泛化理论

第三章 Stability

§3.1.2 Stability-based Bound (proof)

@ 滕佳烨

[ref] Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *The Journal of Machine Learning Research*, *2*, 499-526.
[ref] Hardt, M., Recht, B., & Singer, Y. (2016, June). Train faster, generalize better: Stability of stochastic gradient descent. In *International conference on machine learning* (pp. 1225-1234). PMLR.

Recall:

Algorithmic stability: similar dataset returns similar models

If $\mathcal{D} = \{z_1, \cdots, z_n\}$ and $\mathcal{D}' = \{z_1', \cdots, z_n\}$ differs with only one samples, and the algorithm $\mathcal{A}$ satisfies:

$$\sup_z \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(\mathcal{D}); z) - \ell(\mathcal{A}(\mathcal{D}'); z)] \leq \epsilon,$$

then the algorithm is stable.

Theorem (stability and generalization). If the algorithm $\mathcal{A}$ is ϵ-Uniform-stable, its expected generalization bound (on parameter $\hat{\beta} = \mathcal{A}(D)$) satisfies

$$\mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L \left[L(\hat{\beta}) - \hat{L}(\hat{\beta})\right] \leq \epsilon,$$

where $\mathbb{E}_L$ denotes the expectation on testing point, and $\mathbb{E}_D$ denotes the expectation on training samples.

We will prove the theorem in this section.

Theorem (stability and generalization). If the algorithm $\mathcal{A}$ is ϵ-Uniform-stable,

$$\sup_z \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(\mathcal{D}); z) - \ell(\mathcal{A}(\mathcal{D}'); z)] \leq \epsilon,$$

its expected generalization bound (on parameter $\hat{\beta} = \mathcal{A}(D)$) satisfies

$$\mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L \left[L(\hat{\beta}) - \hat{L}(\hat{\beta}) \right] \leq \epsilon,$$

where $\mathbb{E}_L$ denotes the expectation on testing point, and $\mathbb{E}_D$ denotes the expectation on training samples.

Proof sketch: To go from training error to test error, we need to adjust the evaluated sample in the training set to another example, this causes ϵ loss.

Informally, the training error is $\mathbb{E}_{z_1} \ell(\mathcal{A}(\{z_1, \ldots z_n\}); z_1) = \mathbb{E}_{z_1'} \ell(\mathcal{A}(\{z_1', \ldots z_n\}); z_1')$, and the last equation is close to $\mathbb{E}_{z_1, z_1'} \mathbb{E}_{\mathcal{A}} \ell(\mathcal{A}(\{z_1, \ldots z_n\}); z_1')$, causing an ϵ loss on the test loss. We omit some expectation dependency in the above discussion.

Theorem (stability and generalization). If the algorithm $\mathcal{A}$ is ϵ-Uniform-stable,

$$\sup_z \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(\mathcal{D}); z) - \ell(\mathcal{A}(\mathcal{D}'); z)] \leq \epsilon,$$

its expected generalization bound (on parameter $\hat{\beta} = \mathcal{A}(D)$) satisfies

$$\mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L \left[L(\hat{\beta}) - \hat{L}(\hat{\beta})\right] \leq \epsilon,$$

where $\mathbb{E}_L$ denotes the expectation on testing point, and $\mathbb{E}_D$ denotes the expectation on training samples.

Use some different notations to make the proof clearer.

Let $D = \{z_1, \dots, z_n\}$ denote the training set, and $D' = \{z_1', \dots, z_n'\}$ denote the test set.

The trained parameter is then $\mathcal{A}(D)$. The expected training error is

$$\mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L \hat{L}\big(\mathcal{A}(D)\big) = \mathbb{E}_{D,\mathcal{A}} \frac{1}{n} \sum_i \ell(\mathcal{A}(\{z_1, \dots, z_n\}); z_i)$$

$$= \mathbb{E}_{D,D',\mathcal{A}} \frac{1}{n} \sum_i \ell(\mathcal{A}(\{z_1', \dots, z_n\}); z_i') \leq \mathbb{E}_{D,D',\mathcal{A}} \frac{1}{n} \sum_i \ell\big(\mathcal{A}(\{z_1, \dots, z_n\}); z_i'\big) + \epsilon$$

$$= \mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L L\big(\mathcal{A}(D)\big) + \epsilon$$

Take-away messages

Theorem (stability and generalization). If the algorithm $\mathcal{A}$ is ϵ-Uniform-stable,

$$\sup_z \mathbb{E}_{\mathcal{A}}[\ell(\mathcal{A}(\mathcal{D}); z) - \ell(\mathcal{A}(\mathcal{D}'); z)] \leq \epsilon,$$

its expected generalization bound (on parameter $\hat{\beta} = \mathcal{A}(D)$) satisfies

$$\mathbb{E}_{D,\mathcal{A}} \mathbb{E}_L \left[L(\hat{\beta}) - \hat{L}(\hat{\beta})\right] \leq \epsilon,$$

where $\mathbb{E}_L$ denotes the expectation on testing point, and $\mathbb{E}_D$ denotes the expectation on training samples.

Key idea in proof: training error - change one point - stability - test error.

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

@ 滕佳烨

Thanks!