Friday, December 22nd, 2017

In Order Not to Discriminate, We Might Have to Discriminate

by Christoph Drösser, Journalist in Residence

We all want to be unbiased. We don't want to hire people based on their gender; we don't want people to be incarcerated because of the color of their skin. Laws explicitly forbid discrimination on the basis of a set of "protected categories" like sex, race, religion, or age. The conventional wisdom is that fairness is best achieved when you try to be as blind as possible to these categories – for example, by having people apply for a job without stating their name and age or attaching a photograph. Then the employer would pick candidates solely based on their qualification for the job. But this assumption – the essence of many equality laws and court decisions – might be wrong. Mathematically provably wrong.

In recent years, more and more decisions about our lives are being made by algorithms. Algorithms are used to determine who gets a loan, who is considered for a job, who gets paroled from prison. Proponents argue that those algorithms are not biased like humans, but objective and fair. But research shows that bias can creep into those algorithms even if the programmers try to avoid it. And eliminating this bias is easier said than done – it might come with a price. This was demonstrated by computer scientists at a recent symposium at the Simons Institute for the Theory of Computing. The researchers agreed that the strategy of ignoring protected categories doesn't automatically lead to a fair decision. It might seem paradoxical – but sometimes you have to look at attributes like gender or race in order not to discriminate. As Moritz Hardt, a computer science professor at UC Berkeley, says, "There is no such thing as fairness through unawareness."

A lot of the discussion at the symposium centered on a controversial algorithm, developed by the company Northpointe and used in Broward County, Florida, to assess the recidivism risk (chance that a person will commit a new offense) of defendants awaiting trial. The algorithm produces risk scores used to decide whether a defendant is released on bail or not. The inner workings of the algorithm are proprietary, but the website ProPublica was able to compare the risk scores (on a scale from 1 to 10) of around 7,000 people, with whether the defendants committed a crime or a misdemeanor in the two years after the first arrest. With this comparison, it was possible to test the algorithm's accuracy in every single case and determine if it treated any protected groups in an unfair manner.

A first look at the data seemed to show that the algorithm's predictions were quite accurate: of the people with a low risk score (scoring 1 to 4 on the scale), about a third had a repeat offense. Among the people with high risk, the rate was two thirds.

But what happened when race was considered? Looking at black versus white defendants, the algorithm had classified 59% of blacks as high-risk, but only 35% of whites. Is that discrimination? Not necessarily. Blacks also had a higher overall recidivism rate. For a given risk score, the rates were very similar for blacks and whites. For example, in the high-risk group (with a score from 5 to 10), 37% of black defendants and 41% of white defendants did in fact not go on to commit another crime.

ProPublica's charge of racial bias hinged on a different statistical factor: the "false positives". These are given by the percentage of people who don't reoffend but were classified as high-risk. This share was 45% of black defendants, but only 23% of whites. So of the black defendants who did not go on to commit another crime, almost half received a harsher judgment than they deserved, while among the whites it was only a quarter. A clear case of racial discrimination?

Sam Corbett-Davies, a PhD student in computer science at Stanford University, disagrees. He was at first puzzled by the data from Broward County. Looking at it more carefully, he came to the conclusion that there can be an unresolvable conflict between individual fairness and group fairness.

Assuming that the individual risk scores are not biased, the only way to achieve perfect group fairness with respect to the false positive rate among people who do not reoffend would be to apply a higher risk threshold for the release of black defendants as opposed to white defendants. For example, keep whites with a risk score of at least 5 in jail, but blacks only with a score of 7 or more.

But that would lead to a whole wave of new instances of racial bias. Every white person with a score between 5 and 7 who is detained could argue that he was being held to a stricter standard than a black person with the same score – i.e. discrimination based on race. Additionally, raising the risk threshold for any group by two points would inevitably lead to more crime. In this case, crimes would rise by about 7 percent. And since most crimes are committed within the perpetrators' own communities, in this case it would lead to more crime in minority neighborhoods. People in the affected communities could be argued to be discriminated against by this policy.

Corbett-Davies's conclusion: the only way to achieve a fair algorithm is to try to avoid obvious errors that scientists call miscalibration, redlining, sample and label bias. Once those sources of bias are excluded, apply a common threshold to everybody – and live with the numerical oddities that might arise.

While the focus of the discussion in the Broward County case was on racial bias, Corbett-Davies discovered another striking disparity that had been overlooked: the algorithm clearly discriminated against women. If you compared the risk factor assigned by the computer to the actual recidivism rates, women were rated two classes worse than men. For example, female defendants with a risk factor of 6 recidivated about as often as men with a factor of 4. The algorithm, which doesn't take gender into account, is unfairly calibrated. "This seems like a strong evidence of bias against women in this risk-assessment tool," said Sam Corbett-Davies. To put it another way: there must be something that makes women less likely to commit another crime. And the algorithm misses it. We don't know the 137 pieces of information that go into computing the risk factor, since the software is proprietary. But it's quite possible that this cannot be fixed without introducing gender or a close proxy for it into the equation – and that is forbidden by law. An experienced judge might factor it into her decision and rule more mildly on women, without mentioning it explicitly. An algorithm doesn't have that kind of liberty.

This is not the first time that researchers have come to this conclusion. In an article that appeared in the *University of Pennsylvania Law Review* in 2016, Joshua Kroll and his co-authors state: "There may be cases where allowing an algorithm to consider protected class status can actually make outcomes fairer. This may require a doctrinal shift, as, in many cases, consideration of protected status in a decision is presumptively a legal harm."

Moritz Hardt from Berkeley agrees: "Every single technical study of fairness has come up with this insight that you can't just ignore the sensitive attributes," he said in his talk at the symposium. "But this is still predominantly what a corporate lawyer would want you to do. So there's certainly a tension there."

A tension – and maybe a slippery slope, in some situations. Consider the following example that Sam Corbett-Davies gave: among men, extended periods of joblessness might be a negative predictor for their performance in a future job. Women, on the other hand, might have taken a year or two of maternity leave, which is not related to their performance at all. So how should a hiring algorithm deal with this piece of information?

"Without knowing whether the applicant is a man or a woman, we can't learn to correct that bias," says Corbett-Davies, "but if we know, we can correct it." So the algorithm could in theory incorporate gender in its calculations, give men a negative score for a period without a job and ignore that feature if the applicant is female. But what about the man who took a year of paternity leave? Or the woman who didn't have children but was fired from a job? Isn't the better answer to fix the algorithm in a way that asks the reason for the unemployed period and adjusts its judgment accordingly?

Algorithms as they are applied today look for correlation, not for causation. As long as A can be used as a good predictor of B, they don't care if A causes B, or B causes A, or a third variable C causes both B and A. And an algorithm doesn't judge these correlations. Uncorrected, it might treat a variable that is a proxy for gender, like being a football enthusiast, as a good predictor for being the successful CEO of a company, which is still a predominantly male position. Thus, even when gender is not available as a predictor variable, an algorithm might discover a highly correlated proxy that serves the same purpose. In his talk, Hardt gave two scenarios of classifying people – in this case whether or not you present a job ad to them – in which the "observational properties," meaning all the objective factors that go into an algorithm's computations, were structured in the same way, but we would look at them very differently under a fairness perspective.

Hardt's conclusion is that fairness isn't something that we can easily build into a machine learning algorithm. "I don't think any of this is going to replace human scrutiny and expertise," he said. "The question is: how can we guide and support human experts in understanding these problems?" The old strategy of trying to be blind to protected categories doesn't work in the age of algorithms. We sometimes have to discriminate in order not to discriminate. And for now, there should always be a human in the loop when we make life-changing decisions about people.

Related Articles

- Letter from the Director, Fall 2017