How machine learning can be fair and accurate

October 20 2021, by Aaron Aupperlee


FALAAH ARIF KHAN
Start

Achieving accuracy and fairness in machine learning systems intended for use in social decision making is possible but designing those systems requires venturing off the simple and obvious paths. Credit: Falaah Arif Khan

Carnegie Mellon University researchers are challenging a long-held assumption that there is a trade-off between accuracy and fairness when using machine learning to make public policy decisions.

As the use of machine learning has increased in areas such as criminal justice, hiring, health care delivery and social service interventions, concerns have grown over whether such applications introduce new or amplify existing inequities, especially among racial minorities and people with economic disadvantages. To guard against this bias, adjustments are made to the data, labels, model training, scoring systems and other aspects of the machine learning system. The underlying theoretical assumption is that these adjustments make the system less accurate.

A CMU team aims to dispel that assumption in a new study, recently published in *Nature Machine Intelligence*. Rayid Ghani, a professor in the School of Computer Science's Machine Learning Department and the Heinz College of Information Systems and Public Policy; Kit Rodolfa, a research scientist in ML; and Hemank Lamba, a post-doctoral researcher in SCS, tested that assumption in real-world applications and found the trade-off was negligible in practice across a range of policy domains.

"You actually can get both. You don't have to sacrifice accuracy to build systems that are fair and equitable," Ghani said. "But it does require you to deliberately design systems to be fair and equitable. Off-the-shelf

systems won't work."

Ghani and Rodolfa focused on situations where in-demand resources are limited, and machine learning systems are used to help allocate those resources. The researchers looked at systems in four areas: prioritizing limited mental health care outreach based on a person's risk of returning to jail to reduce reincarceration; predicting serious safety violations to better deploy a city's limited housing inspectors; modeling the risk of students not graduating from high school in time to identify those most in need of additional support; and helping teachers reach crowdfunding goals for classroom needs.

In each context, the researchers found that models optimized for accuracy—standard practice for machine learning—could effectively predict the outcomes of interest but exhibited considerable disparities in recommendations for interventions. However, when the researchers applied adjustments to the outputs of the models that targeted improving their fairness, they discovered that disparities based on race, age or income—depending on the situation—could be removed without a loss of accuracy.

Ghani and Rodolfa hope this research will start to change the minds of fellow researchers and policymakers as they consider the use of machine learning in decision making.

"We want the artificial intelligence, computer science and machine learning communities to stop accepting this assumption of a trade-off between accuracy and fairness and to start intentionally designing systems that maximize both," Rodolfa said. "We hope policymakers will embrace machine learning as a tool in their decision making to help them achieve equitable outcomes."

More information: Kit T. Rodolfa et al, Empirical observation of

negligible fairness–accuracy trade-offs in machine learning for public policy, *Nature Machine Intelligence* (2021). DOI: 10.1038/s42256-021-00396-x

Provided by Carnegie Mellon University