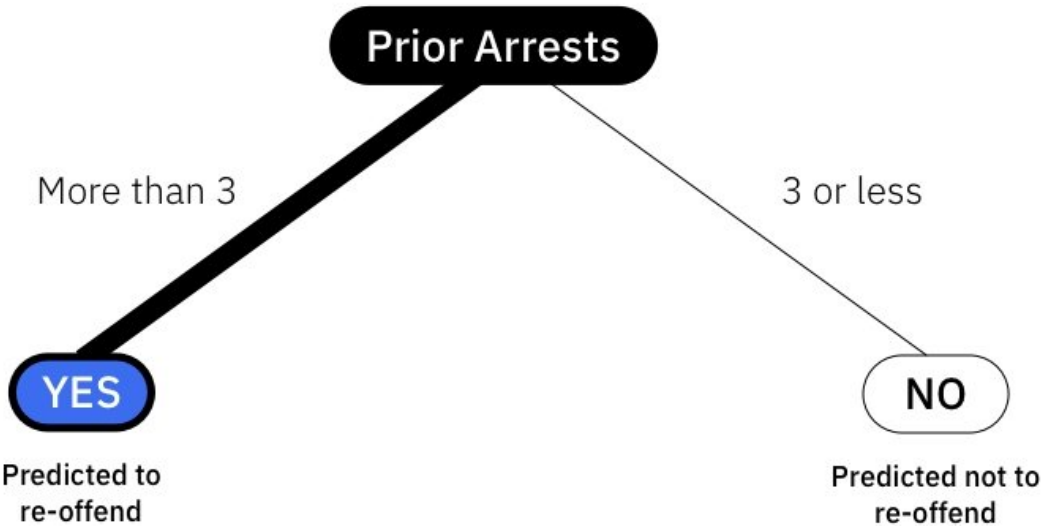
Reducing discrimination in AI with new methodology

December 7 2017



A decision tree from the pre-processed ProPublica COMPAS dataset. Credit: IBM

I finally had a chance to watch Hidden Figures on my long journey to Sydney, where I co-organized the second annual ICML Workshop on Human Interpretability (WHI). The film poignantly illustrates how discriminating by race and gender to limit access to employment and education is suboptimal for a society that wants to achieve greatness. Some of my work published earlier this year (co-authored with L. R.

Varshney) explains such discrimination by human decision makers as a consequence of bounded rationality and segregated environments; today, however, the bias, discrimination, and unfairness present in algorithmic decision making in the field of AI is arguably of even greater concern than discrimination by people.

AI algorithms are increasingly used to make consequential decisions in applications such as medicine, employment, criminal justice, and loan approval. The algorithms recapitulate biases contained in the data on which they are trained. Training datasets may contain historical traces of intentional systemic discrimination, biased decisions due to unjust differences in human capital among groups and unintentional discrimination, or they may be sampled from populations that do not represent everyone.

My group at IBM Research has developed a methodology to reduce the discrimination already present in a training dataset so that any AI algorithm that later learns from it will perpetuate as little inequity as possible. This work by two Science for Social Good postdocs, Flavio Calmon (now on the faculty at Harvard University) and Bhanu Vinzamuri, two research staff members, Dennis Wei and Karthikeyan Natesan Ramamurthy, and me will be presented at NIPS 2017 in the paper "Optimized Pre-Processing for Discrimination Prevention."

The starting point for our approach is a dataset about people in which one or more of the attributes, such as race or gender, have been identified as protected. We transform the probability distribution of the input dataset into an output probability distribution subject to three objectives and constraints:

1. Group discrimination control,
2. Individual distortion control, and
3. Utility preservation.

By group discrimination control, we mean that, on average, a person will have a similar chance at receiving a favorable decision irrespective of membership in the protected or unprotected group. By individual distortion control, we mean that every combination of features undergoes only a small change during the transformation to prevent, for example, people with similar attributes from being compared, causing their anticipated outcome to change. Finally, by utility preservation, we mean that the input probability distribution and output probability distribution are statistically similar so that the AI algorithm can still learn what it is supposed to learn.

Given our collective expertise in information theory, statistical signal processing, and statistical learning, we take a very general and flexible optimization approach for achieving these objectives and constraints. All three are mathematically encoded with the user's choice of distances or divergences between the appropriate probability distributions or samples. Our method is more general than previous work on pre-processing approaches for controlling discrimination, includes individual distortion control, and can deal with multiple protected attributes.

We applied our method to two datasets: the ProPublica COMPAS prison recidivism dataset (an example containing a large amount of racial discrimination whose response variable is criminal re-offense) and the UCI Adult dataset based on the United States Census (a common dataset used by machine learning practitioners for testing purposes whose response variable is income). With both datasets, we are able to largely reduce the group discrimination without major reduction in the accuracy of classifiers such as logistic regression and random forests trained on the transformed data.

On the ProPublica dataset with race and gender as protected attributes, the transformation tends to reduce the recidivism rate for young African-American males more than any other group. On the Adult dataset, the

transformation tends to increase the number of classifications as high income for two groups: well-educated older women and younger women with eight years of education.

Our work contributes to advancing the agenda of ethics and shared prosperity through AI. However, it has a couple of limitations I'd like to point out. First, there are many more dimensions to fairness than the strict sense of procedural equitability or non-discrimination in decision-making that is easy to express mathematically. This broader set includes distributive and restorative justice along with many other notions that we discussed in the Auditing Algorithms workshop I recently participated in. Second, data science and AI pipelines tend to be quite complicated, involving several different entities and processing steps during which it is easy to lose track of the semantics behind the data and forget that the data points represent actual people. These situations call for an end-to-end auditable system that automatically ensures fairness policies as we lay out in this vision (co-authored with S. Shaikh, H. Vishwakarma, S. Mehta, D. Wei, and K. N. Ramamurthy); the optimized pre-processing I've described here is only one component of the larger system.

Provided by IBM