In: Emre Bayamlıoğlu, Irina Baraluic, Liisa Janssens and Mireille Hildebrandt (eds), BEING PROFILED:COGITAS ERGO SUM. 10 Years of Profiling the European Citizen, 2018: Amsterdam University Press, 102-105. DOI 10.5117/9789463722124/CH17.

Preregistration of machine learning research design. Against P-hacking

*Mireille Hildebrandt**

Abstract

This brief provocation targets the mantra of the trade-off between accuracy and interpretability: the higher the accuracy, the lower the interpretability. It seems that this trade-off appeals to a deep-seated desire for magical thinking; the lure of things that work well even if we have no idea why. The suggestion is that in the realm of specific types of machine learning (ML), neither causality nor reasoning matters. Correlation and prediction are all that counts. The story goes that not just lay people, those using an ML application or those targeted by its decisions, but even those who developed the application cannot explain why it gets things right.

I will confront this narrative from the perspective of ML research design, arguing that accuracy depends on the appropriate selection and curation of training and validation data, a properly articulated machine-readable task, a well-developed hypotheses space, and the selection of a relevant performance metric. This entails that accuracy in the realm of data should not be conflated with correctness in the realm of atoms. In other words, if we cannot explain why an ML application gets things right, we cannot be sure that it gets things right.

Keywords: machine learning, research design, methodological integrity, purpose binding, pre-registration, accuracy vs interpretability trade-off

Introduction

ML is based on the idea that intelligence concerns the ability to learn from experience, rather than the ability to apply ready-made knowledge. In that sense it favours inductive rather than deductive inferences. In the domain of artificial intelligence many voices now warn against overestimating the effectiveness of inductive learning, without however disqualifying its potential achievements (Brooks 2017, 2018; Marcus 2018).

The Mechanics of ML

It is interesting to note that human intelligence thrives on what Peirce called abductive inferences (Peirce and Turrisi 1997, 241-56), which are neither inductive nor deductive. Abductive inferencing basically entails an informed guess as to the explanation of a set of observations. Building on Peirce, scientific research can be framed as starting with an abduction based on observation, generating an explanation (theory) from which a hypothesis (prediction) is deduced about subsequent observations, after which the prediction can be inductively tested against new observations. Building on Popper's theory of falsification,[1] hypotheses should be developed in a way that enables the rejection of the explanation – not merely its verification. A theory that explains why all swans are white should not just be verified by detecting ever more white swans, but tested against its potential falsification by searching for black swans.

ML has been defined as 'improving automatically with experience' (Mitchell 1997, 1). More precisely '[a] computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E' (Mitchell 1997, 2). A crucial role is played by the so-called hypotheses space, i.e. a set of mathematical functions to be tested as accurate descriptions of patterns in the data on which the algorithm is trained (the training data). These hypotheses can be seen as abductively generated by the human developer of the ML research design, even if the system can be designed such that it generates further hypotheses (mathematical functions). Feeding the system with 'big data' can be seen as inductive testing. What is missing here, is the explanation. Normally, abduction generates an explanation, from which hypotheses can be deduced, which are then tested. In the case of ML, abduction does not concern an explanation but a set of mathematical functions that may or may not accurately describe statistical patterns in the data. The missing link in

ML is the explanation or interpretation of the world that is supposedly represented by the data (Anderson 2008; Hofman, Sharma, and Watts 2017; Pearl and Mackenzie 2018; Hildebrandt 2015, 37-57).

Machine experience is limited to digital data

To understand what this means, it is pivotal to note that the 'experience' of the machine consists of digital data (when referring to data I mean digital data). Machines do not experience light, temperature, sound or touch, nor do they have any understanding of natural language; they merely process data. Whereas such data may be a trace of, or a representation of light, temperature, sound, touch or text, it should not be confused with what it traces or represents. The choice of data, the kind of translation it implies, the type of error it may contain and the way it has been curated all impact the accomplishments of ML. For instance, developers may use 'low hanging fruit', i.e. data that is easily available but not necessarily relevant or complete. This may result in bad ML applications (garbage in, garbage out or GIGO), and can be remedied either by obtaining other and/or more data, or by accepting that the data needed for the task cannot be harvested.

Before training their learning algorithm ('the learner') on the data, developers will attempt to remove irrelevant or incorrect 'noise', depending on the goal of the operation. They always run the risk of removing highly relevant data, even though the risk can be reduced by testing on differently curated data sets.

However, we must also remind ourselves that data-driven applications necessarily feed on the reduction of real world experience to sensor data or to natural language processing, translating the flux of a life world into variables that enable measurement and calculation. Such translation may lead to computational artefacts (bugs), taking note that any quantification requires prior qualification (as the same type of data). In the case of real-time interaction with data-driven systems this may lead to strange responses, such as taking a pedestrian for a plastic bag – resulting in the death of the pedestrian (Gibbs 2018).

Finally, bias in the real world may be reinforced due to the use of statistics, often resulting in what has been coined 'disparate accuracy', which may further entrench existing discrimination (Barocas and Selbst 2016).

The mathematical target function

A more fundamental point is that the goal of ML can be summarized as detecting relevant 'bias' in a dataset, where 'bias' refers to the patterned deviation from a random distribution (Mitchell 1997, 20-51). Unless a dataset has a random distribution – which is highly improbable – an algorithm that is trained to detect 'bias' will always come up with patterns. The more interesting point then is to figure out whether the bias is either spurious or relevant.

The detection of relevant 'bias' in a dataset can be defined as the approximation of a mathematical target function that best describes the relationship between input and output data. To enable such approximation a so-called hypothesis space is developed with sets of mathematical functions that may or may not succeed in describing this relationship. The better the function expresses this relationship, the higher the accuracy of the system.

Machine learning can thus also be defined as a type of compression. Instead of a huge set of data, we now have a mathematical function that describes the data, noting that the same data can be compressed in different ways, depending on the task and/or the performance metric. As should be clear, the shaping of the hypotheses space is critical for a proper description of the data; a well-developed space is hoped to generate a hypothesis that does well if tested on new data.

A core problem is that a detailed hypothesis space may do very well on the training set, but very bad on out-of-sample test data, as it 'overfits' with the training data in a way that weakens its ability to generalize to new data. A less detailed hypothesis space, however, may generate a function that does well in generalizing to new data, but 'overgeneralizes' in a way that results in overlooking crucial connections, thus missing relevant features. If the environment is static and translates well into data, these problems can be resolved by iterant experimentation. If the environment is dynamic such iteration may not work.

Especially where human agents and societal institutions respond to their behavioural data traces being tested, machine learning algorithms face a double feedback loop as the anticipation of human and societal agents may invalidate the findings of 'the learner'. That is why a game with fixed and closed rules such as Go can be learnt based on the brute force (computing power) of programs such as AlphaZero (Collados 2017), whereas the adaptive nature of complex social phenomena remains elusive even when a system is trained on unprecedented volumes of data. This means that the fundamental assumption that underlies any ML system, i.e. that reality is governed by mathematical functions, does not necessarily hold for human society.

P-hacking

Next to bias in the data and the hypotheses space, the outcome of an ML application may be biased due to cherry picking with regard to the performance metric (P). This metric determines the accuracy of the system, based on the task (T) the system aims to perform and the data (E) it trains on. As one can imagine, if some metric P1 achieves 67% accuracy, whereas another metric P2 achieves 98% accuracy, the temptation to use only P2 and boast high accuracy is formidable. I will call this P-hacking, as it seems to be the twin sister of p-hacking (Gollnick in this volume; Berman et al. 2018). Especially in systems that are difficult to interpret high accuracy does not mean much, as the system may be getting things wrong despite the accuracy. The opacity of the underlying causality (e.g. in the case of medical diagnosis) or reasoning (e.g. in the case of quantified legal prediction) easily hides potential misfits.

For instance, a system that was meant to predict death after pneumonia qualified chest pain, heart disease and asthma as indicators of low risk, contrary to reality (Caruana et al. 2015). Any doctor can tell you that these three indicators correlate with high risk. Nevertheless, the accuracy was very high – within the dataset on which the algorithm was trained. Because the indicators were visible it was easy to figure out what went wrong: patients with chest pain, heart disease or asthma are sent to a hospital and monitored so well that their risk is lowered due to the fact that they are treated as high risk. If, however, the rating had been based on a neural network it might have been far less obvious which of the features caused the system to attribute a low risk. This makes reliance on such systems dangerous, as it may take time (and unnecessary death) before the mistake is uncovered.

So what?

Based on the analysis of ML research design, I propose that whoever puts an ML application on the market should pre-register the research design that was used to develop the application (including subsequent updates). This will contribute to the contestability of claims regarding the safety, security, and reliability of such applications, while also enabling the contestability of decisions based on such applications in terms of potential violations of fundamental rights such as privacy, data protection, freedom of expression, presumption of innocence and non-discrimination. If such preregistration were to become a requirement, e.g. in an updated Machinery Directive,[2] it would also be a good example of 'legal protection by design' (Hildebrandt 2015, 218).

Notes

*Mireille Hildebrandt is a Research Professor on 'Interfacing Law and Technology' at the Faculty of Law and Criminology of Vrije Universiteit Brussel. She also holds the Chair of 'Smart Technologies, Data Protection and the Rule of Law' at the Science Faculty of Radboud University, Nijmegen.

[1] Peirce's fallibilism, as well as Popper's related theory of falsification demand that scientific theory is restricted to explanations that enable testing in a way that enables their refutation.

[2] DIRECTIVE 2006/42/EC of the European Parliament and of the Council of 17 May 2006 on machinery, and amending Directive 95/16/EC (recast).

References

Anderson, Chris. 2008. "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." Wired Magazine 16(7).

Barocas, Solon, and Andrew D. Selbst. 2016. "Big Data's Disparate Impact." California Law Review 104: 671–732.

Berman, Ron, Leonid Pekelis, Aisling Scott, and Christophe Van den Bulte. 2018. 'P-Hacking and False Discovery in A/B Testing'. Rochester, NY: Social Science Research Network. https://papers.ssrn.com/abstract=3204791.

Brooks, Rodney. 2017. "Machine Learning Explained." MIT RETHINK. Robots, AI, and Other Stuff (blog). August 28, 2017. http://rodneybrooks.com/forai-machine-learning-explained/.

——. 2018. "My Dated Predictions – Rodney Brooks." MIT RETHINK (blog). January 1, 2018. https://rodneybrooks.com/my-dated-predictions/.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad. 2015. "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-Day Readmission." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1721–1730. KDD '15. New York, NY, USA: ACM. doi: 10.1145/2783258.2788613.

Collados, Jose Camacho. 2017. "Is AlphaZero Really a Scientific Breakthrough in AI?" Medium (blog). December 11, 2017. https://medium.com/@josecamachocollados/is-alphazero-really-a-scientific-breakthrough-in-ai-bf66ae1c84f2.

Gibbs, Samuel. 2018. "Uber's Self-Driving Car Saw the Pedestrian but Didn't Swerve – Report." The Guardian, May 8, 2018. http://www.theguardian.com/technology/2018/may/08/ubers-self-driving-car-saw-the-pedestrian-but-didnt-swerve-report.

Hildebrandt, Mireille. 2015. Smart Technologies and the End(s) of Law. Novel Entanglements of Law and Technology. Cheltenham: Edward Elgar.

Hofman, Jake M., Amit Sharma, and Duncan J. Watts. 2017. "Prediction and Explanation in Social Systems." Science 355 (6324): 486–88. doi: /10.1126/science.aal3856.

Marcus, Gary. 2018. "Deep Learning: A Critical Appraisal". arxiv.org/abs/1801.00631.

Mitchell, Thomas. 1997. Machine Learning. New York: McGraw-Hill Education.

Pearl, Judea, and Dana Mackenzie. 2018. The Book of Why: The New Science of Cause and Effect. New York: Basic Books.

Peirce, Charles S., and Patricia Ann Turrisi. 1997. Pragmatism as a Principle and Method of Right Thinking: The 1903 Harvard Lectures on Pragmatism. Albany: State University of New York Press.