The no-free-lunch theorems of supervised learning

Tom Sterkenburg
(joint work with Peter Grünwald, CWI/Leiden)

Deutsche Forschungsgemeinschaft

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
FAKULTÄT FÜR PHILOSOPHIE, WISSENSCHAFTSTHEORIE UND RELIGIONSWISSENSCHAFT
MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY

Progic, “Munich”
September 2021

The puzzle

- The **no-free-lunch theorems** of supervised learning suggest a *skeptical* conclusion about machine learning algorithms.
  - "All learning algorithms are equally lacking in epistemic justification."
  - "A standard procedure like empirical risk minimization is just as good as empirical risk *maximization*."

- At the same time, the business of **learning theory** is to show that some possible algorithms *are* better than others.
  - "We can *prove* that empirical risk minimization is a good method (and we couldn't for empirical risk *maximization*)."

- How can these claims co-exist?

The plan

1. An illustration and a reformulation.

2. The road to skepticism.

3. Data-only v. model-dependent.

The no-free-lunch (NFL) theorems

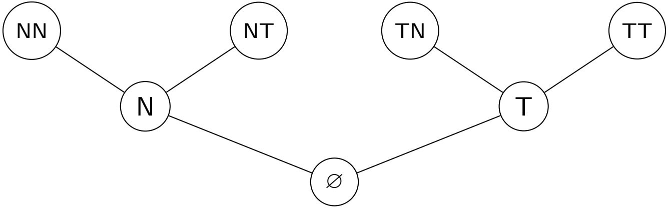
- ▶ Wolpert (1993,1996): "no free lunch theorems for supervised learning."
- ▷ "All learning algorithms are a priori equivalent."
- ▶ Schaffer (1994): "conservation law of generalization performance."

A (*very*) simple version

- Every day we try to predict whether our breakfast will be tasty (T), or not (N).
- Our **learning algorithm** makes a guess whether breakfast will be tasty today, based on the days past.

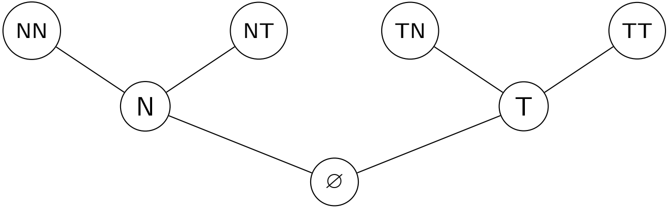
A (*very*) simple version

- Consider histories of two consecutive days.
- There are 2^2 such histories or **learning situations**.



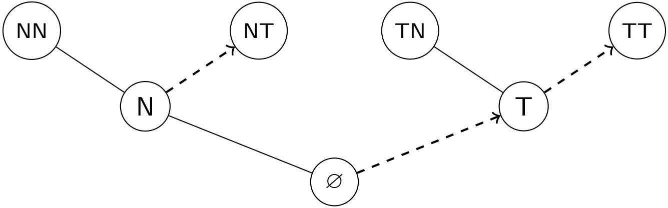
A (*very*) simple version

- Consider histories of two consecutive days.
- There are 2^2 such histories or **learning situations**.
- There are 2^3 different possible **learning algorithms** (functions from $\{\emptyset, \mathrm{T}, \mathrm{N}\}$ to $\{\mathrm{T}, \mathrm{N}\}$).



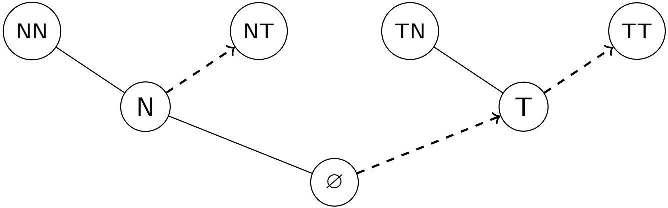
A (*very*) simple version

- Consider histories of two consecutive days.
- There are 2^2 such histories or **learning situations**.
- There are 2^3 different possible **learning algorithms** (functions from $\{\emptyset, \mathrm{T}, \mathrm{N}\}$ to $\{\mathrm{T}, \mathrm{N}\}$).



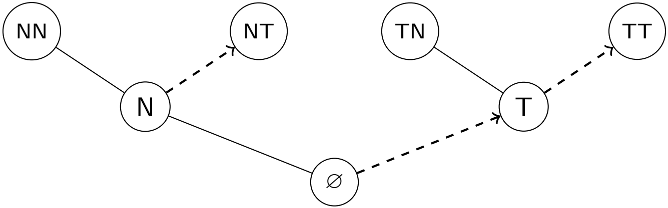
A (*very*) simple version

- ▷ A learning algorithm's **error** in a particular learning situation is its mean number of mistakes.
- ▶ Here, then, is an NFL statement: **every prediction algorithm attains the same error in *equally many* learning situations**.



A (*very*) simple version

- A learning algorithm's **error** in a particular learning situation is its mean number of mistakes.
- Here, then, is an NFL statement: **every prediction algorithm attains the same error in *equally many* learning situations**.
- Assume a *uniform* distribution on learning situations.
- Then we can say that **every learning method has the same expected error 1/2**.



A reformulation

- The assumption of a uniform distribution on learning situations is rather lacking in motivation.
- ▷ The same already holds for *counting* learning situations.
- In fact, this is, for the purpose of learning, really a *worst-case* assumption (cf. Peirce, Carnap, ...)
- ▷ "In a universe where learning is impossible, every learning algorithm is equivalent." Well, sure ...

- But this assumption is actually not essential for a skeptical conclusion...

A reformulation

- For every learning algorithm, there is a learning situation in which it is *not* successful, yet in which *another* learning algorithm *is* successful.
- There is no **universal** learning algorithm.
- ▷ Many modern formulations are of this form (e.g., Shalev-Shwartz & Ben-David, 2014).
- Every learning algorithm must come with some restrictive **inductive bias**.

The road to skepticism

- We are concerned with a limited set of standard, generic, algorithms.
- What justification do we have for these standard learning algorithms?
  - NFL: these algorithms must have specific biases.
  - So, how do we justify these biases..?
- Our universe must have a structure that happens to neatly match these biases...
  - E.g., Giraud-Carrier and Provost's (2005) "weak assumption of machine learning" that "the process that presents us with learning problems . . . induces a non-uniform probability distribution [over learning situations]."
  - OK, but how to justify such an assumption?

The road to skepticism

- Hume's argument for inductive skepticism.
  - Inductive reasoning must proceed upon the supposition that the universe is *induction-friendly*.
  - What reason can we give for this supposition?
  - We certainly cannot give any *deductive*, a priori reason, because it's logically possible that the universe is *not* induction-friendly.
  - But we also cannot give a good *inductive* reason, because that would be circular!
  - Specifically, we cannot conclude from the success of inductive method so far (past evidence for induction-friendliness) that inductive method will remain successful (that the universe is, in fact, induction-friendly).
- So we're stuck.

The road to skepticism

- Our universe must have a structure that happens to neatly match our standard algorithms' biases...
  - ▷ E.g., Giraud-Carrier and Provost's (2005) "weak assumption of machine learning."
  - ▷ OK, but how to justify such an assumption?
  - ▷ . . .
  - ▷ ?
- So we're stuck.

Data-only v. model-dependent

- Let's backtrack.
  - ▷ We don't want to have to defend some grand assumption that the universe is friendly to our machine learning algorithms.
  - ▷ We don't make such assumptions when we actually use machine learning methods...
- Rather, on each use of machine learning methods we rely on *local, context-dependent* factors.
- Even if we use generic machine learning methods, they must in each application still employ—*and thus be provided with*—local assumptions.

Data-only v. model-dependent

- The NFL theorems rely on a conception of learning algorithms as purely data-driven, as **data-only**.
- NFL: There is no universal *data-only* learning algorithm.
- Every *data-only* learning algorithm must come with some restrictive inductive bias.
- Given any such algorithm, we can expose its inductive bias, and question its justification.

Data-only v. model-dependent

- But many standard learning algorithms are better conceived of as **model-dependent**.
  - Such an algorithm does not only take input data, but on each application also requires for input a **model**.
  - On each application, the model represents the bias.
- Crucially, model-dependent algorithms can be given a **model-relative** justification.
- *This* is what learning theory, for many standard learning algorithms, gives us.

Example

- **Empirical Risk Minimization** is a function both of a training sample and of an **hypothesis class** $\mathcal{H}$, a set of classifiers.
  - Given a training sample S and a model $\mathcal{H}$, it returns a classifier that, **among the classifiers in** $\mathcal{H}$, minimizes the empirical error on S.

- A fundamental result of learning theory is that for any $\mathcal{H}$ (that is not too complex), $\mathrm{ERM}+\mathcal{H}$ will with arbitrarily high probability return a classifier that has error arbitrarily close to that of the *best* classifier in $\mathcal{H}$.
  - In contrast, empirical risk *maximization*, for given $\mathcal{H}$, returns with arbitrarily high probability a classifier that has error arbitrarily close to that of the *worst* classifier in $\mathcal{H}$.
- This gives us a model-relative justification for preferring ERM to anti-ERM.

Data-only v. model-dependent

- **Data-only**:
  - Must come with an *inherent* inductive bias.
  - Given any such proposed algorithm, we can expose its inductive bias, and question its justification.
- **Model-dependent**:
  - Itself a *generic* method, that on each application *we* must provide a model.
  - Can be given a model-relative justification, in the form of learning-theoretic guarantees.

To conclude: some caveats/nuances

- We haven't solved Hume's problem of induction.
- We haven't claimed that algorithms with a model-relative justification are *perfect*.
- Not all standard learning algorithms are straightforwardly model-dependent.
  - Nearest neighbor?
  - Neural networks..?

To conclude: take-home

The NFL results show that every *data-only* learning procedure must possess some inductive bias. But many standard learning algorithms are better conceived of as *model-dependent*, and can be given a general *model-relative* justification.

tom.sterkenburg@lmu.de