Unprovability comes to machine learning

Scenarios have been discovered in which it is impossible to prove whether or not a machine-learning algorithm could solve a particular problem. This finding might have implications for both established and future learning algorithms.

LEV REYZIN[*]

During the twentieth century, discoveries in mathematical logic revolutionized our understanding of the very foundations of mathematics. In 1931, the logician Kurt Gödel showed that, in any system of axioms that is expressive enough to model arithmetic, some true statements will be unprovable [1]. And in the following decades, it was demonstrated that the continuum hypothesis — which states that no set of distinct objects has a size larger than that of the integers but smaller than that of the real numbers — can be neither proved nor refuted using the standard axioms of mathematics [2–4]. Writing in Nature Machine Intelligence, Ben-David et al. [5] show that the field of machine learning, although seemingly distant from mathematical logic, shares this limitation. They identify a machine-learning problem whose fate depends on the continuum hypothesis, leaving its resolution forever beyond reach.

Machine learning is concerned with the design and analysis of algorithms that can learn and improve their performance as they are exposed to data. The power of this idea is illustrated by the following example: although it seems hopelessly difficult to explicitly program a computer to determine what objects are in a picture, the Viola–Jones machine-learning system can detect human faces in real time after being trained on a labelled sample of photographs [6]. Today, we regularly interact with machine-learning algorithms, from virtual assistants on our phones to spam filters for our e-mail. But these modern real-world applications trace their origins to a subfield of machine learning that is concerned with the careful formalization and mathematical analysis of various machine-learning settings.

The goal of learning a predictor (a mathematical function that can be used to make predictions) from a database of random examples was formalized in the aptly named probably approximately correct (PAC) learning model [7]. In this model, the aim is to train the predictor to match some true function that labels the data. A different model, called online learning, has the learner making immediate predictions as data arrive — for example, capturing a trading system's task of executing transactions in an ever-changing market. And another model known as multiarmed bandits can simulate clinical trials, in which the medical outcomes that an experimenter observes depend on his or her own choices.

These are only a few examples of the many models used in machine learning. In each case, the basic goal is to perform as well, or nearly as well, as the best predictor in a family of functions, such as neural networks or decision trees. For a given model and function family, if this goal can be achieved under some reasonable constraints, the family is said to be learnable in the model.

[*]Supported by National Science Foundation grant CCF-1848966

Machine-learning theorists are typically able to transform questions about the learnability of a particular function family into problems that involve analysing various notions of dimension that measure some aspect of the family's complexity. For example, the appropriate notion for analysing PAC learning is known as the Vapnik–Chervonenkis (VC) dimension [8], and, in general, results relating learnability to complexity are sometimes referred to as Occam's razor theorems [9]. These notions of dimension happen to be simple enough to leave no room for the spectre of unprovability to manifest itself. But Ben-David and colleagues show that machine learning cannot always escape this fate. They introduce a learning model called estimating the maximum (EMX), and go on to discover a family of functions whose learnability in EMX is unprovable in standard mathematics.

Ben-David et al. describe an example EMX problem: targeting advertisements at the most frequent visitors to a website when it is not known in advance which visitors will visit the site. The authors formalize EMX as a question about a learner's ability to find a function, from a given family, whose expected value over a target distribution is as large as possible. EMX is actually quite similar to the PAC model, but the slightly different learning criterion surprisingly connects it to the continuum hypothesis and brings unprovability into the picture.

The authors' proof involves a beautiful connection between machine learning and data compression that was first observed [10] in the 1980s. The intuition is that, if a training sample labelled by a function from some family can always be compressed, the family must in some sense have low complexity, and therefore be learnable. Moreover, certain learning algorithms can be used to compress data. The authors introduce monotone compression — a variant of compression that they show to be appropriate for characterizing the learnability of particular function families in EMX.

Ben-David and colleagues then prove that the ability to carry out a weak form of monotone compression is related to the size of certain infinite sets. The set that the authors ultimately use in their work is the unit interval, which is the set of real numbers between 0 and 1. Their results imply that the finite subsets of the unit interval have monotone-compression schemes, and therefore are learnable in EMX, if and only if the continuum hypothesis is true, which is known to be unprovable.

Because EMX is a new model in machine learning, we do not yet know its usefulness for developing real-world algorithms. So these results might not turn out to have practical importance. But we do now know that we should be careful when introducing new models of learning. Moreover, we might need to look again at the many subtleties that can come up, even in established learning models.

Machine learning has matured as a mathematical discipline and now joins the many subfields of mathematics that deal with the burden of unprovability and the unease that comes with it. Perhaps results such as this one will bring to the field of machine learning a healthy dose of humility, even as machine-learning algorithms continue to revolutionize the world around us.

Lev Reyzin is in the Department of Mathematics, Statistics, and Computer Science, University of Illinois at Chicago, Chicago, Illinois 60607, USA. e-mail: lreyzin@uic.edu

References

1. Gödel, K. Monatsh. Math. 38, 173–198 (1931).
2. Gödel, K. The Consistency of the Continuum Hypothesis (Princeton Univ. Press, 1940).
3. Cohen, P. J. Proc. Natl Acad. Sci. USA 50, 1143–1148 (1963).
4. Cohen, P. J. Proc. Natl Acad. Sci. USA 51, 105–110 (1964).
5. Ben-David, S., Hrubeš, P., Moran, S., Shpilka, A. & Yehudayoff, A. Nature Mach. Intell. https://doi.org/10.1038/s42256-018-0002-3 (2019).
6. Viola, P. & Jones, M. J. Int. J. Comput. Vis. 57, 137–154 (2004).
7. Valiant, L. G. Commun. ACM 27, 1134–1142 (1984).
8. Vapnik, V. N. & Chervonenkis, A. Y. Theory Probab. Appl. 16, 264–280 (1971).
9. Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. Inf. Process. Lett. 24, 377–380 (1987).
10. Littlestone, N. & Warmuth, M. K. Relating Data Compression and Learnability, tech. rep. (Univ. California Santa Cruz, 1986).