Applied Machine Learning - WA 2

Albin Jansfelt, Amanda Allgurén, Lukas Martinsson

March 2022

The following essay will discuss bias and fairness aspects of machine learning and will be based on the two articles *"A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear"* (Corbett-Davies et al., 2016) and *"A Popular Algorithm Is No Better at Predicting Crimes Than Random People"* (Yong, 2018). Both articles discuss issues of the tool COMPAS which was used for risk-assessment of recidivism and the claims of the algorithm being biased against African Americans.

Corbett-Davies et al. (2016) presents the argument if the algorithm is biased or not from both ProPublica's stance and Northpointe's stance. Depending on how the measure is made, the different actors can be seen as correct. ProPublica noted that individuals with high and medium risk to reoffend, which did not reoffend were 42% for blacks and 22% for whites. Hence being biased. However, looking at blacks and whites separately, each group have the same share of reoffendants in each risk category. If this fact would be any different, it would be considered biased towards one of the groups. Corbett-Davies et al. (2016) means that it is not possible to satisfy both these criterion at the same time.

One additional observation added by Yong (2018) is the overall perception of the algorithm being better and more trusted than human risk-assessments. This, even if the accuracy of a group of 400 random volunteers predicting the recidivism is found higher than COMPAS.

Even if the tool can be found biased, none of the articles suggest how to eliminate these kinds of algorithms in the justice system. Although, Corbett-Davies et al. (2016) suggest replacing bail requirements with electronic monitoring. Instead, they suggest that it is important to be aware of and debate issues like these - because humans do also prove a risk of being biased.

Human biases may be harder to detect and get the same statistics on, especially in the case of justice system. If the human judge find a defendant guilty, then the person will be labelled as such. There is no simple way for checking human biases, as it is for an algorithm to feed it lots of data with the correct labelling. In this case, COMPAS is found baised against blacks if any, but considering that society may also be biased against blacks, the "correct" label of the defendants may be influenced by the judge putting that label on the defendant.

How do we define bias and fairness?

Instead of discussing whether COMPAS is considered fair and unbiased or not, lets generalize the issue. In these cases, what would be considered unbiased and fair, and furthermore, should it even be coveted? By answering this question, a fairer assessment models such as COMPAS could be made, since some sort of framework for deciding fairness and unbiasedness must be established prior to the evaluation.

The discussion regarding bias is probably the easiest discussion to start with. The definition itself is rather specific which is some sort of prejudice for an individual that is attributed to his/her belonging to a certain race, gender, religion et cetera. Which group that belongs to these "protected groups" may be open for debate. However, assume that there exist a well-established protected group, and a model basing its decision on these attributes would be considered biased. In the case of COMPAS this group would be black defendants.

It is significantly harder to determine what constitutes fairness. To generalize as much as possible it would come down to whether one regards fairness to be adjusted for one's potential or not. Most would probably agree that equality of opportunity, that is, everyone should have the same opportunity to achieve or do something would be considered fair. The easiest way would be to assume that everyone has the same potential which would make equality of opportunity on its own be the definition of fairness. Then everyone gets judge on a similar base and hence by definition of equality of opportunity the individual that worked the hardest for an opportunity would most likely get it. However, as all individuals are different this is not the case. For example, individuals with a low socio-economic background with a mild cognitive impairment is at a big disadvantage compared to the rest of the population when it comes to being admitted to a higher education. Since equality of opportunity adjusted for one's potential is extremely complex, not only for the scope of this paper, but in general since there are next to an infinite many factors influencing it this paper assumes fairness through equality of opportunity not adjusted for one's potential.

Do we have to consider bias and fairness to achieve a good model?

As established above, biased can be defined as prejudices against a protected group, which in the COMPAS case is argued to be black defendants. Further, fairness is easiest to define as equality of opportunity. If defined in such a way, the biasness accusation described in the articles can run counter to the equality of opportunity. This is also addressed in Corbett-Davies (2016), where the solution to mitigate the biasness would be to systematically assign white defendants a higher risk score than equally risky black defendants.

However, one question that arises is whether defining bias and fairness is relevant at all. Another example in a similar fashion to the discussion of COMPAS is that more men commit violent crimes than women. A trait correlated with this is aggression, and by looking at a bell curve for aggression comparing men and women one can see that the mean value of the bell curve is slightly

higher for men. This slight difference results in a significant difference in the extremes, i.e. most men and women are about equally aggressive, however there are significantly more extremely violent men than women.(Björkqvist, 2018)(Im et al., 2018) By sheer statistics it would be more efficient to assume that a violent person is a man. There might be a reduction of crime if this assumption could be incorporated into the justice system, even if it would go against the definitions of bias and fairness above. Why not utilize it?

On the one hand, an issue that could arise from this is a negative assumption spiral. If men are treated as more aggressive, even when most men are not, it could lead to repression and through it a realization of what the incorporation of this assumption were supposed to dimmish. On the other hand, since this spiral could only occur if it is noticeable by the general population a solution could be not to present it publicly. This could also bring forth negative consequences which have to be weighed against the positive result of using the assumption (which in this case is not an assumption but a fact).

The way forward

The task of considering bias and fairness in the development of machine learning models such as COMPAS is not as easy as in first glance. As shown in the example of COMPAS, more black individuals were judged to be high risk. This could be justified as discussed above if the false positives (high-risk that would not reoffend) number is similar for black individuals and white individuals, or if the model is better than the alternative, i.e. humans. However, in the case of COMPAS the only similar ratio for these different ethnicities was the ratio between the false positive and true positive and a human comparison was not made.

Utilizing statistics as described previously can result in lowering crime-rates. Yet, it go against the definition of bias fairness stated earlier. Thus, the need for considering bias and fairness needs to be decided on a case-to-case basis. The issue is important and should not be neglected, although the solutions are hard to derive. Both articles understands this, since they are lacking solutions.

References

Björkqvist, K. (2018). *Gender differences in aggression.* Current Opinion in Psychology. Retrieved from `https://doi.org/10.1016/j.copsyc.2017.03.030`

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.* The Washington Post. Retrieved from `https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/`

Im, S., Jin, G., Yeom, J., Jekal, J., Lee, S., Cho, J., ... Lee, C. (2018). *Gender differences in aggression-related responses on eeg and ecg.* 10.5607/en.2018.27.6.526. Retrieved from `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6318556/`

Yong, E. (2018). *A popular algorithm is no better at predicting crimes than random people.* The Atlantic. Retrieved from `https://www.theatlantic.com/technology/archive/2018/01/equivant-compas-algorithm/550646/`