HW 4

Seth Hall

10/22/2024

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?

More information is necessary to assess this classifier according to equalized odds, such as the positive and negative rates of the classifier. We would need the true positive rate (proportion of applicants from each group who were correctly classified as credit-worthy), false positive rate (proportion of applicants from each group who were incorrectly classified as credit-worthy), true negative rate (proportion of applicants who were correctly classified as not credit-worthy), and the false negative rate (proportion of applicants who were incorrectly classified as not credit-worthy). The true positive rate and true negative rate must both be 1, and the false positive rate and the false negative rate must be 0. If any of the categories are different, it would suggest that the classifier treats racial groups differently, leading to potential discriminatory outcomes. We would also have to collect some additional data for each group. This includes the total number of applicants so that we can ensure sufficient sample sizes for statistical comparison. We should also consider the predictions from the classifier and the ground truth (whether the subjects actually are credit-worthy) and compare them to each other.

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

According to the impossibility result, it's impossible for a prediction algorithm to meet all measures of fairness at the same time. However, this rule doesn't apply in two specific fringe cases. The first case is when the classifier is perfect, meaning its predictions always match the ground truth. This entails that it always meets the equalized odds criteria, so the true positive and true negative rates are the same for all groups. It satisfies statistical parity because the predictions match the actual outcomes exactly for each group, so there's no imbalance. Finally, it does not show disparate impact because one protected class is not more negatively affected than another when the algorithm is 100% correct. The second fringe case is when there are perfectly equal proportions of ground truth class labels across the protected variable. In this case, the algorithm would

[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

satisfy equalized odds and statistical parity since it will predict positive outcomes equally across all groups. Since the groups are balanced, there's no conflict between the measures, so disparate impact is also satisfied. So, in these two situations, either a perfect classifier or equal base rates allow both fairness criteria to be achieved without any issues. This shows that the impossibility result doesn't hold in these two fringe cases.

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?

Rawls's Veil of Ignorance is a way to imagine making fair rules by pretending you don't know anything about yourself such as your race, gender, or social status. The idea is that if you don't know these things, you'll make decisions that are fair to everyone. This is because you wouldn't want to risk creating rules that could disadvantage you. Rawl's Veil of Ignorance would define a protected class as something that people would want to anonymize to prevent unfair advantages or disadvantages. These things would include race, gender, or socioeconomic status. By ignoring these characteristics, people would create fair rules that don't give unfair advantages or disadvantages to anyone based on these traits. Even if we remove a protected trait like race from a dataset, it can still affect the algorithm's results through other related data. For example, we discussed in class how information like ZIP code or income might indirectly reveal a person's race. Sometimes, a combination of other features can hint at the protected trait. So, even if we don't include race directly, we might still see unfair patterns in results across different groups. This shows why we need to be careful of hidden biases in data.

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.

Using COMPAS to help judges make decisions is hard to justify because it raises concerns about fairness and bias. COMPAS has been shown to make more misclassifications for some races over others, thus it does ot uphold equalized odds. Not only does COMPAS violate equalized odds, but it also shows clear signs of disparate impact in that one protected group is more negatively affected than another, even though the rules appear to be fair. Beyond statistical measures of fairness, several moral frameworks we have discussed in class would also disagree with COMPAS. From a deontological perspective, which focuses on following fair rules, it's a problem if COMPAS makes biased predictions that affect certain groups unfairly. Even from a utilitarian view, which aims to do the most good for society, COMPAS's biases could make the justice system less fair overall. Without stronger proof that COMPAS is accurate and unbiased, it's tough to say that it should be used in court.