Machine learning and the Continuum Hypothesis

Non impeditus ab ulla sciencia

K. P. Hart

Faculty EEMCS
TU Delft

BLAST2022, 10 August, 2022 (18:30-19:30 UTC)

Headline in Nature

MENU nature

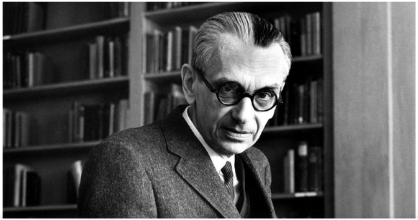
Subscribe

NEWS · 08 JANUARY 2019

Machine learning leads mathematicians to unsolvable problem

Simple artificial-intelligence problem puts researchers up against a logical paradox discovered by famed mathematician Kurt Gödel.

Davide Castelvecchi



PDF version

RELATED ARTICLES

Paradox at the heart of mathematics makes physics problem unanswerable

Enigmatic foundations of maths put to music

That sounds exciting . . .

But, when you read this piece in *Nature* about a paper in *Nature Machine Learning* . . .

“a paradox known as the Continuum Hypothesis”

and no clear (for me) description of what actually happened

So, . . . , what gives?

What's in the paper itself?

In one sentence:

the existence of a certain kind of learning function is equivalent to

$$2^{\aleph_0} < \aleph_\omega$$

Almost what Cantor wrote in 1878:
"Durch ein Induktionsverfahren, auf dessen Darstellung wir hier nicht näher eingehen, wird der Satz nahe gebracht, daß die Anzahl der nach diesem Einteilungsprinzip sich ergebenden Klassen linearer Mannigfaltigkeiten eine endliche und zwar, daß sie gleich Zwei ist."

What's in the paper?

In more detail.

The problem: find a method to pick a finite set that maximizes, within a certain tolerance, a certain expected value.

The difficulty: the probability distributions are unknown.

Approach: work with the family of finite subsets of the unit interval $\mathbb{I}$.

An abstract learning function

Wanted: a function

$$G : \bigcup_{k \in \mathbb{N}} \mathbb{I}^k \to \mathrm{fin}(\mathbb{I})$$

with certain properties.

Look at $\mathbb{P}$, the family of all probability distributions on $\mathbb{I}$ with finite support.
Every finite subset F has an expectation with respect to such a distribution.
We let $\mathrm{Opt}(P) = \sup\{\mathbb{E}_P(F) : F \in \mathrm{fin}(\mathbb{I})\}$.
The objective is to learn/guess(?) as well as possible.

An abstract learning function

G is an (ε, δ)-EMX learning function if there is an integer d (depending on ε and δ) such that

$$\Pr_{S \sim P^d}\left[\mathbb{E}_P\big(G(S)\big) \leqslant \mathrm{Opt}(P) - \varepsilon\right] \leqslant \delta$$

An abstract learning function

Translation to (our kind of) combinatorics:

there is such a function with $\varepsilon = \delta = \frac{1}{3}$

if and only if

there is an $(m+1) \to m$ monotone compression scheme, for some $m \in \mathbb{N}$
In fact $m = \left\lceil \frac{3}{2} d \right\rceil$, where d corresponds to $(\frac{1}{3}, \frac{1}{3})$.

A what?

Monotone compression schemes

What is a $k \to l$ monotone compression scheme?

A function $\eta : [\mathbb{I}]^l \to \mathrm{fin}(\mathbb{I})$ such that

for every $x \in [\mathbb{I}]^k$ there is a $y \in [x]^l$ such that $x \subseteq \eta(y)$.

We reformulate this.

In the above there is an implicit function $\sigma : [\mathbb{I}]^k \to [\mathbb{I}]^l$ with the property that

$$\sigma(x) \subseteq x \subseteq \eta\big(\sigma(x)\big)$$

Monotone compression schemes

We only need σ!

There is an $k \to l$ monotone compression scheme
if and only if
there is a finite-to-one function $\sigma : [\mathbb{I}]^k \to [\mathbb{I}]^l$ such that $\sigma(x) \subseteq x$ for all x

'only if': use η; if $y \in [\mathbb{I}]^l$ then $\sigma(x) = y$ implies $x \subseteq \eta(y)$

'if': define η by $\eta(y) = \bigcup\{x : \sigma(x) = y\}$

Where are the cardinals?

Here:

Theorem

Let X be a set and $k \in \mathbb{N}$;
there is a finite-to-one function $\sigma : [X]^{k+2} \to [X]^{k+1}$ such that $\sigma(x) \subseteq x$ for all x
if and only if

$$|X| \leqslant \aleph_k$$

And there you have it:
there is an $(m+1) \to m$ monotone compression scheme *for some m*
if and only if

$$|\mathbb{I}| < \aleph_\omega$$

An old result of Kuratowski's

Theorem (Kuratowski 1951)

Let X be a set and $k \in \mathbb{N}$; then $|X| \leqslant \aleph_k$ if and only if

$$X^{k+2} = \bigcup_{i<k+2} A_i,$$

where for every $i < k+2$ and every point $\langle x_j : j < k+2 \rangle$ in X^{k+2} the set of points y in A_i that satisfy $y_j = x_j$ for $j \neq i$ is finite;

in Kuratowski's words: "A_i is finite in the direction of the ith axis".

An old result of Kuratowski's: $k = 0$

Look at $\mathbb{N}^2$.
$A_0 = \{\langle m, n\rangle : m \leqslant n\}$ and $A_1 = \{\langle m, n\rangle : m > n\}$.

An old result of Kuratowski's: $k = 1$

This is already non-trivial:

to make A_0, A_1, and A_2 in ω_1^3 choose, simultaneously, for every $\alpha \in \omega_0$ a partition $B_0(\alpha) \cup B_1(\alpha)$ of $(\alpha + 1)^2$ such that $B_i(\alpha)$ is finite on the ith coordinate.

Put $\langle \alpha, \beta, \gamma \rangle$ in A_0 if

- $\max\{\alpha, \beta, \gamma\} = \beta$ and $\langle \alpha, \gamma \rangle \in B_0(\beta)$ or else
- $\max\{\alpha, \beta, \gamma\} = \gamma$ and $\langle \alpha, \beta \rangle \in B_0(\gamma)$

An old result of Kuratowski's: $k = 1$

Put $\langle\alpha, \beta, \gamma\rangle$ in A_1 if

- $\max\{\alpha, \beta, \gamma\} = \alpha$ and $\langle\beta, \gamma\rangle \in B_0(\alpha)$ or else
- $\max\{\alpha, \beta, \gamma\} = \gamma$ and $\langle\alpha, \beta\rangle \in B_1(\gamma)$

Put $\langle\alpha, \beta, \gamma\rangle$ in A_2 if

- $\max\{\alpha, \beta, \gamma\} = \alpha$ and $\langle\beta, \gamma\rangle \in B_1(\alpha)$ or else
- $\max\{\alpha, \beta, \gamma\} = \beta$ and $\langle\alpha, \gamma\rangle \in B_0(\beta)$

An old result of Kuratowski's: $k \geqslant 2$

If you understand the case $k = 1$ you understand these cases too.

There is a connection

We, generally, identify $[X]^n$ with

$$\{x \in X^n : (i < j < n) \to (x_i < x_j)\}$$

(assuming X has a linear order of course).

It is now quite easy to create our function $\sigma : [\omega_k]^{k+2} \to [\omega_k]^{k+1}$ from Kuratowski's decomposition.

There is a connection

Without loss of generality the A_i are pairwise disjoint.

Let $x \in [\omega_k]^{k+2}$, so $x = \langle x_i : i < k+2 \rangle$ with $(i < j < k+2) \to (x_i < x_j)$.

Take the i with $x \in A_i$ and let $\sigma(x) = x \setminus \{x_i\}$.

If $y \in [\omega_k]^{k+1}$ then for each $i < k+2$ there are only finitely many x in A_i with $y = \sigma(x)$.

In fact there are zero x in A_{k+1} with $y = \sigma(x)$

In case $k = 0$ we have $\sigma(x) = x \setminus \{\min x\} = \{\max x\}$

Because: Kuratowski's recipes yield $[\omega_k]^{k+2} \subseteq \bigcup_{i<k+1} A_i$.

There is a connection

Suppose $n > m$ and $\sigma : [\omega_{k+1}]^n \to [\omega_{k+1}]^m$ is finite-to-one and such that $\sigma(x) \subseteq x$ for all x.

The set C of $\delta \in \omega_{k+1}$ that are closed under $\sigma^{\leftarrow}$ is closed and unbounded.

I mean: if $\delta \in C$ and $y \in [\delta]^m$ then $x \in [\delta]^n$ whenever $y = \sigma(x)$.

Take $\delta \in C$ with $\delta \geqslant \omega_k$.
Then $\varsigma : [\delta]^{n-1} \to [\delta]^{m-1}$, defined by

$$\varsigma(x) = \sigma(x \cup \{\delta\}) \setminus \{\delta\}$$

is finite-to-one and satisfies $\varsigma(x) \subseteq x$ for all x.

Summary

We get the following

Theorem

Let X be a set and $k \in \mathbb{N}$. Then the following are equivalent.

1. $|X| \leqslant \aleph_k$
2. $X^{k+2} = \bigcup_{i<k+2} A_i$*, where for every* $i < k+2$ *the set* A_i *is finite in the direction of the* i*th axis*
3. *there is a* $(k+2) \to (k+1)$ *monotone compression scheme for* X*.*

For "3 implies 1" use the previous slide:

if ω_{k+1} has a $(k+2) \to (k+1)$ monotone compression scheme
then ω_k has a $(k+1) \to k$ monotone compression scheme
and ...
and ω_0 has a $1 \to 0$ monotone compression scheme

Extra equivalence

Cichoń and Morayne used Kuratowski's decompositions and generalizations thereof to prove that

$$2^{\aleph_0} \leqslant \aleph_k$$

is equivalent to
the existence of surjections $f : \mathbb{R}^k \to \mathbb{R}^{k+m}$ such that at every $x \in \mathbb{R}^k$ at least k of the coordinate functions are differentiable at x.

In particular (Morayne):
CH is equivalent to the existence of a surjective $f : \mathbb{R} \to \mathbb{R}^2$ such that at every point one of the two coordinates is differentiable.

Are there algorithms?

The functions in the proofs given above and in the paper are quite non-constructive as they involve blatant appeals to the Axiom of Choice.

How about algorithmic/definable/... functions?

Say, continuous, or Borel measurable.

High-brow answer

No.

If $\sigma : [\mathbb{I}]^{m+1} \to [\mathbb{I}]^{m}$ is a Borel measurable function that determines a compression scheme then

after adding $\aleph_{\omega+1}$ Random reals its reinterpretation should still work, which it doesn't.

Elementary answer

Assume $\sigma : [\mathbb{I}]^{m+1} \to [\mathbb{I}]^m$ is a monotone compression scheme.

If σ is continuous then there is a single i such that $\sigma(x) = x \setminus \{x_i\}$ for all x in $[\mathbb{I}]^{m+1}$.

Main Lemma: $O_i = \big\{x : \sigma(x) = x \setminus \{x_i\}\big\}$ is open.

Elementary answer

If σ is Borel measurable the above is almost true:
there are an $x \in [\mathbb{I}]^m$ and a non-meager set A such that $x = \sigma(x \cup \{a\})$ for all $a \in A$.

In either case σ is far from finite-to-one

Consequence

If the learning function G from the beginning is Borel measurable then so is the compression scheme.

So to me this shows that that problem does not look so undecidable after all: there is no algorithm that works.

On the other hand . . .

Instead of $\mathbb{I}$, why not use $\mathbb{Q}$ (or $\mathbb{Q} \cap \mathbb{I}$)?

To begin: the formula $\sigma : x \mapsto \{\max x\}$ not only defines a $2 \to 1$ monotone compression scheme on $\mathbb{N}$, it gives a $k \to 1$ scheme for every k.

The corresponding function η is just $\{n\} \mapsto n + 1$.

The learning function is simply one of these, for a suitably large k.

Maybe some clever enumeration of $\mathbb{Q}$ will lead to a useful learning function but, to reflect the subtitle of the slides: I have no idea whether this is a sensible suggestion.

Light reading

Website: fa.ewi.tudelft.nl/~hart

Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff, *Learnability can be undecidable*, Nature Machine Intelligence **1** (2019), 44–48.

Klaas Pieter Hart,
Machine learning and the Continuum Hypothesis,
Nieuw Archief voor Wiskunde (5), **20** (2019), no. 3, 214–217