HW 4

Jonathan Zhao

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?

Student Input.

For equalized odds, there needs to be a roughly equivalent false positive rate between protected and nonprotected classes. The additional information that would assess this classifier would be the percentage of each racial group that were approved for a loan but were not actually credit-worthy applicants, which is the false positive rate. This percentage should be roughly the same according to equalized odds, so if it isn't, there is a chance of algorithmic bias based on Separation.

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

Student Input For a perfect predicting classifier, it classifies all positive and negative cases correctly, which means the chance of a true positive and a false positive will be consistent throughout all groups, which satisfies Separation. Because it is 100% accurate, it also satisfies independence because the proportion of positive predictions match the base positives. Finally, it satisfies sufficiency because it is completely accurate regardless of the actual group of the data.

For perfectly equal proportions of ground truth labels, the base positive rate is the same across all groups, so independence can be satisfied because there will more likely be an equal proportion of predicted positives. This also allows a higher chance for the true positive rate and false positive rates to be equal since the amount of truth labels are equal for all groups, achieving Separation. Finally, it can satisfy sufficiency because having the same proportion of truth labels makes it easier for an algorithm to have the same positive prediction rate regardless of group.

[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

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?

Student Input

Rawl's Veil of Ignorance defines a protected class as a hiding demographic and socioeconomic information about groups to ensure that decisions are made without bias. This means hiding an individual's identifiers like race, sex, age so that the algorithm would not exhibit biases based on those factors. These variables could still make its way into our interpretation of results from proxy variables. For example, zip code can be a proxy for race, which can be seen in the COMPAS algorithm where it exhibits algorithmic bias against Black defendants without actually being fed data on race.

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.

Student Input

The use of COMPAS to supplement a judge's discretion is not justifiable. From a statistical standpoint, COMPAS violates the Independence fairness criteria. COMPAS predicts that Black defendants are at a higher risk of recidivism than white defendants, which is not fair based on Disparate impact. Black defendants were also more likely to be misclassified than white defendants, so it also violates the Separation fairness criteria because of equalized odds. From a philosophical standpoint, COMPAS isn't fair because it doesn't look at merit, but purely statistics, and thus isn't fair based on Robert Nozick's idea of merit. A person could have a lot of merit and be a very good candidate, but they could be biased against if they were of a certain demographic by COMPAS. COMPAS also violates the moral framework of Deontology, as it views the data collected from individuals as merely statistics to reach an end rather than the ends.