Equivalences in learning: data vs distributions

George Barmpalias – September 2021

with Nan Fang and Frank Stephan

Automata Theory & Applications: Games, Learning, Structures

Institute of Software; Chinese Academy of Sciences

One of the central problems in statistics is:

> given a set of random data, find a distribution with respect to which the given data are random.

Algorithmic aspects studied in learning theories.

- Computational Learning Theory (CLT):

Probably approximately correct (PAC) learning model [Kearns et al., 1994].

- Algorithmic Learning Theory (ALT):

Gold-style learning (identification in the limit) [Vitanyi and Chater 2017, Bienvenu et al., 2014, 2017] .

Our motivation

This work connects:

- classic algorithmic learning theory (learning languages from texts)
- algorithmic learning of probability distributions (from random data)

We provide two equivalence theorems and:

- Reduce results of probabilistic ALT to the classical theory
- Prove many analogues of the classic theory in probabilistic ALT
- Develop probabilistic ALT with the help of the existing ALT

Precursor: Barmpalias/Stephan. Arxiv:1710.11303. Unpublished

Algorithmic Learning Theory (E.M. Gold 1967)

Motivation: a formal model of child language learning.

The main problem in ALT is identification in the limit:

> Given increasingly long initial segments of a formal language L, eventually find a grammar that generates L.

For computable languages (Type 0 in the Chomsky hierarchy):

> Given increasingly long initial segments of a computable $X \in 2^{\omega}$, eventually find a machine that computes X.

Guesses are provided by a learner.

A learner is a function $\mathcal{L} : 2^{<\omega} \to \mathbb{N}$.

Learning concepts

Algorithmic learning comes in many strengths:

- Guesses converge to a single description of the language
- Guesses converge extensionally to the language

  (but not to a fixed description of it)
- Limiting guess is correct up to finitely many mistakes
- A correct description is guessed infinitely often

  (with each of the other guesses occurring finitely often)
- and many more...

  (a zoo of learning concepts)

Learner success notions

We say that learner $\mathcal{L}$:

EX-succeeds on X if $\lim_n \mathcal{L}(X \upharpoonright_n)$ is an index of X

BC-succeeds on X if for almost all n, $\mathcal{L}(X \upharpoonright_n)$ is an index of X

BC*-succeeds on X if for almost all n, $\mathcal{L}(X \upharpoonright_n)$ is approximately an index of X (i.e. modulo finitely many mistakes)

Partially succeeds on X if $\mathcal{L}(X \upharpoonright_n)$ is a fixed index of X for infinitely many n, and any other guess appears finitely often.

ALT studies learnability of classes of languages.

Classic facts in Algorithmic Learning Theory

- The computable reals are not EX or BC learnable. (Gold)
- Learnability is not closed under union. (Blum and Blum)
- The computable reals are BC^* and partially learnable.
  (Harrington and Osherson/Stob/Weinstein)
- An oracle can EX-learn all computable reals iff it is high.
  (Adleman and Blum)
- Computational characterization of the Low for EX oracles
  (Slaman, Solovay, Pleszkoch, Gasarch, Jain)

Learning probability distributions

Given:

- a probability distribution P
- a sufficiently large sample of P-random data

learn or estimate a probability distribution with respect to which the sample has been randomly sampled.

Vitanyi/Chater (2013) formalized this problem by interpreting:

- learning as identification in the limit (ALT)
- randomness in terms of algorithmic information theory

Algorithmic information theory (AIT)

Randomness against effective detection of patterns:

- computable statistical tests (Martin-Löf)
- effective compression (Kolmogorov complexity)
- effective betting strategies (martingales).

AIT (as opposed to classic information theory) allows to:

- refer to specific objects in the sample space as being random
- as opposed to random variables.

Learning probability distributions

Given increasingly long initial segments of $X \in 2^\omega$ which is μ-random for a computable measure μ, eventually find a description of μ' such that X is μ'-random.

A learner is a function $\mathcal{L} : 2^{<\omega} \to \mathbb{N}$. We say that $\mathcal{L}$:

EX-succeeds on X if $\lim_n \mathcal{L}(X \upharpoonright_n)$ is an index of some μ such that X is μ-random.

BC-succeeds on X if there exists a computable μ such that X is μ-random and for almost all n, $\mathcal{L}(X \upharpoonright_n)$ is an index of μ

UD-succeeds on X if for almost all n, $\mathcal{L}(X \upharpoonright_n)$ is an index of some measure μ such that X is μ-random.

Partially succeeds on X if there exists μ such that X is μ-random, $\mathcal{L}(X \upharpoonright_n)$ is an index of X for infinitely many n, and any other approximation appears finitely often.

Learning probability distributions

Given any uniformly computable family of measures $\mathcal{C}$, there exists a computable learner that EX-succeeds on every μ-random real for any $\mu \in \mathcal{C}$.

(Vitanyi/Chater)

There is no computable learner which EX or BC succeeds on all reals that are μ-random for some computable (continuous) μ.

(Bienvenu, Monin, Shen)

There exists a learner which UD-succeeds on every μ-random real for any computable μ.

(Bienvenue, Figueira, Monin, Shen)

There exists a learner which partially succeeds on every μ-random real for any computable μ.

(Barmpalias/Stephan)

Learning classes of measures

Let $\mathcal{C}$ be a class of computable measures.

$\mathcal{C}$ is weakly EX-learnable if there exists a computable learner $\mathcal{L}$ such that for every $\mu \in \mathcal{C}$ and every μ-random X, $\lim_n \mathcal{L}(X \upharpoonright_n)$ exists and equals an index of some μ' such that X is μ'-random.

$\mathcal{C}$ is EX-learnable if there exists computable learner $\mathcal{L}$ such that for every $\mu \in \mathcal{C}$ and μ-random X, $\lim_n \mathcal{L}(X \upharpoonright_n)$ exists and equals an index of some $\mu' \in \mathcal{C}$ such that X is μ'-random.

Similarly for weak BC-learnability.

EX and BC learnability is not closed under subsets.

(weak versions clearly are)

First equivalence theorem (weak learning)

There exists a map from reals to (continuous) measures such that any class of reals is EX/BC learnable iff the corresponding class of measures is weakly EX/BC learnable.

Formally:

There exists a map $Z \to \mu_Z$ from 2^ω to the continuous Borel measures on 2^ω, such that for every class C of computable reals:

C is EX/BC learnable if and only if $\{\mu_Z \mid Z \in C\}$ is a weakly EX/BC learnable class of computable measures, respectively.

Applications of the first equivalence theorem

All previous results by Bienvenu et.al.

An oracle EX-succeeds on all μ-random reals for any computable (continuous) μ iff it is high.

Oracle Z EX-succeeds with finitely many queries on all μ-random reals (for any computable μ) iff $\emptyset'' \leq_T Z \oplus \emptyset'$.

Weak EX/BC measure learnability not closed under union.

Second equivalence theorem (learning)

If a class of measures $\mathcal{B}$ is nicely parametrized by a class of reals $\mathcal{C}$ then $\mathcal{B}$ is EX/BC learnable iff $\mathcal{C}$ is EX/BC learnable.

Formally, let $\mathcal{M}$ be the space of Borel measures on 2^ω:

Let $f : 2^\omega \to \mathcal{M}$ be computable and $\mathcal{D} \subseteq 2^\omega$ effectively closed such that for any $X \neq Y$ in $\mathcal{D}$ the measures $f(X), f(Y)$ are effectively orthogonal.

If $\mathcal{D}^* \subseteq \mathcal{D}$ is a class of computable reals, $\mathcal{D}^*$ is EX-learnable if and only if $f(\mathcal{D}^*)$ is EX-learnable.

The same is true of the BC learnability of $\mathcal{D}^*$.

Proof – from reals to distributions (easier)

- orthogonal class of measures parametrized by reals
- along a random stream, there exists a unique correct guess
- wrong guesses eliminated by estimating the deficiency
- pull-back finite approximations of the current guess to a real that codes them
- continuity ensures that the pull-back real converges to the real coding the correct measure guess
- classic-learner acts on the pull-back, guessing an index
- compute index of the measure, given index of its code

Proof – from distributions to reals (harder)

- While reading a real, we consider approximations of the corresponding measure
- from the measure get approximations of class of random streams
- apply the probability learner on the random data
- take a majority vote amongst the various measure guesses
- pull the majority guess index back to an index of its real code
- Problem: different data streams may point to different indices of the same measure (taking a majority vote becomes tricky)
- EX-case: some (correct) index will be guessed with positive probability
- BC-case: majority is taken extensionally

Applications of the second equivalence theorem

There exist two EX-learnable classes of (Bernoulli) measures whose union is not EX-learnable.

The class of computable Bernoulli measures is EX-learnable with oracle A iff A is high.

Every low for EX (for measures) is low for EX for reals.

Oracles $\not\leq_T 0'$ are not low for EX for measures.

Open problems

Question 1. Suppose that an oracle can BC-learn the class of computable functions.

Can it necessarily learn the class of computable measures?

Continuous computable measures?

Remark: in the classic theory there is no succinct characterization of the oracles that BC-learn the computable functions.

Question 2. We showed that every such oracle is also low for EX-learning in the classical learning theory of reals.

Does the converse hold?

Thanks! – and main references

Barmpalias/Fang/Stephan.
Equivalences between learning of data and distributions.
Information and computation (2018) Arxiv:1801.02566.

Barmpalias/Stephan.
Algorithmic learning of distributions from data in the limit.
Unpublished. Arxiv:1710.11303.

Bienvenu/Figueira/Monin/Shen.
Algorithmic identification of probabilities is hard.
J. Computer and System Sciences 2018 and Alg. Learning Th. 2014

Bienvenu/Monin.
Von Neumann's Biased Coin Revisited. LICS (2012)

Vitanyi/Chater.
Identification of probabilities. J. Mathematical Psycology (2017)