# HW 4

Lynn Check

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?

Equalized odds run on the assumption that the base rates of two groups are representative and were not obtained in a biased manner, until the actual ground truth. Assuming that it is a classifier that is making the predictions, according to equalized odds, additional information about the actual creditworthiness of each applicant is required. Specifically, creditworthiness should be accessed on objective financial indicators. The paper does not provide data in regards to the criteria and whether it was met to receive mortgage approval. Information such as this is important under equalized odds as one of the criteria is based on the actual ground truth. The lack of a true outcome label violates that criterion. Without the information, it would be difficult to determine whether the cause of the legitimacy of the lower approval rates is from the difference in applicant qualifications or due to algorithmic bias. More specifically, in section 4.5.2, it describes the disparity in approval rate across the racial groups, providing only the specific rates for Whites, (71%) Asians (68%), Latinos (62%), and Black (54%). In other words, this data only shows the demographic disparity. However, with this data, it does not give the true positive rates and false positive rates that are necessary to calculating the equalized odds accurately. Furthermore, the idea of "positive" and "negative" causes are unclear. It is necessary to know if positive cases are truly associated with approval for those who are creditworth and negative cases are truly associated with applicants that are involved with denials. Without the clear definition and differentiation between the two, the data is linked to the ground truth label. It would cause the TPR and FPR to be unattainable.

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 under these two fringe cases due to the very definition of them. The perfect predicting classifier implies that the TPR is 100% and the FPR is 0 when it comes to every single group. This implies that there are equalized odds, coming from the fact that the TPR and FPR hold to

[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

be the same across all groups. Furthermore, if the case of perfectly equal proportions of ground truth class labels across the protected variable is met, then each group under the variable will have the same rate in regards to what result is being studied. This means, there won't be any disparities or conflict that could arise among differences with these groups. The impossibliity result arises due to conflicts among different measures of fairness. However, if both these fringe cases are met, they remove the conflicting nature among different measures of fairness, causing different metrics to be met without issue.

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 defined his Veil of Ignorance as a way of thinking. It involves one to sit behind this metaphorical Veil of Ignorance, which would keep them from really knowing their own personal characteristics, allowing them to make unbiased decisions/assumptions about society. Under this Veil of Ignorance, a protected class would be defined as a part of society that faces any sort of discrimination (which could likely be systemic), and would need to be protected in a way that would provide fairness amongst themselves. If we decided to remove this protected variable from consideration before training our algorithm, this does not mean the variable won't be a part of our interpretation. One of the main reasons is that variables correlated with the protected class could exist when analyzed in our algorithm. We could have an independent variable that, after running our algorithm, becomes one of the main points of focus in regards to its correlation with our outcome variable. When analyzing this variable, we might tend to see its correlation with pieces of our protected class, and could find an explanation of certain trends we see explained by that protected class. Furthermore, the algorithm itself could look at patterns amongst variables and external factors before our analysis and see a correlation with the protected class.

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 supplement a judge's discretion is not justifiable, due to both the bias it places along race, as well as the general inaccuracy of the tool. It has been highlighted in multiple cases that COMPAS is biased against racial minority groups, and often produces a lot more false positives for black individuals than false negatives for white individuals (indicating the algorithm's inaccuracies, such as violating indepdence, are shown through the unfairness of its outcomes). The purpose of a classification algorithm like COMPAS is to discern amongst different groups, but act utilitarianism would not justify this. In this context, this type of utilitarianism is dependent on the the pleasure or pain that comes from COMPAS's decision in that moment. However, with the amount of false positives that arise, it becomes hard to state that each outcome can be deemed as good, especially when it comes to racial minority groups that are subject to the biases of this algorithm. Due to this, as well as there being no transparency on how the algorithm works, it is not justifiable to use this in a judge's discretion.