On the Foundations of Machine Learning

KIAS Workshop 2025

Jinsook Kim

jkim2@kias.re.kr
Center for AI and Natural Sciences
KIAS

Introduction

We present a new theory of machine learning.

Main Idea

- Machines **learn** a function when they *succeed* in *computing* it.

When do machines succeed?

- Machines succeed when they satisfy both the *truth* condition and the *belief* condition.

Machine Learning as Successful Computation

- **Definition:** Machines **learn** a function when they *succeed* in *computing* it.

  => *Successful* Computation vs. *Mere* Computation

- Machines *succeed* when they compute it *without fail*.

  => According to Gödel (1992), computation is achieved *without fail* whenever *axiomatic proof* is provided in a rich system.

Successful Computation and ML

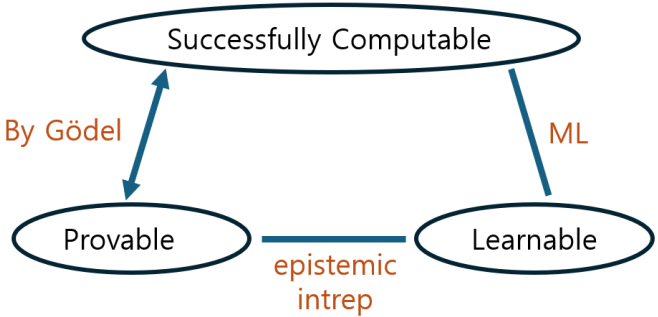
Theorem

*There is a Turing machine M that can **compute** y for the function $f(x_1, x_2, \cdots, x_n)$ when the machine receives $(x_1, x_2, \cdots, x_n)$ as input **if and only if** the mathematical formula $f(x_1, x_2, \cdots, x_n) = y$ is **provable** in a rich system S. (Gödel 1992)*

- Inspired by Gödel 1992 and Feferman 2006, we epistemically interpret successful computation as "machine learning".

  => *Successful* computation leads to *machine learning*

Successfully Computable and Learnable



"With the concept of computability, Turing has succeeded in giving an absolute definition of an epistemological notion" (Gödel 1995)

Successful vs. Mere Computation

- **Definition:** Machines *merely* compute a function when they *happen to* compute it under some crucial *assumption* in the system (without proof)

  (ex) **i.i.d.** assumption (e.g. Vapnik 2000 or Valiant 1984)

  => Kim 2024 proves under what condition machines learn the true probability without relying on the **i.i.d.** assumption.

The Success Criterion for Machines to Learn: Two Conditions

- 1. **Truth** Condition: machines are *correct* most of the time (not necessarily perfectly correct)
- 2. **Belief** Condition: the machines are *self-assured* whenever they indeed satisfy the truth condition

=> Under these conditions, machines attain computational success and thus learning.

Machine Learning on the True Probability

- **Definition:** A **true probability** $P(A_{t+1}|\beta_t)$ is what collectively constitutes a probability space, a triple $(\Omega, \mathscr{F}, p)$ of a joint true probability p by the stochastic data-generating process S_t's.

Theorem

If *machines learn the true probability,* ***then*** $\Pi(A_{t+1}|\beta_t) = P(A_{t+1}|\beta_t)$. *(Kim 2024)*

- If machines **learn**, then what they compute must be **true** to our world.

Truth Condition for Learning the True Probability

Truth Condition

- **Definition:** Machines are (precisely) **correct** when $\Pi(A_{t+1}|\beta_t) = P(A_{t+1}|\beta_t)$
- Machines satisfy the **truth condition** when $\limsup_{t\to\infty} |\Pi(A_{t+1}|\beta_t) - P(A_{t+1}|\beta_t)| < \epsilon$, $\forall \epsilon > 0$.

  => Machines **cannot** be said to learn if they are **wrong** too (= infinitely) often.
  => This is the usual *consistency condition* in learning theory

First Motivation for the Truth Condition

- Why does the truth condition need to be **asymptotic** ?

  => When machines **accidentally happen** to return correct computation once or twice, machines are **not** said to learn.

  - (ex) a broken clock analogy

Theorem

Consider any machine forecast $\alpha \in \Re[0,1]$. If $P(A_{t+1}|ß_t) \neq \alpha$ at least for infinitely many t's with $t \geq n$ for some $n < \infty$ along the stochastic path, then $P(A_{t+1}|ß_t) = \alpha$ at some $t^* < n$ is not equivalent to learning the true probability at t^*. (Kim and Kang 2024)

Second Motivation for the Truth Condition

- Why does the truth condition need to be **precisely correct** ?

  => If machines **cannot** learn **precisely**, machines **cannot** learn **approximately** either.

  (cf) floating point for π

Theorem

Suppose that no machines can learn the true probability function P, denoted f^*_∞, after processing training samples $\{X_i\}_{i=1}^{\infty}$. Then, no machines can assess any finite learning rule $f_{n,m}$ (for $n, m < \infty$) as either a good or bad approximation to the true probability f^*_∞. (Kim and Kang 2024)

What is Approximation?

- **Definition:** Machines do an **approximation** on the true target function f^* by a **learning rule** $f_{n,m}$ when machines **compute** the **distance function** $\ell(f_{n,m}(\cdot), f^*) < \delta$ for a given small bound $\delta < \infty$.

- **Definition:** A **distance function** is a function $\ell : \Psi^2 \to [0, \infty)$ such that

$$\ell\Big(f_{n,m}(\{X_i\}_{i=1}^m, f^*(\{X_i\}_{i=1}^n), \{X_i\}_{i=m+1}^\infty), f^*(\{X_i\}_{i=1}^\infty)\Big).$$

What is Approximation? (Continued)

- **Definition:** A **learning rule** is any sequence of measurable functions, $f_{n,m}$'s, provided by machines, where each

$$f_{n,m} : \Xi^m \times \Psi^n \times \Xi \to \Psi$$

depends on the following data:

(1) the initial training samples of labeled data $\{(X_i, Y_i)\}_{i=1}^{n}$
(2) the unlabeled data $\{X_i\}_{i=n+1}^{m}$
(3) the random variable X_{m+1} for prediction, given that $m > n$ for $n, m \in \mathbb{N} \cup \{0\}$.

What is Approximation? (Continued)

- There are various measures to **approximate** the **true** probability P.

  (1) Variational Distance: $V_P(Q) = sup\{|P(A) - Q(A)| : A \in \sigma(\Sigma)\}$

  (2) Brier Score: $B_P(Q) = \sum_{\sigma \in \Sigma}(P(\sigma) - Q(\sigma))^2$

  (3) Hellinger Distance: $H_P(Q) = \sum_{\sigma \in \Sigma}(P(\sigma) + Q(\sigma) - 2(P(\sigma)Q(\sigma))^{\frac{1}{2}})$

  (4) Kullback-Leibler Divergence: $D_{KL}(P||Q) = \sum_{\sigma \in \Sigma} P(\sigma) log(\frac{P(\sigma)}{Q(\sigma)})$

Belief Condition for Learning the True Probability

Belief Condition

- Machines satisfy the **belief condition** when
  $$\Pi\,(\limsup_{t\to\infty} |\Pi(A_{t+1}|\beta_t) - P(A_{t+1}|\beta_t)| < \epsilon,\ \forall \epsilon > 0) = 1$$
  whenever the true condition is satisfied.

=> **Definition:** Machines *tolerate* an *error* when
$$\Pi(A_{t+1}|ß_t) = P(A_{t+1}|ß_t) \text{ but } \Pi(\{\Pi(A_{t+1}|ß_t) \neq P(A_{t+1}|ß_t)\}) > 0$$

=> Machines *cannot* be said to *learn* if they tolerate errors *too (= infinitely) often*.

First Motivation for the Belief Condition

- Why do machines need the *belief* condition to learn, in addition to the usual (truth) consistency condition?

  => We can construct a stochastic process whose subsequence satisfies the truth condition but whose true underlying true probability machines do not learn.

  (ex) A stochastic path where for any $n > N \in \mathbb{N}$, on $\lfloor \frac{n}{2} \rfloor$ number of the periods t's where $\{X_t = x\}$ occurs, $P(X_t = x|ß_{t-1}) = \alpha$, while $P(X_t = x|ß_{t-1}) = \beta$ on the rest of the periods for any $x \in \{x_1, \ldots, x_k\}$ with $\alpha \neq \beta$.

Second Motivation for the Belief Condition

- According to Lewis 1980, *subjective* probability should be bound by *objective* probability when the objective one is *known*.
- Likewise, *belief* condition should be bound by the *truth* condition when the truth condition is *proven*.

  => It must be that Π (the truth condition is satisfied) $= 1$ when the truth condition has been *proven*, *not* just *assumed*.
- In other words, machines should prove the truth condition without relying on the **i.i.d.** assumption, in order to learn the true probability.

Extending to ML on General Functions

- For machines to learn any *general* functions from *finite* samples, machines should extrapolate an infinite array of potential outcomes from a finite set of inputs, inevitably introducing *uncertainty*.

  => Learning any general functions involves probability
  => To learn any general function, machines learn the true probability first.

- The *truth condition* for machines to learn a function f^*:

  with true probability *P– one*, $\limsup_{m\to\infty} \ell\,(f_{n,m}(\{X_i\}_{i=1}^m, f^*(\{X_i\}_{i=1}^n), X_{m+1}), f^*(X_{m+1})) < \epsilon, \forall\, \epsilon > 0.$

References

- Feferman, S. (2006). “Are there Absolutely Unsolvable Problems? Godels Dichotomy.”. In: *Philosophia Mathemtica III*, pp. 1–19.
- Gödel, K. (1992). *On Formally Undecidable Propositions of Principia Mathematica and Related Systems*. New York: Dover Publications.
- — (1995). *Unpublished Essays and Lectures, Collected Works, Vol. III*. New York.
- Kim, Jinsook (2024). “Can machines learn true probabilities?” In: *International Conference on Machine Learning* PMLR 235.
- Kim, Jinsook and Jinho Kang (2024). “A Theory of Machine Learning”. In: *arXiv* 2407.05520 [cs.LG].
- Lewis, D. (1980). “A subjectivist’s guide to objective chance”. In: *Studies in Inductive Logic and Probability, Volume II, R. Jeffrey (ed.)*, pp. 63–293.
- Valiant, Leslie (1984). “A theory of the learnable”. In: *Communications of the ACM* 27, Nov. Pp. 1134–1142.
- Vapnik, Vladimir (2000). *The Nature of Statistical Learning Theory*. New York, NY: Springer.