MACHINE LEARNING AND THE CONTINUUM HYPOTHESIS

KLAAS PIETER HART

Abstract. We comment on a recent paper that connects certain forms of machine learning to Set Theory.

We point out that part of the set-theoretic machinery is related to a result of Kuratowski about decompositions of finite powers of sets and we show that there is no Borel measurable monotone compression function on the unit interval.

Introduction

In the paper [1] its authors exhibit an abstract machine-learning situation where the learnability is actually neither provable nor refutable on the basis of the axioms of ZFC.

The learnability condition is translated into a combinatorial statement about the family of finite subsets of the unit interval, which is then shown to hold if and only if $2^{\aleph_0} < \aleph_\omega$. This then shows that the existence of a 'learner' is both consistent with and independent from the axioms of ZFC.

This note has two purposes.

The first is to point out that the authors' method is related to a result of Kuratowski from [6] about decompositions of finite powers of sets and that part of their main result follows from that result.

The second is to show that there is no Borel measurable learning function.

The latter is an attempt to address a point already raised by the authors in [1]: the functions they use are quite arbitrary and not related to any recognizable finitary algorithm. One possible way of separating out 'algorithmic' functions is by requiring them to have nice descriptive properties. If 'nice' is taken to mean 'Borel measurable' then the desired functions do not exist.

1. Preliminaries

We begin by describing the combinatorial statement that is equivalent to the existence of a learning function.

We denote the unit interval $[0, 1]$ by $\mathbb{I}$ and we let $\mathcal{F}$ denote the family of finite subsets of $\mathbb{I}$.

Definition 1.1. Let m and d be two natural numbers with $m > d$. An $m \to d$ *monotone compression scheme* is a function $\eta : \mathbb{I}^d \to \mathcal{F}$ such that whenever A is an m-element subset of $\mathbb{I}$ it has a d-element subset B such that $A \subseteq \eta(B)$, where we identify B with the point in $\mathbb{I}^d$ that is the monotone enumeration of B.

Date: Wednesday 30-01-2019 at 22:57:04 (cet).

2010 *Mathematics Subject Classification.* Primary: 03E05; Secondary: 03E15, 54B20, 54C05, 54H05.

Key words and phrases. machine learning, Kuratowski decomposition, finite-to-one map, Borel measurable.

Though the formulation of Definition 2 in [1] leaves open the possibility that $|A| < m$ and that $|B| < d$, as it uses indexed sets, it is clear from the results and their proofs that our definition captures the essence of the notion.

There is a second function implicit in Definition 1.1: the choice of the subset B of A. In [1] it does not get a name, we call it σ. So our schemes consist of a pair of functions: $\sigma : [\mathbb{I}]^m \to [\mathbb{I}]^d$ and $\eta : \mathbb{I}^d \to \mathcal{F}$; they should satisfy $A \subseteq (\eta \circ \sigma)(A)$ for all A.

There is nothing special about the unit interval in this definition; we can and shall apply this definition to other sets as well.

One final simplification is the following: our sets will come with a linear order. Therefore we can, in a linearly ordered set $(X, \prec)$, identify an m-element subset of X with the point in X^m that is its monotone enumeration, so for us $[X]^m = \{x \in X^m : (i < j < m) \to (x_i \prec x_j)\}$.

In the case of $\mathbb{I}$ this makes it possible to speak of continuity or Borel measurability of the functions σ and η. For the latter we consider the standard Vietoris topology on the space of finite subsets, see [3, 2.7.20, 4.5.23].

The statement that is shown to be equivalent to $2^{\aleph_0} < \aleph_\omega$ is:

Weak compressability. For some $m \in \mathbb{N}$ there is an $(m+1) \to m$ monotone compressablity scheme for the finite subsets of $\mathbb{I}$.

The equivalence of this statement with $2^{\aleph_0} < \aleph_\omega$ follows from the following set of equivalences.

Theorem 1.2 ([1, Theorem 1])**.** *Let $k \in \mathbb{N}$ and let X be a set. Then there is a $(k+2) \to (k+1)$ monotone compression scheme for the finite subsets of X if and only $|X| \leqslant \aleph_k$.*

Indeed, $2^{\aleph_0} < \aleph_\omega$ if and only if $|\mathbb{I}| = \aleph_k$ for some $k \in \mathbb{N}$.

The amount of Set Theory needed to appreciate this result and follow the proofs in the next section is not too large. The first chapter of Kunen's book [5] more than suffices.

2. Compression and decompositions

In this section we give an equivalent description of monotone compression schemes that does not mention the function η. This shows that it is σ that is doing the compressing. We then use this description to connect the existence of compression schemes to a result of Kuratowski from [6] that also characterizes when a set has cardinality at most $\aleph_k$.

2.1. **Reformulating compression.** Though the function η may have an important role to play in concrete compression schemes, they are not really needed in existence theorems.

Proposition 2.1. *Let m and d be natural numbers and let X be a set. There is an $m \to d$ monotone compression scheme for the finite subsets of X if and only if there is a finite-to-one function $\sigma : [X]^m \to [X]^d$ such that $\sigma(x) \subseteq x$ for all x.*

Proof. If the pair $\langle \eta, \sigma \rangle$ determines an $m \to d$ monotone compression scheme then σ is finite-to-one. For let $y \in [X]^d$ then $\sigma(x) = y$ implies $x \subseteq \eta(y)$, hence there are at most $\binom{M}{m}$ such x, where $M = \bigl|\eta(y)\bigr|$.

Conversely, if σ is as in the statement of the proposition then we can let $\eta(y) = \bigcup\{x : \sigma(x) = y\}$. $\square$

2.2. **Kuratowski's decompositions.** The following theorem, proved by Kuratowski in [6] provides one direction in his characterization of when a set has cardinality at most $\aleph_k$.

Theorem 2.2. *The power ω_k^{k+2} can be written as the union of $k+2$ sets, $\{A_i : i < k+2\}$, such that for every $i < k+2$ and every point $\langle x_j : j < k+2\rangle$ in ω_k^{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".*

The case $k = 0$ is easy: $A_0 = \{\langle m, n\rangle : m \leqslant n\}$ and $A_1 = \{\langle m, n\rangle : m > n\}$.

The case $k = 1$ and every later induction step involves a blatant application of the Axiom of Choice: for every ordinal α in an interval $[\omega_n, \omega_{n+1})$ one needs a well-order $<_\alpha$ in order type ω_n. This same choice is used in the original proof of Theorem 1.2, which indicates that the resulting compression schemes have no simple description.

We now show how Theorem 2.2 can be used to prove sufficiency in Theorem 1.2.

Constructing a compression scheme from a decomposition. From a decomposition as in Theorem 2.2 we construct a finite-to-one function $\sigma : [\omega_k]^{k+2} \to [\omega_k]^{k+1}$ such that $\sigma(x) \subseteq x$ for all x. We assume, without loss of generality, that the sets A_i are disjoint.

Let $x \in [\omega_k]^{k+2}$ (so $i < j < k+2$ implies $x_i < x_j$). Take (the unique) i such that $x \in A_i$ and let $\sigma(x)$ be the point in ω_k^{k+1} that is x but without its coordinate x_i. In terms of sets we would have set $\sigma(x) = x \setminus \{x_i\}$.

This function is finite-to-one: if $y \in [\omega_k]^{k+1}$ then for each $i < k+2$ there are only finitely many x in A_i with $y = \sigma(x)$. □

As mentioned above Kuratowski's result works both ways: if X^{k+2} admits a decomposition as above for ω_k^{k+2} then $|X| \leqslant \aleph_k$. This suggests that the necessity in Theorem 1.2 is related to the converse of Theorem 2.2. This is indeed the case: one can construct a Kuratowski-type decomposition from a compression scheme, but because of our definition of the schemes we only get a decomposition of the subset $[\omega_k]^{k+2}$ of the whole power. This can be turned into one for the whole power but the process is a bit messy so we leave it be.

The proof of necessity from [1] closes the circle of implications that proves the following.

Theorem 2.3. *For a set X and a natural number k the following are equivalent:*

(1) $|X| \leqslant \aleph_k$,
(2) X^{k+2} *admits a Kuratowski-type decomposition into $k+2$ sets,*
(3) *there is a $(k+2) \to (k+1)$ monotone compression scheme for the finite subsets of X.*

3. Continuity and Borel measurability

In this section we show that there does not exist an $(m+1) \to m$ monotone compression scheme for the finite subsets of $\mathbb{I}$ where the function σ is Borel measurable. Sp, let m be a natural number and let $\sigma : [\mathbb{I}]^{m+1} \to [\mathbb{I}]^m$ be a function such that $\sigma(x) \subseteq x$ for all x.

If σ is continuous then σ is not finite-to-one. One can apply [4, Theorem VI.7] and deduce that there is a point y such that the fiber $\sigma^{\leftarrow}(y)$ is one-dimensional, but in this case there is an elementary and more informative argument.

To this end let $x \in [\mathbb{I}]^{m+1}$ and assume for notational convenience that $\sigma(x) = \langle x_i : i < m\rangle$, i.e., that the coordinate x_m is left out of x when forming $\sigma(x)$.

Let $\varepsilon = \frac{1}{3}\min\{x_{i+1} - x_i : i < m\}$ and let $\delta > 0$ be such that $\delta \leqslant \varepsilon$ and for all $y \in [\mathbb{I}]^{m+1}$ with $\|y - x\| < \delta$ we have $\|\sigma(y) - \sigma(x)\| < \varepsilon$.

Now if $y \in [\mathbb{I}]^{m+1}$ and $\|y - x\| < \delta$ then $|y_i - x_i| < \varepsilon$ for all $i \leqslant m$. Also, when $i < j$ we have $x_j - x_i > 3\epsilon$. It follows that $y_m - x_i > \epsilon$ for all $i < m$. This implies that $\sigma(y) = \langle y_i : i < m\rangle$ for all y with $\|y - x\| < \delta$.

This shows that for every i the set $O_i = \bigl\{x \in [\mathbb{I}]^{m+1} : \sigma(x) = x \setminus \{x_i\}\bigr\}$ is open. Because $[\mathbb{I}]^{m+1}$ is connected there is one i such that $O_i = [\mathbb{I}]^{m+1}$. This shows that σ cannot be finite-to-one. $\square$

The above proof can be used/adapted to show that if σ is Borel measurable it is not finite-to-one either.

If σ is Borel measurable then σ is not finite-to-one. There is a dense G_δ-set G in $[\mathbb{I}]^{m+1}$ such that the restriction of σ to G is continuous, see [7, § 31 II].

Let $x \in G$. As in the previous proof we assume $\sigma(x) = \langle x_i : i < m\rangle$ and we obtain a $\delta > 0$ such that $\sigma(y) = \langle y_i : i < m\rangle$ for all $y \in G$ that satisfy $\|y - x\| < \delta$.

By the Kuratowski-Ulam theorem, [8], we can find a point y in G with $\|y - x\| < \delta$ such that the set of points t in the interval $(x_m - \delta, x_m + \delta)$ for which $y_t = \sigma(y) * \langle t\rangle$ belongs to G is co-meager. But for every such point we have $\sigma(y_t) = \sigma(y)$ and this shows that σ is not finite-to-one. $\square$

4. Remarks

In [1] a learning function is a function G from the union $\bigcup_{k\in\mathbb{N}} \mathbb{I}^k$ to the family of finite subsets of $\mathbb{I}$. We can call such a function continuous or Borel measurable if its restriction to each individual power is.

In the construction of an $(m+1) \to m$ compression scheme from a learning function the authors use its restriction to just one of these powers $\mathbb{I}^d$, where $d \leqslant m$. The definition of $\eta(S)$ involves taking the union of $G(T)$ for all d-element subsets T of S, hence a union of $\binom{m}{d}$ many sets.

The definition of σ involves choosing one m-element subset with a certain property from of a given $m+1$-element set.

The latter choice can be made explicit using a Borel linear order on the family of all finite subsets of $\mathbb{I}$, or even $[\mathbb{I}]^m$.

An analysis of this procedure shows that if G is Borel measurable then so are σ and η.

The results of this section then imply that a Borel measurable learning function does not exist. In this author's opinion that means that the title of [1] should be emended to “EMX-learning is impossible”.

References

The paper by Kuratowski is the second of three consecutive papers — [11], [6] and [12] — that deal with the same type of partition problem.

References [2] and [9] are commentaries in *Nature* on [1].

[1] Shai Ben-David, Pavel Hrubeš, Shay Moran, Amir Shpilka, and Amir Yehudayoff, *Learnability can be undecidable*, Nature Machine Intelligence **1** (2019), 44–48, DOI 10.1038/s42256-018-0002-3.

[2] Davide Castelvecchi, *Machine learning leads mathematicians to unsolvable problem*, Nature **565** (2019), 277, DOI 10.1038/d41586-019-00083-3.

[3] Ryszard Engelking, *General topology*, 2nd ed., Sigma Series in Pure Mathematics, vol. 6, Heldermann Verlag, Berlin, 1989. Translated from the Polish by the author. MR1039321

[4] Witold Hurewicz and Henry Wallman, *Dimension Theory*, Princeton Mathematical Series, v. 4, Princeton University Press, Princeton, N. J., 1941. MR0006493

[5] Kenneth Kunen, *Set theory*, Studies in Logic and the Foundations of Mathematics, vol. 102, North-Holland Publishing Co., Amsterdam-New York, 1980. An introduction to independence proofs. MR597342

[6] Casimir Kuratowski, *Sur une caractérisation des alephs*, Fundamenta Mathematicae **38** (1951), 14–17, DOI 10.4064/fm-38-1-14-17 (French). MR0048518

[7] K. Kuratowski, *Topology. Vol. I*, New edition, revised and augmented. Translated from the French by J. Jaworowski, Academic Press, New York-London; Państwowe Wydawnictwo Naukowe, Warsaw, 1966. MR0217751

[8] C. Kuratowski and St. Ulam, *Quelques propriétés topologiques du produit combinatoire*, Fundamenta Mathematicae **19** (1932), 247–251. ZBL 0005.18301

[9] Lev Reyzin, *Unprovability comes to machine learning*, Nature **565** (2019), 166–167, DOI 10.1038/d41586-019-00012-4.

[10] Wacław Sierpiński, *Hypothèse du continu*, Monografie Matematyczne, vol. 4, PWN - Panstwowe Wydawnictwo Naukowe, Warszawa, 1934 (French). JFM 60.0035.01. ZBL 0009.30201

[11] ———, *Sur quelques propositions concernant la puissance du continu*, Fundamenta Mathematicae **38** (1951), 1–13, DOI 10.4064/fm-38-1-1-13 (French). MR0048517

[12] Roman Sikorski, *A characterization of alephs*, Fundamenta Mathematicae **38** (1951), 18–22, DOI 10.4064/fm-38-1-18-22. MR0048519

Faculty EEMCS, TU Delft, Postbus 5031, 2600 GA Delft, the Netherlands
E-mail address: k.p.hart@tudelft.nl
URL: http://fa.its.tudelft.nl/~hart