HW 4

Nikita McClure

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?

In the case discussed in section 4.5.2, to assess this classifier according to equalized odds, means we wish to ensure that the outcomes are independent of race. This means we are checking that the predictions do not vary by race. To do this, we would want the true credit-worthiness of each applicant, as well as the prediction for each. The true credit-worthiness to be compared would need to be based on other factors such as past debts and how/if they were paid off, if they have steady income, etc. With this information we could make a confusion matrix that shows true positives, true negatives, false positives, and false negatives for each race. If the distribution of the rates are similar among each racial group, it can be reasonably assumed that the classifier was independent of race, and that it followed equalized odds. Additionally, we would check that the types or amounts of loans are even across racial groups in accordance to their true credit-worthiness. This means that for each group, controlling for credit-worthiness we would make a confusion matrix of what loan they were predicted to get and what loan they got, to see if there is disparity in the rates or distribution of true or false negatives for different racial groups. This should probably be done for each bank being sued individually, since location/ regional income can be a factor in deciding trustworthiness.

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

A) A perfect classifier would mean all results are true so true positive and true negative are maximized and false positive and false negative are zero. This means the proportions of false are consistent between each group (zero all around), but true positive and negative proportions could vary between different groups. However, since it's completely accurate, meaning there is no misclassification for any reason, it can be seen to have achieved fairness.

B) In class, the impossibility result is said to state that it is impossible to be fair according to all three fairness types (predictive equality, predictive parity, and false negative rate) unless "the fraction of class 1 labels is

[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

consistent across protected and unprotected groups". This means that the classifier can only be completely fair, or fulfill all 3 fairness types, if the class proportions of positive and negative are the same across each group. This is the case that is described by "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?

Rawl's Veil of Ignorance would define a protected class as a group who has systematically worse outcomes or lower success, despite having similar ability and ambition. This means protected classes would be those that work as hard but have worse results due to factors out of their control such as race and sex. While these factors can be removed before training the algorithm, they can sneak their way back in through proxy data. Proxy data means variables which are not the protected class but that closely correlate to it. This is the case in algorithms like COMPAS that exclude race, but keep factors such as geographic location and social ties, which can strongly correlate to race. While race is not directly accounted for, the proxy data is closely enough related that it is as if race were accounted for, so the results are biased against certain racial groups even though race is not a factor in the algorithm.

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 decision is not justifiable. While statistically the algorithm is considered accurate for binary classification, it is not highly accurate, certainly not to the degree that it should help decide the fate of someone's life. Additionally, COMPAS is biased against black individuals, as it is more likely to predict them to reoffend than white individuals, meaning it does not fit both the statistical parity and equalized odds fairness criteria. While the argument can be made that this only supplements a judge's decision and does not make it for him, I would rebut that COMPAS would sway his decision, potentially in the wrong direction, so it is still a significant factor. Additionally, an algorithm cannot reflect on its decisions, nor can the decisions be questioned or defended, especially as it is a black-box situation. This makes it less fair to the defendant as they cannot question the reasons behind their sentencing. Lastly, there are few justifications for using an algorithm such as COMPAS when there are more reliable, less biased, options available such as LSI-R available. One philosophical framework that supports this opinion is consequentialism which values the consequences or outcomes of an action much more than the causes or intentions behind them. Consequentialism would argue that the consequences of unjustly punishing individuals outweighs potential benefits of supplementing a judge's discretion.