_{2,1} \leq b_i\}.$$

Then

$$\Re_s(\mathcal{F}) \leq \frac{c}{\sqrt{n}} \cdot \left(\prod_{i=1}^{r} k_i\right) \cdot \left(\sum_{i=1}^{r} \frac{b_i^{\frac{2}{3}}}{k_i^{\frac{2}{3}}}\right)^{\frac{3}{2}}.$$

For more details, please refer to (Bartlett et al., 2017).