泛化理论

序章 泛化理论简介

§0.3.2 PAC Learning

@ 滕佳烨

[ref] Understanding Machine Learning: From Theory to Algorithms, Shai Shalev-Shwartz and Shai Ben-David (2014)

Recall:

No-free-lunch: there is no universal learner!

For any algorithm A, there exists an adversarial distribution D, such that although there exists a good predictor, with high prob, we cannot find it.

Insight: we need some prior knowledge!

How to get the prior knowledge?
PAC learning (Provably Approximately Correct)!

PAC learning (Provably Approximately Correct)!

- Probably: for any distribution satisfying the prior,
- Approximately: with small error,
- Correct: we can learn the “best” predictor.

DEFINITION 3.1 (PAC Learnability) A hypothesis class $\mathcal{H}$ is PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$, for every distribution $\mathcal{D}$ over $\mathcal{X}$, and for every labeling function $f : \mathcal{X} \to \{0,1\}$, if the realizable assumption holds with respect to $\mathcal{H}, \mathcal{D}, f$, then when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$ and labeled by f, the algorithm returns a hypothesis h such that, with probability of at least $1-\delta$ (over the choice of the examples), $L_{(\mathcal{D},f)}(h) \leq \epsilon$.

Recall: realizable: there exists $h \in \mathcal{H}$ such that $L_{(D,f)}(h) = 0$.

Prior knowledge: realizable assumption. The distribution includes those which can be realized by class $\mathcal{H}$ (not all the distribution).

Agnostic PAC learning

DEFINITION 3.3 (Agnostic PAC Learnability) A hypothesis class $\mathcal{H}$ is agnostic PAC learnable if there exist a function $m_{\mathcal{H}} : (0,1)^2 \rightarrow \mathbb{N}$ and a learning algorithm with the following property: For every $\epsilon, \delta \in (0,1)$ and for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, when running the learning algorithm on $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ i.i.d. examples generated by $\mathcal{D}$, the algorithm returns a hypothesis h such that, with probability of at least $1 - \delta$ (over the choice of the m training examples),

$$L_{\mathcal{D}}(h) \leq \min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h') + \epsilon.$$

No need for "realizable".

What is the prior knowledge under the agnostic PAC? The adaptive loss $\min_{h'} L_D(h')$.

Note that this loss corresponds to the function class $\mathcal{H}$.

Take-away messages

(a) PAC learning: introduce the prior knowledge via “realizable” over a function class.
(b) Agnostic PAC learning: introducing the prior knowledge via adaptive loss.
ALL the algorithms we study later fall into the PAC learning framework…

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

@ 滕佳烨

Thanks!