HW 4

Quyen Dang

10/29/2024

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?

Specific performance data on true and false favorable rates for each racial group is needed to assess the classifier's predictions according to equalized odds. This is crucial to evaluate in order to determine whether the classifier meets the equalized odds criterion, which requires rates to be similar across all groups. Additionally, the classifier needs to confirm whether attribution like race is being explicitly accounted for or systematically. Socioeconomic backgrounds and any features that might correlate with race are necessary to control for potential cofounders.

Show or argue that the impossibility result discussed in class does not hold when our two fringe cases [a) perfect predicting classifier and b) perfectly equal proportions of ground truth class labels across the protected variable] are met.

The impossibility result does not hold when the two fringe cases are met. If a classifier has perfect predictive power, each group's actual positive rate and the false positive rate will align with the actual distribution. This concept satisfies equalized odds, as the actual positive and false favorable rates would be zero for negatives and one hundred percent for positives. Additionally, when the underlying class proportions are the same across groups, a classifier can result in equal acceptance rates across groups without compromising on equalized odds.

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?

Under Rawl's concept, a protected class would be defined as any group with characteristics that could lead to a disadvantage. The Viel of Ignorance claims that decisions should be made as if the decision-maker is unaware of their position. Even if a variable like race is removed from preprocess data, it can still influence the model indirectly. "Proxy variables" can still correlate with protected variables such as zip code. These

[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

proxies can allow the model to reconstruct some of the original bias. Despite removing protected variables, the algorithm's outcomes can still result in societal inequalities.

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.

COMPAS can be beneficial in terms of its consistency in risk assessment, but it goes against fairness and ethical issues. COMPAS has been known for producing false favorable rates, especially for Black defendants. This disparity undermines Rawl's principle of justice because it fails to comply with the equality of opportunity for fair treatment. From a utilitarian perspective, COMPAS can increase efficiency, but it also comes with costs, such as potentially harming disadvantaged groups. For these concerns, using COMPAS without adjustments for bias is challenging to justify ethnically.