HW 4

Sofia Zhang

10/29/2024

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

1

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?

Equalized odds is a fairness criterion to evaluate whether a model treats different groups fairly， such as gender and race. A model satisfies equalized odds if its true positive rate (TPR) and false positive rate (FPR) are the same across all groups for a given decision threshold. In section 4.5.2 of this paper, the paper mentioned the difference of mortgage approval among different racial groups. To assess equalized odds, the TPR and FPR should be calculated separately for each racial group under consideration. Information on the decision threshold for mortgage approval is essential. Also, it would be helpful to know if different thresholds are being applied across racial groups, as this would directly affect TPR and FPR.

2

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

The impossibility result applies when there are inherent differences in the ground truth distributions or prediction errors across groups. When a perfect prediction classifier is met, all predictions align with the ground truth which means that it has 100% accuracy. Since the model always predicts the correct class, it satisfies equalized odds because TPR and FPR across groups are equal (both are either 1 or 0 for each group). Thus, it contradicts the impossible results. Case b applies when each protected group has the same proportion of positive and negative labels in the ground truth. In this situation, TPR and FPR can be balanced across groups, allowing the classifier to achieve equalized odds. Therefore, there is no source of conflict, as the classifier doesn't face differing probabilities. Therefore, impossibility results does not hold in this case.

3

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 John Rawls's concept of the Veil of Ignorance, the protected class is the group of people unaware of their own disadvantaged characteristics. However, even though a protected variable is removed, correlated variables can still indirectly present the influence of this protected variable, as known as proxy variables, such as ZIP code,

education level, or income. For instance, geographic location can often correlate strongly with race due to historical and socioeconomic factors. These proxy variables can inadvertently introduce the influence of the protected variable into the model and affect the results.

4

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.

The use of COMPAS to support judicial discretion is not acceptable. Statistically, COMPAS demonstrates issues with predictive parity, as its false positive rates vary across racial groups, leading to disproportionately high risk assessments for certain minorities, which challenges the system's commitment to equal protection. From a utilitarian standpoint, the correct action should maximize pleasure and minimize pain. Even if COMPAS could improve efficiency, its inherent biases likely produce more pain than benefit by reinforcing racial disparities, ultimately decreasing public trust in the justice system. This is the opposite of utilitarian argument.

---

1. https://link.springer.com/article/10.1007/s00146-023-01676-3 (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↩