Epistemology and theory of machine learning

Part 2: Statistical learning theory and beyond

Tom Sterkenburg

CWI
Centrum Wiskunde & Informatica

Emmy
Noether-
Programm
DFG Deutsche
Forschungsgemeinschaft

LMU
LUDWIG-
MAXIMILIANS-
UNIVERSITÄT
MÜNCHEN

FAKULTÄT FÜR PHILOSOPHIE, WISSENSCHAFTSTHEORIE UND RELIGIONSWISSENSCHAFT
MUNICH CENTER FOR MATHEMATICAL PHILOSOPHY

HLRS Summer School Trust and ML
Stuttgart, July 2023

Recall: 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..?
- The world must have a structure that neatly matches 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]."
  - Analogous to Hume's "principle of the uniformity of nature."
- OK, but how to justify such an assumption..?
- So we're stuck.

The postulate of general induction-friendliness

- ► Let's backtrack.
- ▷ We don't want to have to defend some grand assumption that the world is friendly to induction / to our machine learning algorithms.

- ► Okasha (2001, 2005): this is the weak point in Hume's argument.
- ▷ Inductive inference does not rely on some universal uniformity assumption.
- ► Sober (1988, 2015): Hume commits a "quantifier-shift fallacy".
- ▷ There is no universal assumption that every inductive inference requires; rather, every inductive inference requires a specific "local" assumption.

General inductive rules

- Okasha: rejecting general postulates/notions of induction-friendliness, we must also reject the possibility of general rules for induction.
- This is also the core of Norton's (2003, 2021) "material theory of induction."

- But in machine learning theory, we *do* study generic rules for induction—generic learning algorithms.
- How to square this with the view that every inductive inference requires particular, *local*, assumptions?

- Even if we use generic machine learning methods, they must in each application still employ—*and thus be provided with*—local assumptions.

The plan

1. Model-relative justification.
2. Statistical learning theory.
3. Application: Occam's razor.
4. The "generalization puzzle": the way forward?

The plan

1. **Model-relative justification.**
2. Statistical learning theory.
3. Application: Occam's razor.
4. The "generalization puzzle": the way forward?

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 more naturally conceived of as explicitly **model-dependent**.
  - Such an algorithm does not only take input data, but on each application also requires for input an **inductive model**.
  - On each application, the inductive model represents the inductive 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 a **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.
- The fundamental theorem of learning theory says that for any $\mathcal{H}$ (that is sufficiently simple), $\mathrm{ERM}(\mathcal{H})$ 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 empirical risk maximization.

Summary: two conceptions of learning algorithms

- **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 general yet model-relative justification, in the form of learning-theoretic guarantees.

“Defensive epistemology”

- The original epistemological project is one where our concern is the (ultimate) basis of our knowledge.
  - This is a project of tracing the justificatory basis for a statement or belief of interest.
  - We ask, what is the basis for trusting what our learning algorithm returns?
  - Since the algorithm must have a particular inductive bias, we are lead to ask, what is the basis for this inductive bias?
  - Since no ultimate basis is forthcoming, we are led to skepticism.
- Many authors in the philosophy of science have arrived at a view that this general epistemological project is simply a dead end.
  - Van Fraassen speaks of “defensive epistemology,” Levi of “predigree epistemology.”

“Forwards-looking epistemology”

- The alternative project that arises is a pragmatist one that takes seriously that we will always start with a body of beliefs or presuppositions that we do not seriously or actively doubt.
- The interesting question is not whether we actually have an ultimate justification for these beliefs.
- The interesting question is how to *proceed* from these beliefs: how to *improve* these beliefs.
- This is still an interesting question of justification!

“Forwards-looking epistemology”

- Moreover, this project aligns well with (traditional) machine learning.
  - Russell (1991): “the picture that is currently fashionable in machine learning is that of an agent that *already knows something* and is trying to learn some more.”
  - Domingos (2012): “induction (what learners do) is a knowledge lever: it turns a small amount of input knowledge into a large amount of output knowledge.”
- The model-relative guarantees from learning theory serve exactly this project.
  - There are provably better and worse algorithmic ways of proceeding.
- Learning theory thus provides a normative component to a forwards-looking epistemological perspective on machine learning methods.

The plan

1. Model-relative justification.
2. **Statistical learning theory.**
3. Application: Occam's razor.
4. The "generalization puzzle": the way forward?

The framework of SLT

- We do *classification*.
  - For instance, we seek to learn whether our Käsespätzle will be tasty (T) or not (N).

  - We assume there is some unknown distribution $\mathcal{D}(\mathcal{X} \times \mathcal{Y})$ that governs the relation between instances in $\mathcal{X}$ (given by attributes like temperature, color, smell) and labels in $\mathcal{Y} = \{\mathsf{T}, \mathsf{N}\}$ (tasty or not).

The framework of SLT

- We do *classification*.
- For instance, we seek to learn whether our Käsespätzle will be tasty (T) or not (N).

- We assume there is some unknown distribution $\mathcal{D}(\mathcal{X} \times \mathcal{Y})$ that governs the relation between instances in $\mathcal{X}$ (given by attributes like temperature, color, smell) and labels in $\mathcal{Y} = \{\mathsf{T}, \mathsf{N}\}$ (tasty or not).
- We draw a *training sample S* from this unknown distribution $\mathcal{D}$.
- Based on the training sample, our algorithm learns a *classifier* or *hypothesis h* : $\mathcal{X} \rightarrow \{\mathsf{T}, \mathsf{N}\}$ that is a function from all possible Käsespätzles (combinations of attribute values) to labels.

The framework of SLT

- ▶ We want to say something about what makes for a good method—in a model-relative sense!
- ▶ In SLT, inductive bias enters in the form of a class $\mathcal{H}$ of hypotheses.
- ▷ So a learning algorithm is a function $A_{\mathcal{H}} : \mathcal{H} \times S \to \mathcal{H}$ from samples to hypotheses in given $\mathcal{H}$.
- ▶ The goal is to learn a hypothesis that is the (near) best in the class.
- ▷ That is, a hypothesis h that has a *true risk*

$$L_{\mathcal{D}}(h) = \mathrm{Prob}_{(x,y)\sim\mathcal{D}}\left[h(x) \neq y\right]$$

that is not much worse than that of the best in the class,

$$\min_{h' \in \mathcal{H}} L_{\mathcal{D}}(h').$$

Learnability

- We now formulate a guarantee of "probably-approximately-correct" (PAC) learnability.
  - Pick an accuracy term ϵ to bound the distance from the best hypothesis in the class.
  - Pick a confidence term δ to bound the probability of finding a near-best hypothesis.

Definition (Learnability)

A hypothesis class $\mathcal{H}$ is learnable if there exists a learning method $A_{\mathcal{H}} : \mathcal{S} \to \mathcal{H}$ and a sample size function $m_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ such that for all $\epsilon, \delta \in (0,1)$, for all $m \geq m_{\mathcal{H}}(\epsilon, \delta)$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

$$\mathrm{Prob}_{S \sim \mathcal{D}^m} \left[L_{\mathcal{D}}(A_{\mathcal{H}}(S)) \leq \min_{h \in \mathcal{H}} (L_{\mathcal{D}}(h)) + \epsilon \right] \geq 1 - \delta. \tag{1}$$

- What makes a hypothesis class learnable?

Uniform convergence

- The cornerstone of SLT—or the original Vapnik-Chervonenkis theory—is a **uniform** version of the law of large numbers.
  - Specifically, this is the convergence, *for each hypothesis simultaneously*, of the *empirical error* to the *true* risk.

Definition (Uniform convergence)

A hypothesis class $\mathcal{H}$ *has the uniform convergence property if there exists a sample size function* $m^{\mathrm{uc}}_{\mathcal{H}} : (0,1)^2 \to \mathbb{N}$ *such that for all* $\epsilon, \delta \in (0,1)$*, for all* $m \geq m^{\mathrm{uc}}_{\mathcal{H}}(\epsilon, \delta)$ *and any distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$ *we have*

$$\mathrm{Prob}_{S\sim\mathcal{D}^m}\left[(\forall h \in \mathcal{H})\left(|L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon\right)\right] \geq 1 - \delta. \qquad (2)$$

- What you see is what you get!

Uniform convergence, ERM, and learnability

- Uniform convergence motivates the learning method of **empirical risk minimization** (ERM).
  - ERM$_{\mathcal{H}}$ on sample S selects a hypothesis with smallest empirical error,

$$h \in \min_{\mathcal{H}}(L_S(h)).$$

  - Uniform convergence gives a bound on true risk in terms of empirical error,

$$(\forall h \in \mathcal{H})\,(L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon)\,,$$

    and ERM can be seen to explicitly minimize this bound.
- If $\mathcal{H}$ has the uniform convergence property, then $\mathcal{H}$ is learnable by ERM$_{\mathcal{H}}$.
- But when does a hypothesis class have the uniform convergence property?

VC dimension

- It turns out that the uniform convergence property can be characterized by a purely combinatorial property of the hypothesis class.
- This is the property of finite **Vapnik-Chervonenkis dimension**.
- The VC dimension of $\mathcal{H}$ is the largest size k of a subset X of instances such that each of the 2^k possible labelings of X is predicted by some $h \in \mathcal{H}$. (It is infinite if there is no largest size.)
- It is a measure of the "richness" of a hypothesis class—the extent to which it covers possibilities.
- We call a hypothesis class with finite VC dimension a **VC class**.

The fundamental theorem

Theorem (Fundamental theorem of statistical learning theory)

The following are equivalent:

- $\mathcal{H}$ *has the uniform convergence property;*
- $\mathcal{H}$ *is uniformly learnable;*
- $\mathcal{H}$ *is uniformly learnable by* $ERM_{\mathcal{H}}$*;*
- $\mathcal{H}$ *is a VC class.*

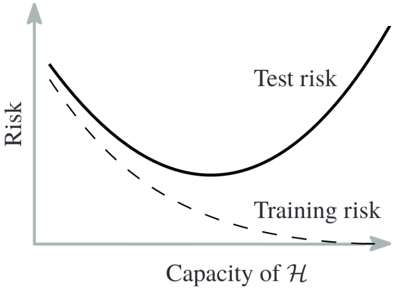
Underfitting v. overfitting

- The fundamental theorem concerns estimation error, which is one side of the infamous **bias-complexity trade-off**.

$$L_{\mathcal{D}}(\hat{h}) = \underbrace{\min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\text{approx. error}} + \underbrace{L_{\mathcal{D}}(\hat{h}) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)}_{\text{est. error}}.$$

  - The fundamental theorem is about model-relatively preventing *overfitting*—this is avoided if $\mathcal{H}$ is VC class.
  - The *absolute error* of an hypothesis also depends on how good our inductive model $\mathcal{H}$ is—whether $\mathcal{H}$ is not *underfitting*.

Underfitting v. overfitting



The plan

1. Model-relative justification.
2. Statistical learning theory.
3. **Application: Occam's razor.**
4. The "generalization puzzle": the way forward?

The justification for Occam's razor

- Is it a good idea to prefer simplicity in inductive inference?
- Old debate in machine learning and in particular in statistical learning theory (and older still in philosophy of science).

- Can statistical learning theory offer a justification for preferring simplicity?

The notion of simplicity

- No attempt at giving some robust definition of the complexity of *individual hypotheses* can be considered successful.
- However, learning theory has developed robust notions of the complexity or **capacity** of *hypothesis classes*.
- In particular, in statistical learning theory, the VC dimension is a robust notion of the complexity of a hypothesis class.
- We can call a hypothesis class *simple* iff it is a VC class.
- (I'm simplifying here. Finite size still matters, so really a graded notion.)

“Means-ends epistemology” (Formal learning theory)

- Kelly, Schulte, et al.: inductive problems call for a context-dependent means-ends analysis of what epistemic notions of success (ends) are attainable with what assumptions and methods (means).
  - Fits a forward-looking epistemological perspective, where the analysis gives a model-relative justification for methods that solve the problem.
- But we can also fix a learning problem and notion of success, and ask what assumptions are required for a method to possibly solve the problem.
  - Here we are after characterization results that give necessary and sufficient conditions for the attainability of the relevant notion of success.
  - Kelly (1996): “To revive Kant’s expression, such results may be thought of as *transcendental deductions* for reliable inductive inference, since they show what sort of knowledge is necessary if reliable inductive inference is to be possible.”
- For instance, the fundamental theorem shows that a necessary and sufficient condition for learnability is a hypothesis class of finite VC dimension.

The justification for Occam's razor

- A qualified model-relative means-ends justification.
  - We come to a certain problem of classification, that we are prepared to cast as a problem in statistical learning.
  - We come to this problem with further prior knowledge still, and we are interested in doing well relative to this prior knowledge.
  - Now the fundamental theorem tells us that for the formal guarantee of learnability (a formal expression of "doing well"), the hypothesis class (a formal encoding of our prior knowledge) must be a VC class—must be sufficiently *simple*.
- This gives us a means-ends model-relative justification for modeling, *if we can*, our prior knowledge in the shape of a simple class of hypotheses.
- The *if we can* is a pretty strong qualification—but can be weakened in a analogous argument based on the more general method of **structural risk minimization**.

The plan

1. Model-relative justification.
2. Statistical learning theory.
3. Application: Occam's razor.
4. **The "generalization puzzle": the way forward?**

A backwards-looking story . . . ?

- Does all of this still fit modern machine learning?
- Classical learning theory can't actually seem to account for the observed generalization behavior of certain much-used algorithms.
- Specifically, the model-relative guarantees from SLT are of the form,

  if the model is not too complex (has sufficiently low capacity),
  then we have a certain generalization guarantee.

- But the inductive models in these algorithms are way too complex!

Zooming in

- Indeed, a clear separation between inductive model and learning algorithm seems lost in learning algorithms for deep neural nets.
    - The relevant ERM-approximating algorithms (versions of stochastic gradient descent) appear to themselves have some implicit inductive bias.
- But maybe these algorithms implement some “implicit regularization” that still gives a low “effective capacity,” so that the SLT story still applies?
    - That is, maybe the antecent in “if sufficiently small capacity, then generalization” is still satisfied?

Zhang et al. (2017).

- More specifically, what SLT gives us are "what-you-see-is-what-you-get" ("wysiwig") bounds:

  if sufficiently small capacity, then wysiwig.

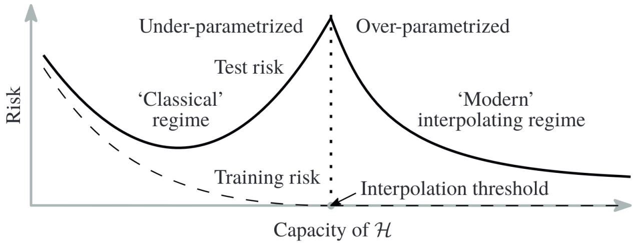
  - The paper by Zhang et al. (2017), "Understanding deep learning requires rethinking generalization," drives a wedge in the consequent.

- Specifically, they empirically showed the following:

1. A neural network applied to a "natural" dataset of labeled images attains low training error and low test error (**small generalization gap**).
2. The same network applied to **the same dataset but with the labels randomly shuffled** still attains low training error (indeed, perfect fit!) but (by definition—the data is random!) does not attain low test error (**large generalization gap**).

  - So training error *is no indication for* generalization error—directly against wysiwyg.

The double-descent phenomenon (Belkin et al., 2019)



The way forward?

- Belkin (2021): need for a "new framework for a theory of induction."
  - With again a fundamental role for a philosophical principle of Occam's razor: "Select the smoothest function, according to some notion of functional smoothness, among those that fit the data perfectly" (p. 218).
  - Rather akin to a "principle of simplicity of nature"...
- A different theory—and a different epistemology?

To conclude: take-home

Classical learning theory offers a general yet model-relative justification for standard learning algorithms.

However, classical learning theory has trouble accounting for modern generalization behavior—and it's not fully clear what a new learning theory and its justificatory story might look like.

tom.sterkenburg@lmu.de