Workarounds in decision support systems

(work in progress)

Tom Slee

August 2, 2019

Researchers from Harvard and MIT warn us about "Adversarial attacks on medical machine learning", in which "a small, carefully designed change in how inputs are presented to a system… completely alter its output, causing it to arrive at manifestly wrong conclusions." (Link) I think that's only half the story…

Contents

Adversarial attacks and workarounds in medical machine learning

It's not like the Harvard/MIT article is *purely* about adversarial attacks. Early on the authors point out that the American medical claims approval process is plagued by many competing financial interests, "providers [i.e. doctors and hospitals] seeking to maximize and payers [insurance companies] seeking to minimize reinbursement."

But the remainder of the paper focuses almost exclusively on the problem of adversarial attacks by healthcare providers. Deep learning algorithms may be uncannily accurate over a well-defined data set, but they are also fragile if a sample steps in an unexpected direction, and this fragility leaves them open to "attack". Most blatantly, the authors show, doctors could modify medical images to change the diagnosis, even as the image appears unchanged to the human eye. Then there are grey areas, like rotations of an image, or a careful choice of words in notes, that pushes the system to give unexpected results. How should the medical system handle a techique that seems so accurate and yet so easily misled?

Workarounds

Adversarial attacks are "inputs to a machine learning model that are intentionally crafted to force the model to make a mistake": that is, you take an input that the model classifies correctly and you tweak it so that it is classified incorrectly. *Workarounds* are the opposite of adversarial attacks: inputs that are intentionally crafted to force the model to *correct* a mistake. That is, you take an input that is incorrectly classified by the model and tweak it so that it is correctly classified.

Even the best machine learning systems are, after all, statistical engines with non-zero error rates, so incorrectly classified examples will always be with us. And of course in the real world there are ambiguities in any data set, to the boundaries between classes are not as sharp as a labelled training set may make it seem.

Adversarial examples may be novel, but in the automated workplace workarounds have become routine, even if they are often overlooked. Automated systems shift decision-making from "front line" workers to a system, but as these systems have entered the workplace, many professionals have learned a new and often unacknowledged skill: they have become experts at making the system work properly.

As just one example, legal scholar Jennifer Raso explored the changing jobs of Ontario case workers when a new automated decision system was introduced. Instead of making their own judgements about client needs, case workers are now to enter data into the system, which would make a decision according to well-defined rules. She writes about how the social service case workers responded:

#+begin$_{\text{kquote}}$ While new technologies may attempt to deskill and decentre front-line decision-makers, transforming them into data entry clerks, caseworkers learn how to expertly translate and input client data to produce decisions that more closely match their interpretation of clients' needs and welfare laws. #+end$_{\text{quote}}$

The idea will be familiar to anyone who has read James C. Scott's Seeing Like a State: certain schemes to improve the human condition assume, and impose, a simplified and idealized way of working and living that neglects the unruly irregularities of real life. Managed systems commonly rely on a certain level of workaround to even function: hence the effectiveness of the "work to rule" as a form of protest.

American healthcare systems have their own workarounds. The rigidities of automated

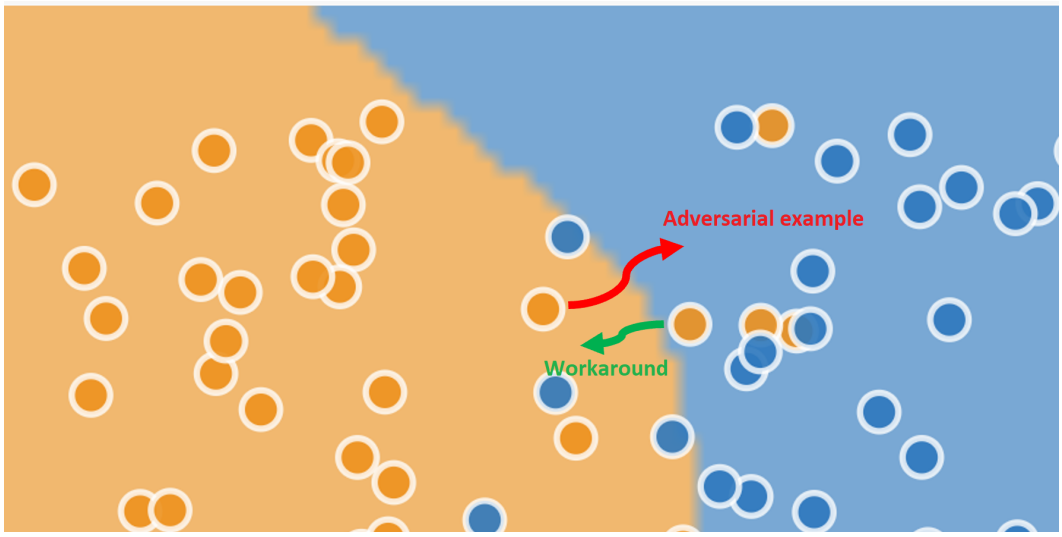

Figure 1: Schematic view of a machine learning model. The data points are shown with their "true" values of blue or orange, and the background shade marks the areas that the model categorizes as "blue" or "orange". The adversarial example changes a correctly-classified "orange" data point into an incorrectly-classified one, while the *workaround* changes an incorrectly-classified data point into a correctly classified one. The image was built from Tensorflow Playground)

billing systems recently produced a memorable essay by Atul Gawande entitled Why Doctors Hate Their Computers, documenting many of the challenges experienced by physicians as new decision support systems have been introduced.

I said that adversarial attacks and workarounds are opposites, but in another way they are very similar. Surprisingly there is no way to distinguish unambiguously between an adversarial attack and a workaround in real-world environments within machine learning theory. Attempts to define "attacks" refer to "intent", but who is to say what that is? Many examples focus on cases where "ground truth" is unambiguous (a picture of a bus is "obviously" not a picture of a monkey). Examples that are clearly artificial and intentional are reflections of lab research, not of how systems will play out in the wild. Some machine learning practitioners have gone so far as to define "attacks" as any attempt by a subject to influence the outcome of a machine learning system, which makes an attacker of anyone who has put effort into writing their resume in the hopes of getting a job. But don't take all this from me: if you want to hear an expert opinion on this definition problem, see this talk by David Evans.

When it comes down to it, both adversarial examples and workarounds are deliberate changes to inputs in order to generate desired outputs. Machine learning systems are statistical, and however accurate they may be over a large data set, statistics will not say whether any one individual answer is correct. For complex tasks such as medical diagnosis certified "correct" answers are often not available. And if the "correct" answer is not well-defined, who is to say whether an intentional change is correcting a mistake or corrupting a correct answer?

Will workarounds be needed for the next generation of deep learning medical systems? We can't say for sure yet, but history suggests that they will. There is a long history of technical innovations that are designed around idealized and simplified models of behaviour, and which underestimate the complexities of "edge cases" in real life. Witness the over-enthusiasm around self-driving cars a couple of years ago, that is now being confronted with the messiness of reality.

The article is not completely silent on the topic of workaroudns: it does reference a paper from almost 20 years ago titled "Physician Manipulation of Reimbursement Rules for Patients: Between a Rock and a Hard Place", which says this:

> It has been suggested that some insurers are "gaming" patients and physicians—tricking them into paying for covered services by routinely denying coverage but then approving services that are subsequently appealed, knowing that time and other constraints will prevent some appeals.

So doctor workarounds may be responses not only to technical deficiencies in the system, but also to real or perceived bad faith on the part of the insurers (or their suppliers) who are responsible for the system itself. And why should doctors have faith in these systems? Big money is at work on both sides, and the incentives for insurance companies to "optimize" is at least as strong as for the doctors. They face incentives to fix inaccuracies that lead to overbilling, but not those blind spots that lead to underpayments; to label edge cases in training sets in such a way as to minimize payments; to redefine payment schedules around the observed behaviour of the systems.

When it comes to interventions, Finlayson et al make two recommendations. The first is to procrastinate: to avoid stifling innovation by prematurely enforcing demands for robustness. The second is to increase supervision of medical practitioners, to check that they do not "tamper" with the data. Such an approach not only neglects any consideration

of insurance companies, but it also rules out workarounds. It removes any room for doctors' discretion and judgement, and reduces the role of physician to that of a managed and supervised data entry technician.

As I've written elsewhere, most machine learning systems of any interest are *incentive incompatible*. The subjects who provide the inputs and the consumers of the outputs have different and conflicting interests. And in such an arrangement additional rules are not just likely, but inevitable. The problem with procrastination is that it favours the consumers and, more than anyone, the providers of the system. A natural response to problems is to demand more complete data, better data, and closer supervision of data entry. Without a check on insurance companies, medical machine learning systems will not be a cure for a damaged health system.