LIZ KUBALL

Researchers at TAE Technologies in California and at Google are using machine learning to optimize equipment that produces a high-energy plasma.

Three pitfalls to avoid in machine learning

As scientists from myriad fields rush to perform algorithmic analyses, Google's **Patrick Riley** calls for clear standards in research and reporting.

Machine learning is driving discovery across the sciences. Its powerful pattern finding and prediction tools are helping researchers in all fields — from finding new ways to make molecules and spotting subtle signals in assays, to improving medical diagnoses and revealing fundamental particles.

Yet, machine-learning tools can also turn up fool's gold — false positives, blind alleys and mistakes. Many of the algorithms are so complicated that it is impossible to inspect all the parameters or to reason about exactly how the inputs have been manipulated. As these algorithms begin to be applied ever more widely, risks of misinterpretations, erroneous conclusions and wasted scientific effort will spiral.

These problems are not new. The machine-learning field has chastened itself for decades with the 'tank problem'. The ▶

ATUL LOKE/THE NEW YORK TIMES/RED/EYEVINE

An eye examination at Aravind hospital in Madurai, India, where staff and Google researchers are trying to automate diagnoses of blindness caused by diabetes.

original study seems to have arisen in the 1960s (ref. 1 is the earliest plausible reference known for this study; with thanks to software engineer Jeff Kaufman) and is obscured by the mists of time, but the story goes like this. Researchers wrote an algorithm to spot tanks in photographs provided by the military. The model found the tanks successfully in test images. But it failed later with future real photos in the field. Why? The details vary in the retelling, but the pictures it was trained on contained other patterns — tanks emerging in the morning light, or under clouds. So, it was other factors such as these that drove the algorithm, not the presence of tanks.

Similar confusions are causing soul-searching today[2]. Many machine-learning papers fail to perform an adequate set of experiments. Standards for review are inconsistent. And competition is encouraging some researchers to cut corners and skip checks once they think they have the answer they want.

We cannot predict all the difficulties that will arise with each analysis. But, as a minimum, researchers bringing machine learning to their fields should familiarize themselves with the common pitfalls and the practices they can use to detect and avoid them.

To illustrate, I highlight three problems in machine-learning analyses that we have faced and overcome in the Google Accelerated Science team.

THREE PROBLEMS

Splitting data inappropriately. When building models, machine-learning practitioners typically break data into training and test sets. The training set teaches the model, and the model's performance is evaluated by how well it describes the test set. Researchers typically split the data at random. But data in real life are rarely random. They might contain trends in time — such as from changes in the way the data were gathered, or from varying choices over what information to collect.

Such historical patterns are buried in data sets on molecules, for example, which are being screened virtually by machine-learning algorithms to find candidates for drugs. The challenge is to predict how effectively a hypothetical molecule will, for example, be absorbed into the body or decrease inflammation. Screening starts with data on molecules that either do or do not have the desired effect. But the contexts in which the data were collected might be different from how the machine-learning model is to be used.

"Competition is encouraging some researchers to cut corners and skip checks."

For example, a model might be built on a set of molecules that is publicly available, but then used on a different, proprietary set. And chemists' gazes often switch from certain groups of molecules to others, when promising leads are examined and discarded. Thus, researchers often overestimate how well the model will do in practice[3]. This can lead to inflated expectations and it wastes time and money on poorly chosen molecules. Many model builders (myself included) have fallen into this trap.

In other words, the question you want to answer should affect the way you split your data. For the model to predict the effect of adding a couple of atoms to a molecule, each molecule in the test set should have a partner in the training set that is a couple of atoms different. If you want to get good predictions on chemically diverse molecules, each molecule in the test set should be unlike everything in the training set. The 'right' way to split data might not be obvious, but careful consideration and trying several approaches will give more insight.

Hidden variables. In an ideal experiment, the researcher changes only the variables of interest and fixes all others. This level of control is often impossible in the real world.

The accuracy of equipment drifts over time, batches of reagents differ, one experimental condition is performed before another, and results can even be skewed by the weather. Such uncontrolled variables can be pernicious in machine-learning models.

For example, my team at Google has been working with the nuclear-fusion start-up firm TAE Technologies in Foothill Ranch, California, to optimize an experiment for producing high-energy plasma[4]. We built models to try and understand the best equipment settings for the plasma machine. There were hundreds of control parameters, from when to energize electrodes to which voltage to set on the magnets. A range of measurements was recorded, including temperatures and spectra.

We took data from thousands of runs of the plasma machine over many months. The settings varied as the device was tuned and modified and as components wore out and different ideas were tried. We were pleased when we arrived at a model that predicted well, for given settings, whether the plasma's energy would be high. Soon, it became obvious that our predictions were not based on what we thought.

When we trained the model again, with the time of the experiment as the only input, rather than all the settings of the machine, we got similar predictive power. Why? We think that our first model locked on to time trends, rather than physical phenomena. As the experiments ran, there were periods of time when the machinery was functioning well and periods when it wasn't. Therefore, the time at which the experiment was done gives you some information about whether the plasma produced was high energy or not. Furthermore, it's possible to predict roughly when an experiment was done from the setting of the control parameters — there were time trends in how those were varied, too.

Hidden variables can also stem from the layout of experiments. For example, we are working with many collaborators on interpreting microscope images, including the New York Stem Cell Foundation Research Institute in New York City. The images include arrays of biological experiments on plates — typically a grid of wells containing cells and liquids. The goal is to spot wells with certain characteristics, such as a change in appearance of the cells after a chemical treatment. But biological variation means that each plate will always look slightly different. And there can be variation across a single plate. The edges often look different from the centre, for example, if more liquid has evaporated in peripheral wells or if the plate was tilted.

A machine-learning algorithm can easily pick up on these unintentional variations. For instance, the model might just identify which wells are on the edge of the plate. A simple way to check if this has happened is to ask the model to predict other things, such as the location on the plate, which plate it is and which batch the image is from. If it can do this, be suspicious of your results.

The take-home lesson is: use multiple machine-learning models to detect unexpected and hidden variables. One model focuses on the question you care about — is the plasma high or low energy; are the cells healthy or sick? Other models flush out the confounders. If the latter result is strong, normalize your data, run further experiments or temper your conclusions.

Mistaking the objective. Machine-learning algorithms require researchers to specify a 'loss function', which determines the severity of various errors — such as whether it is better to make two errors of 1% each, or a single error of 2%. Practitioners tend to use a small set of functions that can fail to capture what they really care about.

For example, we have been using machine learning to assist in solving partial differential equations[5]. These formulae are common across the sciences, including in fluid dynamics, electromagnetism, materials science, astrophysics and economic modelling. Often, they must be solved numerically, and we trained models to provide better accuracy at limited resolution.

> *"Uncontrolled variables can be pernicious in machine-learning models."*

We started with an equation to describe how water waves propagate in one dimension. The algorithm was tasked with repeatedly predicting the next time step from the current one. We had two slightly different formulations and trained models on both. According to our loss functions, the two models were equally good. However, one produced nonsense while the other stayed close to the desired result.

Why? The loss function controlling the learning was considering only the error of the next step, not the validity of the solution over many steps, which is what we really want.

Diverging goals also cropped up in our work on machine screening for diabetic retinopathy[6], a complication of diabetes and a leading cause of preventable blindness in the world. The condition can be treated effectively if it is detected early, from images of the back of the eye. As we gathered data and had ophthalmologists offer diagnoses based on the images, we asked our machine-learning tools to predict what the ophthalmologist would say. Two issues emerged.

First, the ophthalmologists often disagreed on the diagnosis. Thus, we realized that we could not base our model on a single prediction. Nor could we use a majority vote, because, when it comes to medical accuracy, sometimes the minority opinion is the right one. Second, the diagnosis of a single disease was not actually the real objective. We should have been asking: 'should this patient see a doctor?' We therefore expanded our goal from the diagnosis of a single disease to multiple diseases.

It is easy for machine-learning practitioners to become fixated on an 'obvious' objective in which the data and labels are clear. But they could be setting up the algorithm to solve the wrong problem. The overall aim must be kept in mind, or we will produce precise systems that answer the wrong questions.

WHAT NEXT?

First, machine-learning experts need to hold themselves and their colleagues to higher standards. When a new piece of lab equipment arrives, we expect our lab mates to understand its functioning, how to calibrate it, how to detect errors and to know the limits of its capabilities. So, too, with machine learning. There is no magic involved, and the tools must be understood by those using them.

Second, different disciplines need to develop clear standards for how to perform and report on machine learning in their areas. The appropriate controls, soundness checks and error measurements will vary from field to field, and these need to be spelt out clearly so that researchers, reviewers and editors can encourage good behaviour.

Third, the education of scientists in machine learning needs to include these broader issues. Although some resources exist (such as http://ai.google/education), we need to do more. We often teach the algorithms and tools, but students need to learn more about how to apply their algorithms and question them appropriately.

We are at an amazing point — computational power, data and algorithms are coming together to produce great opportunities for discoveries with the assistance of machine learning. It is our responsibility as a scientific community to ensure that we use this opportunity well. ■

Patrick Riley *is a principal engineer and the senior researcher on the Google Accelerated Science team at Google, Mountain View, California, USA.*
e-mail: pfr@google.com

1. Kanal, L. N. & Randall, N. C. In *Proc. 1964 19th ACM National Conf.* 42.501–42.5020 (ACM, 1964).
2. Lipton, Z. C. & Steinhardt, J. Preprint at arXiv. http://arxiv.org/abs/1807.03341 (2018).
3. Sheridan, R. P. *J. Chem. Inform. Model.* **53,** 783–790 (2013).
4. Baltz, E. A. *et al. Sci. Rep.* **7**, 6425 (2017).
5. Bar-Sinai, Y., Hoyer, S., Hickey, J. & Brenner, M. P. *Proc. Natl Acad. Sci. USA* https://doi.org/10.1073/pnas.1814058116 (2019).
6. Gulshan, V. *et al. J. Am. Med. Assoc.* **316**, 2402–2410 (2016).