HW 4

Ava Klissouras

10/23/2024

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 states that the difference between the false positive rate in the protected class, and the false positive rate in the non-protected class, is less than some predetermined value (legally it is defined as 0.2). Consider a positive event to be denied a mortage. In order to assess this classifier according to equalized odds, we would first need to define a protected class; let's define it as all non-Caucasian races.

Now, we would need to determine the false positive rate within each class (protected and non-protected). This is the probability of the algorithm predicting someone is denied a mortgage, given they are in the (non-)protected class and they were actually granted a mortgage.

So, we would need a data set to test the algorithm on, that contains information about whether applicants were approved or denied for a mortgage, so we can determine the false positive rate for each class.

After testing the algorithm, we'd need to compare the false positive rate in the protected and non-protected classes, and see if their difference exceeds the predetermined threshold. If it does, then the algorithm is biased according to Equalized Odds.

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

Let the symbol $\perp$ denote statistical independence. I could not figure out how to type the symbol we used in class to denote independence.

Let $\hat{Y}_A$ be the value predicted by the algorithm, Y be the true value, and S be the protected variable.

Suppose we have a perfect predicting classifier. Then $\hat{Y}_A = Y$.

Want to show independence, separation, and sufficiency can be satisfied simultaneously.

[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

First, suppose separation is satisfied. Then $\hat{Y}_A \perp S|Y$.

Now, consider that $\hat{Y}_A = Y$. The above statement may be rewritten as $Y \perp S|\hat{Y}_A$. This is precisely the definition of sufficiency, so sufficiency is satisfied.

Now, suppose we have perfectly equal proportions of the ground truth. This means $Y \perp S$.

Since earlier we concluded that $\hat{Y}_A = Y$, substituting $\hat{Y}_A$ for Y in the above equation yields $\hat{Y}_A \perp S$.

Therefore, independence is satisfied.

Since we have satisfied all three statistical fairness criteria, the impossibility result discussed in class does not hold when our two fringe cases are met.

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 is a thought experiment that entails individuals sitting behind an imaginary "veil" that precludes them from knowing whether they are advantaged or disadvantaged; therefore, when asked to determine how to allocate resources, they will allocate them in favor of the least-advantaged, because they can imagine falling into this category. By "walking a mile in another's shoes," they will thus aim to ensure everyone has a baseline quality of life, by allocating resources based on need. As a result, Rawl's Veil of Ignorance would define a protected class as one that is most negatively impacted by disparities in resource allocation (the least-advantaged class).

Even if removed from the data before training our algorithm, the protected variable could make its way into our interpretation of results by way of proxies in the data. For instance, if the protected class is race, variables such as zip code and income may be correlated with the protected class, particularly if wealth/geographical disparities are present. If other variables that remain in the data set are highly correlated with the protected class, the algorithm may utilize them in generating predictions, and predictions may reflect the protected variable as a result.

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 the arguments discussed in class, the use of COMPAS to supplement a judge's discretion is not justifiable. Based on statistical measures of fairness, COMPAS violates independence and separation. In the case of independence, disparate impact is violated because the ratio of the proportion of recidivism among the unprotected class (White race) to the the proportion of recidivism among the protected class (Black race) does not meet the legal threshold of 0.8. However, one might argue that this criteria does not take into account the "ground truth" proportions– what if Black individuals, on a population level, have a greater proportion of recidivism? Now, we can examine separation, which takes this into account. The COMPAS algorithm violates equalized odds, because the false positive rate among Black defendants exceeds that of White defendants above the legal threshold of 0.2. Therefore, the COMPAS algorithm predicts recidivism will occur disproportionately for Black individuals, and is unfair. Now, examining philosophical measures of fairness, fairness as equality is violated because the benefits of COMPAS do not extend equally to White and Black defendants. Black defendants are more likely to be predicted to recidivate than Black defendants under the algorithm, resulting in a greater proportion of unfair convictions/incarcerations. Lastly, based on principles of Utilitarianism, the

use of COMPAS is not justifiable. According to Utilitarianism, to be morally sound, an act must maximize pleasure and minimize pain. However, the drawbacks of using the COMPAS algorithm outweigh the benefits. Although it can be used to aid judges, potentially improving efficiency, the disparate negative impact it creates for Black individuals outweighs this. Moreover, if it used to supplement, not replace, a judge's decision, it will not circumvent the judge's human biases; it may, in fact, augment their impact by introducing algorithmic bias as well. Therefore, the use of COMPAS is not morally justifiable, as it does not have enough benefits to outweigh its negative impacts. As a counterargument, one could argue that the use of COMPAS is justifiable because it satisfies statistical fairness as sufficiency (meaning the ground (recidivism) is independent of Race, given the algorithm's prediction); however, this alone is not enough to justify its use, in consideration with the above arguments presented.