HW 4

María Rubio Navarro

10/10/2024

This homework is designed to give you practice working with statistical/philosophical measures of fairness.

The paper linked below[1] discusses potential algorithmic bias in the context of credit. In particular, banks are now regularly using machine learning algorithms to do an initial screening for credit-worthy loan applicants. In section 4.5.2, this paper reports the rates at which various racial groups were granted a mortgage. If we assume that it is a classifier making these predictions[2] what additional information would be necessary to assess this classifier according to equalized odds?

Perhaps an additional piece of information needed would be the error rate for each racial group. Knowing how often the classifier incorrectly approves someone who is ineligible and how often it incorrectly rejects someone who is eligible in each group would allow us to compare the classifier's fairness and see if it treats all groups equally.

Show or argue that the impossibility result discussed in class does not hold when our two fringe cases[3] are met.

On the one hand, if the classifier predicts flawlessly, there are no false positives or false negatives, which prevents conflicts between fairness metrics: all groups present the same perfect results. On the other hand, if the ratio of true results is constant across all groups, the fairness metrics also align more easily, since all groups start with the same baseline of results.

How would Rawls's Veil of Ignorance define a protected class? Further, imagine that we preprocessed data by removing this protected variable from consideration before training out algorithm. How could this variable make its way into our interpretation of results nonetheless?

According to Rawls's Veil of Ignorance, any group that may experience inconvenience without individuals knowing which group they belong to is considered a protected class, thus promoting equitable decisions for all. Even though we discard this protected variable when training the algorithm, its impact can be maintained indirectly through correlated variables, making it easier for subtle biases to influence the results.

[1] https://link.springer.com/article/10.1007/s00146-023-01676-3

[2] It is unclear whether this is an algorithm producing these predictions or human

[3] a) perfect predicting classifier and b) perfectly equal proportions of ground truth class labels across the protected variable

Based on all arguments discussed in class, is the use of COMPAS to supplement a judge's discretion justifiable. Defend your position. This defense should appeal to statistical and philosophical measures of fairness as well as one of our original moral frameworks from the beginning of the course. Your response should be no more than a paragraph in length.

Based on what we have seen in class, it seems to me that using COMPAS to supplement a judge's discretion is not justifiable. Statistically, COMPAS shows a 64% accuracy rate but violates fairness principles like statistical parity (independence) and equalized odds (separation) by producing racially biased outcomes through proxies like zip code. Philosophically, it should be highlighted Rawls' perspective which said that this lack of fairness contradicts a just system where one's treatment should not depend on immutable characteristics. Additionally, as a black-box model, COMPAS lacks transparency, preventing accountability. Thus, relying on COMPAS compromises fairness and justice, undermining the ethical standards expected in judicial decision-making to me.