泛化理论

第二章 一致收敛

§2.3.1 Rademacher Complexity

@ 滕佳烨

[ref] Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: a statistical viewpoint. Acta numerica, 30, 87-201.
[ref] Shalev-Shwartz, S., & Ben-David, S. (2014). Understanding machine learning: From theory to algorithms. Cambridge university press.

Recall:

Uniform convergence: Decouple the dependency via "sup" over function class.

$$L(\hat{f}) - \hat{L}(\hat{f}) \leq \sup_{f \in \mathcal{F}} \left| L(f) - \hat{L}(f) \right|.$$

Today's topic:

Rademacher Complexity: symmetric matters!

In the following of this chapter, we abuse the notation $f = l(f)$ (loss function).
Denote $\widehat{\mathbb{E}} f = \hat{L}(f)$ as the training error and $\mathbb{E} f = L(f)$ as the test error.

Rademacher complexity

We consider the set $\mathcal{F} \circ S = \{(f(z_1), \dots, f(z_n)) : f \in \mathcal{F},\ \{z_i\} = S\}$, then

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

where σ is Rademacher Random Variable $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = 1/2$.

Intuitively, for any noise σ, considering the best function f in the function class $\mathcal{F}$, how much does it fit the noise?

For a complex function class $\mathcal{F}$, it would have more ability to fit the noise → large Rademacher complexity.

More generally, we can define Rademacher complexity on a set $A = \{a\}$.

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

Rademacher complexity and Uniform Convergence

Uniform Convergence: $\mathbb{E}f - \widehat{\mathbb{E}}f \leq \sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right|$

Rademacher complexity: $R(\mathcal{F}\circ S) \coloneqq \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \sigma_i f(z_i)\right]$.

Theorem: with prob $\geq 1-\delta$ (prob over datasets S),

$$\frac{1}{2}\mathbb{E}_S R(\mathcal{F}\circ S) \leq \mathbb{E}_S \sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| \leq 2\mathbb{E}_S R(\mathcal{F}\circ S).$$

Note that the lower bound requires $\mathbb{E}f = 0$, or centered version $\frac{1}{2}\mathbb{E}_S R\left(\bar{\mathcal{F}}\circ S\right)$, $\bar{\mathcal{F}} = \{f - \mathbb{E}f\}$. And the concentration results:

$$\mathbb{E}_S \sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| - \sqrt{\frac{\log\frac{2}{\delta}}{2n}} \leq \sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| \leq \mathbb{E}_S \sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}.$$

Therefore, if $\mathbb{E}_S R(\mathcal{F}\circ S) \to 0$, we have $\sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| \to 0$ (**sufficient condition!**).

Take-away messages

(a) Rademacher Complexity $R(\mathcal{F} \circ S) := \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}} \sum_{i=1}^n \sigma_i f(z_i)\right]$

(b) Measuring how function class fit the noise.

(c) If $\mathbb{E}_S R(\mathcal{F} \circ S) \to 0$, we have $\sup_{f\in\mathcal{F}} \left|\mathbb{E}f - \widehat{\mathbb{E}}f\right| \to 0$

All the slides will be available at www.tengjiaye.com/generalization soon.

@ 滕佳烨

Thanks!