HW 4

Hunter Busick

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?

Section 4.5.2 gives us the mortgage approval rates for different demographics, but we still need more information to assess whether or not this classifier violated equalized odds. To assess whether or not this classifier violated equalized odds, we need the True Positive rates for each racial group and the False Positive rates for each racial group. These numbers will tell us the percentage of each racial group that was correctly granted a loan when they deserved one, and when non-worthy applicants received a loan, respectively.

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 will not hold when we have an oracle classifier as there are no inaccuracies to measure. Any metric that we try to run on a perfect classifier will just come back as satisfied as our confusion matrix will only contain values in the two cells that indicate assigning a value to its correct classification. Additionally, when we have perfectly equal proportions of ground truth class labels across the protected variable, it is also possible to violate this impossibility result. This is because each protected group has the same probability of being selected for the training set and ensures that the protected variable is not one that is swaying the outcome of the model. Thus, we do not have to account for proportions of each group when performing statistical measures of fairness, allowing multiple measures to be satisfied at once.

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?

[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

Rawl's veil of ignorance asks the agent to consider themselves as being without an identity such that the characteristics of oneself are unknown, such as how much money you make or what neighborhood you live in. The goal of this is to encourage one to think as if they had no identity. Knowing this, the Veil of Ignorance would define a protected class as those that should not skew outcomes in an inherently biased manner. These often consist of race, ethnicity, gender, and other variables that we would wish to exclude when attempting to model or utilize an unbiased classifier. In many cases, it is actually important that these protected classes work their way into interpretations of the results so that we can assess whether or not our methods are producing results that do not disproportionately discriminate against any particular class. They may also work their way into the model if a variable is not protected but has a high correlation with one that is protected.

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 software to supplement in a judge's decision on granting parole is justifiable given that the judge does not use it as their sole deterministic factor in a decision. When considering fairness as equality, merit, and need, blending a judge with COMPAS creates a mixture of all three. While the judge is able to consider the merit and need of a person eligible for parole, this process is not always independent because of the difference in biases found between judges. Introducing COMPAS allows for an independent agent to evaluate the case, while still leaving the ultimate decision up to the judge. While some may argue that COMPAS struggles to pass statistical measures of fairness, it is important to keep in mind that the benchmarks set for statistical fairness, such as $\epsilon = .2$, are general rules of thumb and it should not be encouraged to alter training data or results in a fashion similar to *p-hacking*. The use of COMPAS software may also hedge against judges that appeal to any of our 3 primary philosophical theories from earlier. It may cancel out a judge that is overly rooted in evaluating virtuous actions, or may contradict the opinion of a harsh judge that believes he must follow the sentencing guidelines without fault, ignoring good deeds.