UC Irvine Machine learning and philosophy CFA
15 January 2018

Name: Robert Long
Affiliation: NYU Philosophy
Contact: rl2898@nyu.edu
Title: Fairness in machine learning: beware false positive equality
Topics: ethics, machine learning, statistical discrimination, disparate impact

UC Irvine Machine learning and philosophy CFA
15 January 2018

Fairness in machine learning: beware false positive equality

Suppose that you have been arrested for shoplifting and are awaiting trial. You are given a questionnaire to fill out. Questions range from: how many of your friends have been arrested, to how many times you have moved in the past two years, to how you feel about stealing from rich people. In addition to this data, police pull information from your criminal record. Then, a machine learning algorithm takes this data, weighs it, [1] and generates a score that predicts your likelihood of committing another crime in the near future—your risk of "recidivism". The use of these scores, known as risk assessments, is on the rise, spurred by prison overcrowding and by the hope that machine learning can help make criminal justice decisions more swift, accurate, and impartial. What risk score you receive is very important: defendants who receive a high risk score[2] are far more likely to be detained while awaiting trial [1].

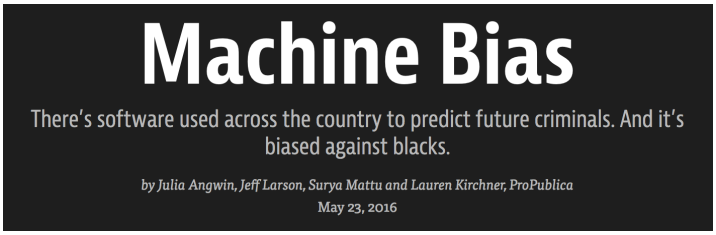
This algorithm, called COMPAS, is one of the most widely discussed cases in the burgeoning literature in "machine learning ethics" [2-7]. This is due in large part to an exposé of COMPAS by investigative journalism outfit Pro Publica. Pro Publica revealed that COMPAS produces higher rates of false positives for blacks than for whites—it manifests what I'll call "false positive inequality". In this context, a false positive is a high risk score for someone who does not recidivate. On the basis of this finding, Pro Publica charged that COMPAS is "biased against blacks". In the ensuing discussion, however, other commentators stressed an important sense in which COMPAS is intuitively racially fair or neutral: COMPAS is equally accurate for blacks and for whites when it flags defendants as risky. That is, its "positive predictive value" is the same for whites and blacks; it manifests what I'll call "predictive parity". However, because there are genuinely different base rates of rearrest among blacks and whites, an algorithm with predictive parity will in fact be statistically guaranteed to produce false positive inequality.

This is an instance of a more general tradeoff that I call the **central impossibility**:

> Given unequal base rates between two groups and less than perfect prediction, then a single decision rule applied to these two groups cannot have all of the following:
> Predictive parity: equal positive predictive value (PPV)
> Error rate parity: equal false positive rates (FPR) and equal false negative rates (FNR)[3]

[1] COMPAS was created by a private company called Northpointe, and its methodology is proprietary and thus unknown, which itself is problematic. However, observers who have replicated COMPAS's results speculate that COMPAS is either a regression model, or *should* be a regression model, since more complicated methods (neural nets, SVM) do not significantly boost accuracy. The ethics of model transparency is fascinating in its own right. See, for example, Jung et al. (2017) for an argument that decision trees should be used whenever possible due to their comprehensibility to practitioners.
[2] COMPAS gives people a score from 1-10. Many observers collapse these into "high risk"/"low risk" or "risky"/"not risky" for ease of exposition and analysis.
[3] Part of this paper explains, using the common statistical tool of "confusion tables", exactly why this is the case—having a good grip on this mathematical result is key to discussing its ethical significance.

In short: some defend COMPAS as fair because it has predictive parity; others attack it as unfair because of its false positive inequality. In essence, people are disagreeing about what makes an algorithm fair or unfair.



The headline of the Pro Publica article [1]

The state of the literature is this: many machine learning papers canvass technical tools for combatting false positive inequality, while not treating this ethical question of whether false positive equality is a "type" of fairness that should be optimized for. As the central impossibility implies, these technical tools ultimately involve either: sacrificing predictive parity for one group, or setting different detention thresholds for each group (to adjust the false negative rates). This paper gives an ethical framework for considering these sorts of changes.

I argue that false positive inequality (FPI) generated via predictive parity is not objectionable. However, I accommodate the intuition that it is by pointing out that there are many things that are often *associated* with false positive inequality that are objectionable. In fact, in the COMPAS case the FPI may very well be a symptom of real ethical problems: for example, that COMPAS is in fact miscalibrated because it makes use of arrest rates, which are a racially biased proxy for crime, or that the threshold for detention does not take into account the interests of blacks. But I argue that it is these symptoms that should be addressed, and not false positive inequality itself. In general, combatting false positive inequality is at best indirect and at worst counterproductive, not to mention in conflict with legal mandates of, and philosophical principles of, impartiality.

While the COMPAS case presents some ethical issues that are distinctive to criminal justice—punishment, rights, authority—grappling with the central impossibility will be important to many machine applications. In fact, a large prior literature about false positive inequality comes from employment law [11] and psychometrics [12]. This paper seeks to draws attention to these literatures, which seem to have been completely neglected as machine learning ethicists consider the central impossibility "from scratch".

The paper is also a test case for a question of interest for workshop participants: what philosophical issues are distinctive to machine learning per se, and what are instances of more general issues about (e.g.) statistical prediction [8-10]? As it happens, many issues in my paper are more general issues, albeit issues that will be especially relevant to machine learning practitioners as their role in constructing decision-rules continues to grow.

Selected works cited

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. ProPublica, 2016
[2] Chouldechova, A. (2016) "Fair Prediction With Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." arXiv:1610.075254v1 [stat.AP]
[3] Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2017). Fairness in Criminal Justice Risk Assessments: The State of the Art. arXiv preprint arXiv:1703.09207.
[4] Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2016). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. arXiv preprint arXiv:1610.08452.
[5] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.
[6] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. arXiv preprint arXiv:1701.08230.
[7] Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S. (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear. Washington Post.
[8] Lippert-Rasmussen, K. (2011). "We are all Different": Statistical Discrimination and the Right to be Treated as an Individual. *The Journal of ethics*, *15*(1-2), 47-59.
[9] Eidelson, B. (2015). Discrimination and Disrespect. Oxford University Press.
[10] Schauer, F. F. (2009). Profiles, probabilities, and stereotypes. Harvard University Press.
[11] Wax, A. L. (2011). Disparate Impact Realism. Wm. & Mary L. Rev., 53, 621.
[12] Jensen, A. R. (1980). Bias in mental testing. New York. Chicago