Methods that help users decide whether to trust a machine-learning model's predictions can perpetuate biases

June 1 2022, by Adam Zewe

Credit: Pixabay/CC0 Public Domain

When the stakes are high, machine-learning models are sometimes used

to aid human decision-makers. For instance, a model could predict which law school applicants are most likely to pass the bar exam to help an admissions officer determine which students should be accepted.

These models often have millions of parameters, so how they make predictions is nearly impossible for researchers to fully understand, let alone an admissions officer with no machine-learning experience. Researchers sometimes employ explanation methods that mimic a larger model by creating simple approximations of its predictions. These approximations, which are far easier to understand, help users determine whether to trust the model's predictions.

But are these explanation methods fair? If an explanation method provides better approximations for men than for women, or for white people than for Black people, it may encourage users to trust the model's predictions for some people but not for others.

MIT researchers took a hard look at the fairness of some widely used explanation methods. They found that the approximation quality of these explanations can vary dramatically between subgroups and that the quality is often significantly lower for minoritized subgroups.

In practice, this means that if the approximation quality is lower for female applicants, there is a mismatch between the explanations and the model's predictions that could lead the admissions officer to wrongly reject more women than men.

Once the MIT researchers saw how pervasive these fairness gaps are, they tried several techniques to level the playing field. They were able to shrink some gaps, but couldn't eradicate them.

"What this means in the real-world is that people might incorrectly trust predictions more for some subgroups than for others. So, improving

explanation models is important, but communicating the details of these models to end users is equally important. These gaps exist, so users may want to adjust their expectations as to what they are getting when they use these explanations," says lead author Aparna Balagopalan, a graduate student in the Healthy ML group of the MIT Computer Science and Artificial Intelligence Laboratory (CSAIL).

Balagopalan wrote the paper with CSAIL graduate students Haoran Zhang and Kimia Hamidieh; CSAIL postdoc Thomas Hartvigsen; Frank Rudzicz, associate professor of computer science at the University of Toronto; and senior author Marzyeh Ghassemi, an assistant professor and head of the Healthy ML Group. The research will be presented at the ACM Conference on Fairness, Accountability, and Transparency.

High fidelity

Simplified explanation models can approximate predictions of a more complex machine-learning model in a way that humans can grasp. An effective explanation model maximizes a property known as fidelity, which measures how well it matches the larger model's predictions.

Rather than focusing on average fidelity for the overall explanation model, the MIT researchers studied fidelity for subgroups of people in the model's dataset. In a dataset with men and women, the fidelity should be very similar for each group, and both groups should have fidelity close to that of the overall explanation model.

"When you are just looking at the average fidelity across all instances, you might be missing out on artifacts that could exist in the explanation model," Balagopalan says.

They developed two metrics to measure fidelity gaps, or disparities in fidelity between subgroups. One is the difference between the average

fidelity across the entire explanation model and the fidelity for the worst-performing subgroup. The second calculates the absolute difference in fidelity between all possible pairs of subgroups and then computes the average.

With these metrics, they searched for fidelity gaps using two types of explanation models that were trained on four real-world datasets for high-stakes situations, such as predicting whether a patient dies in the ICU, whether a defendant reoffends, or whether a law school applicant will pass the bar exam. Each dataset contained protected attributes, like the sex and race of individual people. Protected attributes are features that may not be used for decisions, often due to laws or organizational policies. The definition for these can vary based on the task specific to each decision setting.

The researchers found clear fidelity gaps for all datasets and explanation models. The fidelity for disadvantaged groups was often much lower, up to 21 percent in some instances. The law school dataset had a fidelity gap of 7 percent between race subgroups, meaning the approximations for some subgroups were wrong 7 percent more often on average. If there are 10,000 applicants from these subgroups in the dataset, for example, a significant portion could be wrongly rejected, Balagopalan explains.

"I was surprised by how pervasive these fidelity gaps are in all the datasets we evaluated. It is hard to overemphasize how commonly explanations are used as a 'fix' for black-box machine-learning models. In this paper, we are showing that the explanation methods themselves are imperfect approximations that may be worse for some subgroups," says Ghassemi.

Narrowing the gaps

After identifying fidelity gaps, the researchers tried some machine-learning approaches to fix them. They trained the explanation models to identify regions of a dataset that could be prone to low fidelity and then focus more on those samples. They also tried using balanced datasets with an equal number of samples from all subgroups.

These robust training strategies did reduce some fidelity gaps, but they didn't eliminate them.

The researchers then modified the explanation models to explore why fidelity gaps occur in the first place. Their analysis revealed that an explanation model might indirectly use protected group information, like sex or race, that it could learn from the dataset, even if group labels are hidden.

They want to explore this conundrum more in future work. They also plan to further study the implications of fidelity gaps in the context of real-world decision making.

Balagopalan is excited to see that concurrent work on explanation fairness from an independent lab has arrived at similar conclusions, highlighting the importance of understanding this problem well.

As she looks to the next phase in this research, she has some words of warning for machine-learning users.

"Choose the explanation model carefully. But even more importantly, think carefully about the goals of using an explanation model and who it eventually affects," she says.

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.