泛化理论

序章 泛化理论简介

§0.3.1 No-free-lunch

@ 滕佳烨

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

Recall:

Generalization: Measuring how model performs on *unseen* data.

Relationship between generalization gap and population risk.
Generalization gap v.s. excess risk (label noise).

Can we find an universal learner without any prior knowledge on the problem?

Can we find an universal learner?

- **Universal learner**: no prior knowledge, can be challenged by any tasks.

- Whether there ***exist*** a learning algorithm A and a training set size m, such that for ***every*** distribution $\mathcal{D}$, if A receives m i.i.d. samples from $\mathcal{D}$, there is a high chance that it outputs a predictor h that has a low (population) risk.

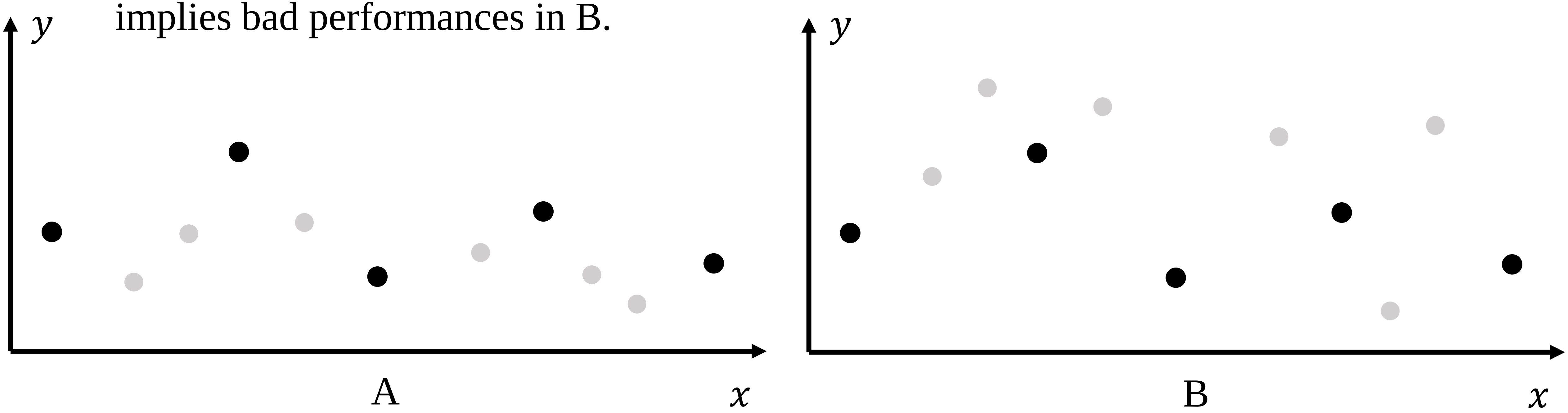
Intuitively, no.

Can we find an universal learner?

- Whether there ***exist*** a learning algorithm A and a training set size m, such that for ***every*** distribution $\mathcal{D}$, if A receives m i.i.d. samples from $\mathcal{D}$, there is a high chance that it outputs a predictor h that has a low (population) risk.

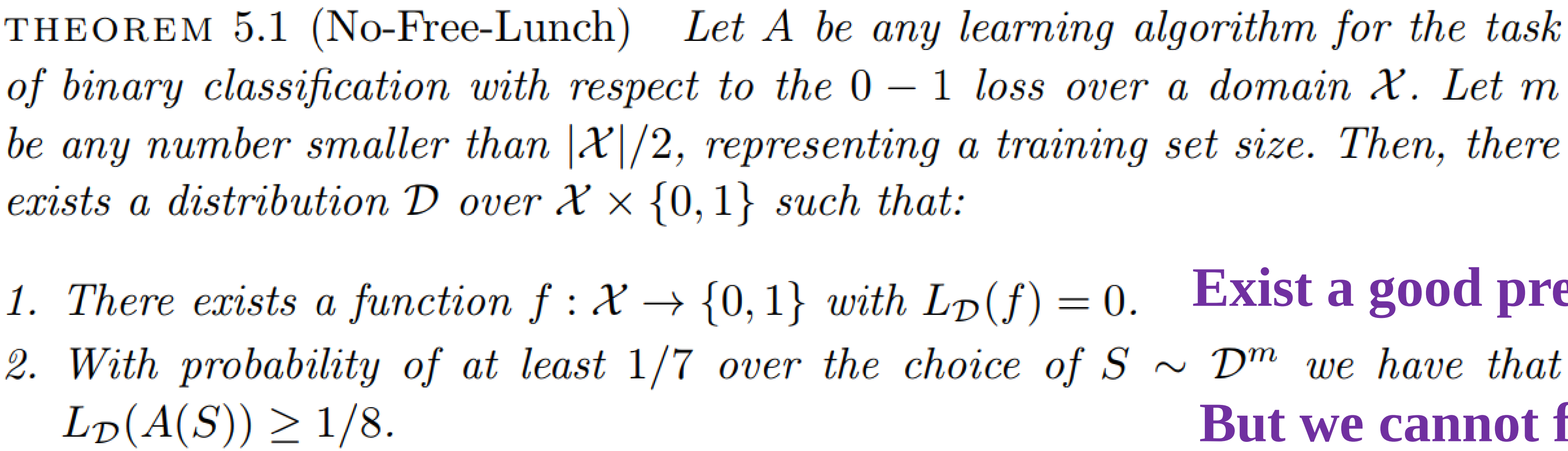
Intuitively, no.

- Hint: for any trained classifier, it can satisfy **either** A **or** B. Good performances in A implies bad performances in B.



Can we find an universal learner?

No-free-lunch Theorem: (one of the most basic and important theorems in generalization!)



THEOREM 5.1 (No-Free-Lunch) *Let A be any learning algorithm for the task of binary classification with respect to the $0-1$ loss over a domain $\mathcal{X}$. Let m be any number smaller than $|\mathcal{X}|/2$, representing a training set size. Then, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$ such that:*

1. *There exists a function $f : \mathcal{X} \to \{0,1\}$ with $L_{\mathcal{D}}(f) = 0$.*

Exist a good predictor

2. *With probability of at least $1/7$ over the choice of $S \sim \mathcal{D}^m$ we have that $L_{\mathcal{D}}(A(S)) \geq 1/8$.*

But we cannot find it universally

Hint:
find a worst D (distribution, correlated with f) with average S (training set) > find a best S with average D > For each S, for a fixed sample v not in S, it cannot predict well in both f_i and f_i' (they only differ in v), average over D leads to the results.

Take-away messages

(a) There exists no universal learner! (No-Free-Lunch Theorem)
(b) We need to do something to restrict the learning process, e.g., prior knowledge.
(c) What is prior knowledge? PAC (next lecture!)

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

@ 滕佳烨

Thanks!