泛化理论

第三章 Stability

§3.1.1 Stability-based Bound

@ 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:

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

Bad news: not algorithm-dependent; a loose bound…

Today's topic:
algorithmic stability: another approach to generalization

Algorithmic Stability

Intuition on stability: similar datasets return similar models

ϵ-Uniform-stability (not the unique definition):

If $\mathcal{D} = \{z_1, \cdots, z_n\}$ and $\mathcal{D}' = \{z_1', \cdots, z_n\}$ differs with only one sample, 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.

Note: the algorithm takes dataset as input and returns estimated parameter (or, model).

1. Sup on z: for any testing point → uniform
2. Expectation on $\mathcal{A}$ → random initialization, random operation …
3. "similar" dataset → the dataset has only one different example
4. "similar" model → the model has similar *loss* on any testing point

Stability and Generalization

One sentence: algorithmic stability leads to generalization bound

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

Note that different from previous analysis, here we use expectation form $\mathbb{E}_D$ instead of the high probability bound (with high probability, $\mathbb{E}_L\left[L(\hat{\beta}) - \hat{L}(\hat{\beta})\right] \leq \epsilon$).
We will talk about this in the following sections.

The proof will be shown in 3.1.2.

Take-away messages

(a) Algorithmic stability: similar dataset returns similar models
(b) Stability implies generalization (ϵ-uniform-stability)

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

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

@ 滕佳烨

Thanks!