When is Inductive Inference Possible? (Neurips 2024, spotlight)

Zhou Lu
Department of Computer Science, Princeton University

PRINCETON UNIVERSITY

A Philosophy Problem

There are two ways of human reasoning: deductive and inductive.

Deduction: conclusions follow necessarily from the stated premises (theorems proven under axioms and assumptions). Rigorous by logic.

Induction: general laws or axioms are formulates based on limited empirical observations (the evolution of physical models). No rigorousness is guaranteed.

The problem of induction is that extrapolations based on past experiences cannot reliably predict the unexperienced: there can always be black swans. The basic philosophy question we study is thereby

How to model induction, and what guarantees can it offer?

Background

The most representative mathematical abstraction of induction is **inductive inference**. In inductive inference, a learner aims to deduce the ground-truth hypothesis h^* from a hypothesis class $\mathcal{H}$ based on an infinite observation sequence $\{x_t\}$.

Any $h \in \mathcal{H}$ is a mapping $\mathcal{X} \to \{0,1\}$. At each round t, the learner makes a binary prediction y_t based on all previous information after observing x_t, then the true outcome $h^*(x_t)$ is revealed and the learner makes an error if $y_t \neq h^*(x_t)$.

Previous works such as [1][2] showed that when $|\mathcal{H}| \leq \aleph_0$, the learner can guarantee that only a finite number of errors are made. Clearly this is not a necessary condition: consider $\mathcal{H} = \{h_c | h_c(x) = 1_{x=c}, c \in \mathbb{R}\}$, we can always predict 0 until an error. We make a connection to online learning theory to provide a sufficient and necessary condition.

In online learning [3], instead of fixing h^*, Nature can change the ground-truth with time. We say a class $\mathcal{H}$ is online learnable if the learner can make at most m errors for some integer m depending only on $\mathcal{H}$. We can this number the Littlestone dimension $\textbf{Ldim}(\mathcal{H})$.

Classic online learning protocol

1: Given domain $\mathcal{X}$ and hypothesis class $\mathcal{H}$
2: **for** $t = 1, \ldots, \infty$ **do**
3: Nature presents observation x_t to Learner
4: Learner predicts y_t
5: **Nature selects a consistent** $h_t \in \mathcal{H}$
6: Nature reveals the true label $h_t(x_t)$
7: **end for**
8: **Goal: a uniform error bound**

Non-uniform Online Learning

We propose a new learning framework which subsumes previously considered inductive inference as special cases.

Non-uniform online learning protocol (inductive inference)

1: Given domain $\mathcal{X}$ and hypothesis class $\mathcal{H}$
2: **Nature selects ground-truth** $h^* \in \mathcal{H}$
3: **for** $t = 1, \ldots, \infty$ **do**
4: Nature presents observation x_t to Learner
5: Learner predicts y_t
6: Nature reveals the true label $h^*(x_t)$
7: **end for**
8: **Goal: error bound can depend on** h^*

It is a variation of online learning by (1) requiring Nature to fix a ground-truth h^* in advance, and (2) considering non-uniform error bounds depending on h^* (and $\mathcal{H}$).

Denote $err_{\mathcal{A}}(h, x) \triangleq \sum_{t=1}^{\infty} |\hat{y}_t - h(x_t)| \in \mathbb{N} \cup \infty$, as the number of errors made by a learning algorithm $\mathcal{A}$ when Nature chooses h and presents x to the learner, we define non-uniform online learnability:

Definition: We say a hypothesis class $\mathcal{H}$ is non-uniform online learnable, if there exists a deterministic learning algorithm $\mathcal{A}$, such that

$$\exists m : \mathcal{H} \to \mathbb{N}^+, \forall h \in \mathcal{H}, \forall x \in X, err_{\mathcal{A}}(h, x) \leq m(h).$$

Theorem (main): $\mathcal{H}$ is non-uniform online learnable iff $\mathcal{H}$ can be written as a countable union of online learnable classes.

Stochastic Observations

The previous setting makes no assumption on how Nature chooses x_t (it can choose adaptively). An easier (for the learner) setting is each x_t is drawn iid from some unknown μ fixed by Nature in advance.

Definition: We say a hypothesis class $\mathcal{H}$ is non-uniform stochastic online learnable, if there exists a deterministic learning algorithm $\mathcal{A}$, such that

$$\exists m : \mathcal{H} \to \mathbb{N}^+, \forall h \in \mathcal{H}, \forall \mu, \mathbb{P}_{x \sim \mu^\infty} \left(err_{\mathcal{A}}(h, x) \leq m(h) \right) = 1.$$

Theorem: $\mathcal{H}$ is non-uniform stochastic online learnable iff $\mathcal{H}$ can be written as a countable union of online learnable classes.

The two theorems together imply that for both the strongest (adaptive) and weakest (stochastic) Nature's choice on x, the characterization for learnability is the same, therein applies to any other setting in between.

The Agnostic Setting

Preciously we make the realizable assumption $h^* \in \mathcal{H}$. Here we consider the agnostic setting, in which we pose no constraint on how Nature presents y_t, with Learner's objective to be performing as good as the best hypothesis in $\mathcal{H}$ (regret minimization).

Definition: We say a hypothesis class $\mathcal{H}$ is agnostic non-uniform online learnable with rate $r(T)$, if there exists a learning algorithm $\mathcal{A}$, such that

$$\exists m : \mathcal{H} \to \mathbb{N}^+, \forall x \in X, \forall y \in \{0,1\}^\infty, \forall h \in \mathcal{H}, \forall T \in \mathbb{N}^+,$$

$$\mathbb{E}\left[\sum_{t=1}^{T} 1_{[\hat{y}_t \neq y_t]} - \sum_{t=1}^{T} 1_{[h(x_t) \neq y_t]} \right] \leq m(h) r(T).$$

The following trichotomy provides a complete characterization of agnostic non-uniform online learnability, including degenerate, typical, and arbitrarily slow rates.

Theorem: only three possible rates in the agnostic setting.

- $\mathcal{H}$ is learnable at rate $0 \iff |\mathcal{H}| = 1$.
- $\mathcal{H}$ is learnable at rate $\tilde{\Theta}(\sqrt{T}) \iff \mathcal{H}$ is a countable union of online learnable classes.
- $\mathcal{H}$ requires arbitrarily slow rates $\iff \mathcal{H}$ isn't a countable union of online learnable classes.

Proof Idea of the Main Result

Warm up: In the case $|\mathcal{H}| \leq \aleph_0$, we index the hypotheses and adopt a Bayesian approach. We predict w.r.t. the hypothesis which has the smallest index among all the hypotheses that are correct so far. Then if $h^* = h_n$, we make at most n errors.

Our algorithm for the main theorem is a natural generalization of this: we index each online learnable class in the countable partition. Every such class $\mathcal{H}_n$ will be excluded if the online learning algorithm run on it makes at least $\textbf{Ldim}(\mathcal{H}_n) + 1$ errors.

References

[1] E Mark Gold. Language identification in the limit. Information and control, 10(5):447–474, 1967.

[2] Ray J Solomonoff. A formal theory of inductive inference. part i. Information and control, 7(1):1–22, 1964a.

[3] Nick Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. Machine learning, 2:285–318, 1988.