Machine learning and the Continuum Hypothesis

Tá scéilín agam

K. P. Hart

Faculty EEMCS
TU Delft

Winchester, 28 June, 2019: 11:30–12:30

Headline in Nature

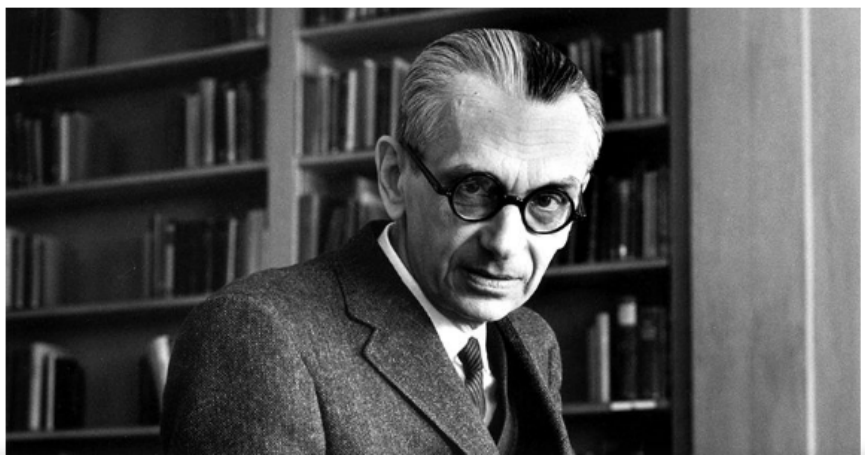
MENU nature Subscribe

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

Sounds exciting

Some quotes from the description in Nature

A team of researchers has stumbled on a question that is mathematically unanswerable because it is linked to logical paradoxes discovered by Austrian mathematician Kurt Gödel in the 1930s that can't be solved using standard mathematics.

`https://www.nature.com/articles/d41586-019-00083-3`

Some quotes from the description in Nature

The mathematicians, who were working on a machine-learning problem, show that the question of 'learnability' — whether an algorithm can extract a pattern from limited data — is linked to a paradox known as the continuum hypothesis. Gödel showed that the statement cannot be proved either true or false using standard mathematical language.

`https://www.nature.com/articles/d41586-019-00083-3`

A quote from the paper itself

Our proof utilizes one of the most revolutionary mathematical discoveries in the past century: Gödel's incompleteness theorems. Roughly speaking, they state that there are mathematical questions that cannot be resolved. Theorems stating the impossibility of resolving mathematical questions are called independence results.

`https://www.nature.com/articles/s42256-018-0002-3`

Balderdash . . .

- Gödel did not "discover paradoxes"
- The Continuum Hypothesis is not a paradox
- Gödel 'merely' showed that it cannot be proved false
  Cohen showed it cannot be proved true
- The proof does not use (hence certainly does not utilise)
  Gödel's Incompleteness theorems

So . . . , what gives?

What's in the paper itself?

In one sentence:

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

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

Sounds familiar?

Here is what Cantor wrote in *Ein Beitrag zur Mannigfaltigkeitslehre* (Crelles Journal für Mathematik **84** (1878) 242–258):

"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."

[There is a finite number (in fact there are two) equivalence classes of infinite subsets of $\mathbb{R}$ under 'there is a bijection between them'.]

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

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

for every (finitely supported) probability distribution P over $\mathbb{I}$.

EMX: Estimate the MaXimum

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 $m \in \mathbb{N}$

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 $y \in [x]^l$ with $x \subseteq \eta(y)$.

We reformulate this.

In the above there is an implicit (choice) 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 $x \subseteq \eta(y)$
(the preimage of y has at most $2^{|\eta(y)|}$ points)

'if': define η by $\eta(y) = \bigcup\{x : \sigma(x) = y\}$
(a union of finitely many finite sets)

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 $m \in \mathbb{N}$
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

Example 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

Example 1: look at ω_1^3.

to make A_0, A_1, and A_2 in ω_1^3 choose, simultaneously, for every $\alpha \geqslant \omega_0$ a well-order $\prec_\alpha$ of $\alpha + 1$ in type ω.

Exercise: for every $\alpha \geqslant \omega_0$ use $\prec_\alpha$ to write

$$(\alpha + 1)^2 = A_0(\alpha) \cup A_1(\alpha)$$

and manufacture A_0, A_1, and A_2 out of these sets.

An old result of Kuratowski's

Partial solution:

One puts $\langle \alpha, \beta, \gamma \rangle$ into A_0

- if β is the largest coordinate and $\langle \alpha, \gamma \rangle \in A(\beta, 0)$ or
- if γ is the largest coordinate and $\langle \alpha, \beta \rangle \in A(\gamma, 0)$.

So, for fixed $\langle \beta, \gamma \rangle$ we have $\langle \alpha, \beta, \gamma \rangle \in A_0$ iff

- $\gamma \leqslant \beta$ and $\alpha \preccurlyeq_\beta \gamma$ or
- $\beta < \gamma$ and $\alpha \preccurlyeq_\gamma \beta$

so that is finitely many αs

There is a connection

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

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

(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 $y = \sigma(x)$.

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 is 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.

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 $\aleph_{\omega+1}$ Cohen reals its reinterpretation should still work,

which it does not.

Elementary answer

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

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

Exercise: show that 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 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.

But . . .

do we really need the unit interval?

Why not use the rationals?

Then we have an easy $2 \to 1$ monotone compression scheme: order $\mathbb{Q}$ in type ω, by $\preccurlyeq$ say, and put

$$\sigma\big(\{q\}\big) = \{p : p \preccurlyeq q\}$$

What kind of EMX-learning function this produces I don't know.

Light reading

Blog: `hartkp.weblog.tudelft.nl`

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,
`https://arxiv.org/abs/1901.04773`
(to appear in *Nieuw Archief voor Wiskunde*).

And a big

‘Thank You’

to the organisers!