A technique to improve both fairness and accuracy in artificial intelligence

July 20 2022, by Adam Zewe

Credit: Pixabay/CC0 Public Domain

For workers who use machine-learning models to help them make decisions, knowing when to trust a model's predictions is not always an easy task, especially since these models are often so complex that their

inner workings remain a mystery.

Users sometimes employ a technique, known as selective regression, in which the model estimates its confidence level for each prediction and will reject predictions when its confidence is too low. Then a human can examine those cases, gather additional information, and make a decision about each one manually.

But while selective regression has been shown to improve the overall performance of a model, researchers at MIT and the MIT-IBM Watson AI Lab have discovered that the technique can have the opposite effect for underrepresented groups of people in a dataset. As the model's confidence increases with selective regression, its chance of making the right prediction also increases, but this does not always happen for all subgroups.

For instance, a model suggesting loan approvals might make fewer errors on average, but it may actually make more wrong predictions for Black or female applicants. One reason this can occur is due to the fact that the model's confidence measure is trained using overrepresented groups and may not be accurate for these underrepresented groups.

Once they had identified this problem, the MIT researchers developed two algorithms that can remedy the issue. Using real-world datasets, they show that the algorithms reduce performance disparities that had affected marginalized subgroups.

"Ultimately, this is about being more intelligent about which samples you hand off to a human to deal with. Rather than just minimizing some broad error rate for the model, we want to make sure the error rate across groups is taken into account in a smart way," says senior MIT author Greg Wornell, the Sumitomo Professor in Engineering in the Department of Electrical Engineering and Computer Science (EECS)

who leads the Signals, Information, and Algorithms Laboratory in the Research Laboratory of Electronics (RLE) and is a member of the MIT-IBM Watson AI Lab.

Joining Wornell on the paper are co-lead authors Abhin Shah, an EECS graduate student, and Yuheng Bu, a postdoc in RLE; as well as Joshua Ka-Wing Lee SM '17, ScD '21 and Subhro Das, Rameswar Panda, and Prasanna Sattigeri, research staff members at the MIT-IBM Watson AI Lab. The paper will be presented this month at the International Conference on Machine Learning.

To predict or not to predict

Regression is a technique that estimates the relationship between a dependent variable and independent variables. In machine learning, regression analysis is commonly used for prediction tasks, such as predicting the price of a home given its features (number of bedrooms, square footage, etc.) With selective regression, the machine-learning model can make one of two choices for each input—it can make a prediction or abstain from a prediction if it doesn't have enough confidence in its decision.

When the model abstains, it reduces the fraction of samples it is making predictions on, which is known as coverage. By only making predictions on inputs that it is highly confident about, the overall performance of the model should improve. But this can also amplify biases that exist in a dataset, which occur when the model does not have sufficient data from certain subgroups. This can lead to errors or bad predictions for underrepresented individuals.

The MIT researchers aimed to ensure that, as the overall error rate for the model improves with selective regression, the performance for every subgroup also improves. They call this monotonic selective risk.

"It was challenging to come up with the right notion of fairness for this particular problem. But by enforcing this criteria, monotonic selective risk, we can make sure the model performance is actually getting better across all subgroups when you reduce the coverage," says Shah.

Focus on fairness

The team developed two neural network algorithms that impose this fairness criteria to solve the problem.

One algorithm guarantees that the features the model uses to make predictions contain all information about the sensitive attributes in the dataset, such as race and sex, that is relevant to the target variable of interest. Sensitive attributes are features that may not be used for decisions, often due to laws or organizational policies. The second algorithm employs a calibration technique to ensure the model makes the same prediction for an input, regardless of whether any sensitive attributes are added to that input.

The researchers tested these algorithms by applying them to real-world datasets that could be used in high-stakes decision making. One, an insurance dataset, is used to predict total annual medical expenses charged to patients using demographic statistics; another, a crime dataset, is used to predict the number of violent crimes in communities using socioeconomic information. Both datasets contain sensitive attributes for individuals.

When they implemented their algorithms on top of a standard machine-learning method for selective regression, they were able to reduce disparities by achieving lower error rates for the minority subgroups in each dataset. Moreover, this was accomplished without significantly impacting the overall error rate.

"We see that if we don't impose certain constraints, in cases where the model is really confident, it could actually be making more errors, which could be very costly in some applications, like health care. So if we reverse the trend and make it more intuitive, we will catch a lot of these errors. A major goal of this work is to avoid errors going silently undetected," Sattigeri says.

The researchers plan to apply their solutions to other applications, such as predicting house prices, student GPA, or loan interest rate, to see if the algorithms need to be calibrated for those tasks, says Shah. They also want to explore techniques that use less sensitive information during the model training process to avoid privacy issues.

And they hope to improve the confidence estimates in selective regression to prevent situations where the model's confidence is low, but its prediction is correct. This could reduce the workload on humans and further streamline the decision-making process, Sattigeri says.

More information: Abhin Shah et al, Selective Regression Under Fairness Criteria. arXiv:2110.15403v3 [cs.LG], arxiv.org/abs/2110.15403

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

Provided by Massachusetts Institute of Technology