No-free-lunch theorem for machine learning

Michikazu Hirata

September 1, 2025

Abstract

This entry is a formalization of the no-free-lunch theorem for machine learning following Section 5.1 of the book *Understanding Machine Learning: From Theory to Algorithms* [1] by Shai Shalev-Shwartz and Shai Ben-David. The theorem states that for binary classification prediction tasks, there is no universal learner, meaning that for every learning algorithms, there exists a distribution on which it fails.

Contents

1 No-Free-Lunch Theorem for ML

theory *No-Free-Lunch-ML*
imports
  HOL−Probability.Probability
begin

1.1 Preliminaries

lemma *sum-le-card-Max-of-nat*:*finite A*
$\Longrightarrow$ *sum f A* $\leq$ (*of-nat* :: - $\Rightarrow$ - ::{*semiring-1*,*ordered-comm-monoid-add*}) (*card A*) $*$ *Max* (*f* ‘ *A*)
  ⟨*proof*⟩

lemma *card-Min-le-sum-of-nat*: *finite A*
$\Longrightarrow$ (*of-nat* :: - $\Rightarrow$ - ::{*semiring-1*,*ordered-comm-monoid-add*}) (*card A*) $*$ *Min* (*f* ‘ *A*) $\leq$ *sum f A*
  ⟨*proof*⟩

The following lemma is used to show the last equation of the proof of the no-free-lunch theorem in the book [1].

Let A be a finite set. If A is divided into the pairs $(x_1, y_1), \ldots, (x_n, y_n)$ such that $f(x_i) + f(y_i) = k$ for all $i = 1, \ldots, n$. Then, we have $\sum_{x \in A} f(x) = k * |A|/2$.

lemma *sum-of-const-pairs*:
  fixes k :: *real*
  assumes *A*:*finite A*
    and $fst\ `\ B \cup snd\ `\ B = A$ $fst\ `\ B \cap snd\ `\ B = \{\}$
    and *inj-on fst B inj-on snd B*
    and *sum*: $\bigwedge x\ y.\ (x,y) \in B \Longrightarrow f\ x + f\ y = k$
  shows $(\sum x{\in}A.\ f\ x) = k * real\ (card\ A)\ /\ 2$
  ⟨*proof*⟩

lemma(**in** *prob-space*) *Markov-inequality-measure-minus*:
  assumes $u \in$ *borel-measurable M* **and** *AE x in M*. $0 \leq u\ x$ *AE x in M*. $1 \geq u\ x$
    and [*arith*]: $0 < (a{::}real)$
  shows $\mathcal{P}(x\ in\ M.\ u\ x > 1 - a) \geq ((\int x.\ u\ x\ \partial M) - (1 - a))\ /\ a$
⟨*proof*⟩

1.2 No-Free-Lunch Theorem

In our implementation, a learning algorithm of binary clasification is represented as a function $A : nat \Rightarrow (nat \Rightarrow {}'a \times bool) \Rightarrow {}'a \Rightarrow bool$ where the first argument is the number of training data, the second argument is the training data ($S\ n{=}\ (x_n, y_n)$ denotes the nth data for a training data S), and $A\ m\ S$ is a predictor. The first argument, which denotes the number of training data, is normally used to specify the number of loop executions in learning algorithm. In this formalization, we omit the first argument because we do not need the concrete definitions of learning algorithms.

Let X be the domain set. In order to analyze the error of predictors, we assume that each data (x, y) is obtained from a distribution $\mathcal{D}$ on $X \times \mathbb{B}$. The error of a predictor f with respect to $\mathcal{D}$ is defined as follows.

$$\begin{aligned} \mathcal{L}_{\mathcal{D}}(f) &\stackrel{\text{def}}{=} \mathop{\mathrm{P}}_{(x,y)\sim\mathcal{D}}(f(x) \neq y) \\ &= \mathcal{D}(\{(x, y) \in X \times \mathbb{B} \mid f(x) \neq y\}) \end{aligned}$$

In these settings, the no-free-lunch theorem states that for any learning algorithm A and $m < |X|/2$, there exists a distribution $\mathcal{D}$ on $X \times \mathbb{B}$ and a predictor f such that

- $\mathcal{L}_{\mathcal{D}}(f) = 0$, and
- $\displaystyle \mathop{\mathrm{P}}_{S\sim\mathcal{D}^m}\left(\mathcal{L}_{\mathcal{D}}(A(S)) > \frac{1}{8}\right) \geq \frac{1}{7}$.

theorem *no-free-lunch-ML*:
fixes X :: $'a$ *measure* **and** m :: *nat*
and A :: $(nat \Rightarrow 'a \times bool) \Rightarrow 'a \Rightarrow bool$
assumes *X1*:*finite* (*space* X) $\Longrightarrow 2 * m <$ *card* (*space* X)
and *X2*[*measurable*]:$\bigwedge x.\ x \in$ *space* $X \Longrightarrow \{x\} \in$ *sets* X
and m[*arith*]:$0 < m$
and A[*measurable*]: $(\lambda(s,x).\ A\ s\ x) \in (PiM\ \{..<m\}\ (\lambda i.\ X \bigotimes_M$ *count-space* (*UNIV* :: *bool set*))) $\bigotimes_M X$
$\rightarrow_M$ *count-space* (*UNIV* :: *bool set*)
shows $\exists \mathcal{D}$ :: $('a \times bool)$ *measure*. *sets* $\mathcal{D}$ = *sets* ($X \bigotimes_M$ *count-space* (*UNIV* :: *bool set*)) $\wedge$
prob-space $\mathcal{D}$ $\wedge$
$(\exists f.\ f \in X \rightarrow_M$ *count-space* (*UNIV* :: *bool set*) $\wedge \mathcal{P}((x, y)$ *in* $\mathcal{D}.\ f\ x \neq y) = 0) \wedge$
$\mathcal{P}(s$ *in* $Pi_M\ \{..<m\}\ (\lambda i.\ \mathcal{D}).\ \mathcal{P}((x, y)$ *in* $\mathcal{D}.\ A\ s\ x \neq y) > 1\ /\ 8) \geq 1\ /\ 7$
$\langle proof \rangle$

end

References

[1] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.