HW 4

Student Name

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?

We cannot make an assumption that the bank's AI assistant is unfair just by looking at the high-rates that a certain racial group was denied. The most important statistic that we need to make an assessment on the fairness of these algorithms is the percentage of people that were identified as credit-worthy, but were denied due to their race. Therefore, to properly evaluate the fairness of the algorithm, we need to compare the approval rates among different racial groups for applicants with similar creditworthiness.

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

The impossibility result does not hold in the two fringe cases. First, if the classifier is perfectly accurate, it makes no errors, so fairness measures like equalized odds or demographic parity are fully satisfied. Second, if there are perfectly equal base rates across protected groups, the classifier can achieve equal true positive and false positive rates among all groups without sacrificing accuracy or fairness. Thus, in both cases, the conditions that create conflicts between fairness criteria are absent, so the impossibility result does not apply.

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, a protected class is any group that could end up disadvantaged when society's rules are set without knowing one's own position. Even if we remove the protected class from our data before training the algorithm, it can still influence results because other variables might be closely related to it. These connected features can act as proxies, so the protected characteristic can unintentionally affect our interpretations.

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.

[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

Artificial Intelligence is advancing every day, and we are finally starting to see its influence in our lives by efficiently speeding up tasks, from recommending your next meal to helping diagnose medical conditions. Even though the decision-making seems to be rational, there are ethical disadvantages that come with integrating AI into every step of life. The higher the stakes of a certain decision, the more the users have to worry about the algorithm's fairness. Therefore, the use of COMPAS to supplement a judge's discretion is unjustifiable because of the its long history of yielding biased predictions. Such decisions, which can significat impact on people's lives, should not be left to an imperfect algorithm that has been shown to disproportionately label minority groups as higher-risk. Even though the developers have closely trained COMPAS with all the data throughout the years, we still cannot trust on AI to account for individual circumstances of a certain case. An important concept that we learned earlier in the course, deontology, is closely related with the argument against the use of COMPAS, as we aim to focus on the morality of our actions regardless of the outcome. To conclude, only the judge is capable of making a fair assessment on a case, and relying on AI in this particular field can further advance discrepancies in our justice system.