HW 4

Tara Hinton

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?

In the case of classifiers, equalized odds means that predictions are conditionally independent of a protected/sensitive feature given its label. So, to satisfy equalized odds, we need both model predictions that are independent of sensitive group/protected class and classes that have equal false positive rates and true positive rates. Faber (2018) found different mortgage approval rates for members of different demographic groups (71% of whites, 68% of Asians, 63% of Latinos, and 54% of blacks—demographic disparity). In this case, we have the difference between groups, but we do not know the false positive and true positive rates for each group. We would need this information in some form (a confusion matrix would be best) in order to assess equalized odds.

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

Our impossibility theorem states that it is impossible for one classifier to be fair by independence, separation, and sufficiency criteria – except when there is a perfect predicting classifier or the proportion of ground truth class labels is consistent across protected and unprotected groups.

a. Fringe case #1: When we have a perfect predicting classifier, independence can be met as statistical parity will be satisfied (no disparate impact). We will not see differing ratios of false positives between groups because there will be no false positives in a perfect classifier! Separation will also be met because, without false positives, we will have equal true positive rates (100% correct) and equal false positive rates (0% error). For sufficiency, our choices will reflect the same accuracy per sensitive group because there is 100% accuracy.

[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

b. Fringe case #2: Proportion of ground truth class labels consistent: Broadly, equal proportions of class labels across protected and unprotected groups means that any machine learning approach on a training set will not be subject to concerns about over or under-represention of class labels between groups. Independence will be met, because perfectly equal proportions of class labels across protected and unprotected groups means that our false positives will be equal. Our true positive rates and false positive rates will virtually not differ between protected/unprotected groups (because we have equal representation in the training set.). For sufficiency, our groups will have the same amount of accuracy, too.

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?

In Rawls's conception, the Veil of Ignorance was meant to prevent biased judgement by prompting individuals to imagine that they do not know facts about themselves (including demographic ones) or society. By existing behind a "veil of ignorance," Rawls posited the we could avoid the influence of bias in just decision-making. In this scenario, a protected class might be a feature/variable that gets in the way or biases our impartial decision making – it is information that we would choose to blind ourselves of in the veil of ignorance approach. Unfortunately, this principle can not account for proxy variables, which may not be directly protected but can introduce some bias nonetheless. Our classic example of this was zip code as a proxy for race in COMPAS. Here, the developers did not include race in their algorithm explicitly, but zip code largely indicated race based on the existing societal biases and organization. #

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.

Based on statistical measures of fairness and a deontological moral framework, COMPAS should not be used to supplement a judge's discretion.

COMPAS violates basic statistical principles of independence and separation. For independence, we see that COMPAS violates statistical parity because it returns more false positives among Black defendants than white defendants; there is a disparate impact as we see deviations from statistical parity between sensitive groups. For the principle of separation, we see a violation of equalized odds. That is, Black and white defendants do not have equal true positive rates (higher for white defendants) and equal false positive rates (higher false positive rates in Black defendants than white). Further, it produces confusion matrices that show only around 64% accuracy – is this really accurate enough to guarantee much insight for the judge? From a deontology perspective, the use of COMPAS violates Kant's categorical imperative that moral agents should not be used purely as a means to an end. By using the data of arrested individuals purely to decide on parole of another moral agent, COMPAS uses these individuals as a means to an end of deciding to grant parole or not. It does not recognize individuals as free moral agents and instead ties them to histories of recidivism (that they did not, in fact, commit). Similarly, we might argue that, if scaled up, the use of COMPAS would again violated Kant's categorical imperative. If all court systems used machine learning to supplement decision-making, there might be a risk of reproducing historical inequalities in the prison system, thus defeating the purpose of individual, contextual parole hearings. Overall, statistically and morally, the use of COMPAS (even in as a supplement) is unjust.