HW 5

Razmin Bari

11/08/2024

This homework is meant to give you practice in creating and defending a position with both statistical and philosophical evidence. We have now extensively talked about the COMPAS [1] *data set, the flaws in applying it but also its potential upside if its shortcomings can be overlooked. We have also spent time in class verbally assessing positions both for and against applying this data set in real life. In no more than two pages* [2] *take the persona of a statistical consultant advising a judge as to whether they should include the results of the COMPAS algorithm in their decision making process for granting parole. First clearly articulate your position (whether the algorithm should be used or not) and then defend said position using both statistical and philosophical evidence. Your paper will be grade both on the merits of its persuasive appeal but also the applicability of the statistical and philosophical evidence cited.*

The COMPAS algorithm should not be used by a judge in their decision making process for granting parole. The main point of using COMPAS would be to curb the inherent bias of the current human processes, as in overcoming racial or other biases. This would require COMPAS itself to be completely unbiased. As evident by calculations with the COMPAS confusion matrices [3], the algorithm exhibits disparate impact. Even though race is a agreed-upon protected class, COMPAS predicts that defendants of one racial group are more likely

[1] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis
[2] knit to a pdf to ensure page count
[3] presented in class

to re-offend if they are granted parole. Hence COMPAS violates the fairness criteria of independence.

It may be argued that the algorithm need not predict equal proportions of re-offenders across groups as there may be group-wise differences in actuality. However, even when accounting for the ground truth, COMPAS does not show equalized odds across the racial groups. The algorithm *wrongly* predicts black defendants to be re-offenders at a higher rate, violating the separation criteria for fairness. This would play out in real life with many people who were already rehabilitated to be denied parole when they already deserve it.

It is worth noting that COMPAS fulfills the sufficiency fairness criteria as the algorithm groups similar individuals together and decides which attributes are good predictors of the outcome of re-offending if granted parole. With the zip-code variable acting as a proxy for race however, race does not remain a protected variable within the algorithm. Historical racism has thus been programmed into COMPAS. An individual should only be judged based on their own actions, not using data from previous unrelated defendants with similar traits. The way COMPAS works violates the first categorical imperative in deontology which states that an act is permissible only if it can be universalized without logical contradiction. An individual being judged not for their own actions but by virtue of what group he belongs to inherently is discrimination. It also violates the idea of "innocent until proven guilty" since a person may be assumed guilty right from the start due to, for example, their race.

COMPAS could potentially be used if it was retrained on data with larger sample sizes across racial groups so that equalized odds could be reached. At the current stage, it only serves to perpetrate systemic racism. Assuming that COMPAS has only been proposed for use as a guiding tool, it may end up simply being a source of confirmation bias for judges that were biased to begin with. If the COMPAS output is different from the judge's decision, it *may* counter human biases but COMPAS' accuracy is only 63.6%. If a judge is aware of COMPAS' limitations, they may be less inclined to trust it over their own thoughts and

intuitions. The judge would not be able to explain his reasoning for the verdict anyway. COMPAS is a black-box algorithm and it cannot be reasonably expected that the judge would be able to explain the workings of such an algorithm. This is a problem since judges are expected to provide reasoning for their verdicts such that it is explainable to the public.