Stop gambling with black box and explainable models on high-stakes decisions

May 14 2019, by Ken Kingery

Cynthia Rudin. Credit: Duke University

As the buzzwords "machine learning" continue to grow in popularity, more industries are turning to computer algorithms to answer important questions, including high-stakes fields such as healthcare, finance and criminal justice. While this trend can lead to major improvements in these realms, it can also lead to major problems when the machine learning algorithm is a so-called "black box."

A black box is a machine learning program that does not explain how it reaches its conclusions, either because it is too complicated for a human to understand or because its inner workings are proprietary. In response to concerns that these types of models may include unjust inner workings—such as racism—another growing trend is to create additional models to "explain" these black boxes.

In a new editorial published in *Nature Machine Intelligence*, Cynthia Rudin, associate professor of computer science, electrical and computer engineering, mathematics, and statistical science at Duke University, argues that black box models must be abandoned for high-stakes decisions. Even when so-called explanation models are created, she says, decision-makers should be opting for interpretable models, which are completely transparent and easily understood by its users.

Explainable models are wrong

"Explainable" machine learning models are built in an attempt to understand what's going on inside of a black box. If it can produce the same results, people assume it's an accurate representation.

But it's not.

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Explainable algorithms provide explanations that are inaccurate by definition—otherwise the complex black box inner workings would be unnecessary. While an explanation model may produce similar or even exact results to the original black box algorithm, there is no way to know if it is using the same parameters or not.

"If a tenth of the explanations are wrong, you can't trust the explanations, and thus you can't trust the original black box," says Rudin. "If we can't know for certain whether our explanation is correct, we can't know whether to trust either the explanation or the original model."

More does not equal better

People typically assume that just because a machine learning algorithm is complicated, that it is more accurate than a simple one. But this belief is unfounded.

For example, Rudin and collaborators Elaine Angelino, Margo Seltzer, Nicholas Larus-Stone and Daniel Alabi have created a simple interpretable model for criminal recidivism based on only age, sex and prior record. Not only does it follow three simple rules that anyone can understand, it predicts the likelihood of future arrests just as well as the controversial "COMPAS" program, which is widely employed in the U.S. Justice System. And besides being a black box that many suspect employs racist biases, COMPAS uses more than 130 pieces of information to make its predictions, which is a major problem of its own accord.

"If the people entering this data make a typographical error just one percent of the time, then more than 1 out of every 2 surveys on average will have at least one mistake," says Rudin. "Plus an overly complicated black box model may be flawed without anyone realizing it, because it's difficult to troubleshoot."

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	only age, priors, (optional) gender no other information free, transparent

The Propublica example

ProPublica recently accused the COMPAS recidivism black box algorithm of being racially biased because they created an explainable model based on race that reproduces COMPAS's results. But because societal pressures have created a criminal justice system where criminal history and age are correlated with race in every dataset, the actual COMPAS black box might actually only be relying on the first two variables. But then again, it could also be using race as a factor just as ProPublica claims. The problem is that it's impossible to tell because COMPAS is an (expensive) proprietary black box that nobody buts its owners can peer into.

Rudin also points out several other contemporary problematic examples. The proprietary black box BreezoMeter told users in California their air quality was perfectly fine when the air quality was dangerously bad according to multiple other models. A machine learning model designed for reading x-rays was picking up on the word "portable" within an X-ray image, representing the type of X-ray equipment rather than the medical content of the image, and thus reporting bad conclusions.

"There is a conflict of responsibility in the use of black box models for high-stakes decisions. The companies that profit from these models are not necessarily responsible for the quality of individual predictions," says Rudin. "A prisoner serving an excessively long sentence due to a mistake entered in an overly-complicated risk score could suffer for years, whereas the company that constructed this model profits from its complexity and propriety. In that sense, the model's designers are not incentivized to be careful in its design, performance and ease of use. These are some of the same types of problems affecting the credit rating agencies who priced mortgages in 2008."

"I'm hoping that people realize the risks in explainable models and that they don't actually need black boxes at all. They can use models that are completely interpretable," says Rudin. "I would like to see a system in which no black box algorithm is used for a high-stakes decision unless there is no equally accurate interpretable model. I've worked on many different applications—medical care, energy, credit risk, criminal recidivism, computer vision—and I've never found an application where a black box is actually needed."

More information: Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, *Nature Machine Intelligence*, DOI: 10.1038/s42256-019-0048-x , www.nature.com/articles/s42256-019-0048-x

Provided by Duke University