HW 4

Aarjav Jain

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?

-> The text currently provides us with information about the approval rates for different racial groups (Faber, 2018). This is given as follows: Whites (71%), Asians (68%), Latinos (63%), and Blacks (54%). To analyze the classifier in the context of equalized odds, we will require more information. Specifically, we will require data on the "Ground Truth" labels i.e. who was approved and actually repaid their loan (the true positives) as well as who was approved but did not pay back their loan (the false positives). Further, this needs to be broken down by racial group. Both the true positive rate and the false positive rate are important to determine if the equalized odds criterion is to be met since it states that the TPR/FPR across racial groups must be the same. IF it is not, then we can say that there is some form of classifier bias against or for a certain racial group. The different approval rates alone could possibly be justified if they accurately represent the different base rates of loan repayment, or they could represent discrimination - in either case, only one out of two can be achieved: equalized odds or perfect prediction accuracy, but not both.

2

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

Case A

In this case, we have a perfect classifier. When we have a perfect classifier, the True Positive Rate will always be 100% (or 1) for all groups. This is because only the people who will pay back their loan will get approved. The false positive rate will be 0 since anyone who will default their payment will not get approved. This perfect predictor would lead to the TPR being 1 and the FPR being 0 across all races. Therefore we will have satisfied equalized odds while having perfect accuracy.

Case B

In this case, we have perfectly equal proportions of ground truth class labels. This means that the base rate across all groups would be the same. Therefore, when the classifier predicts whether a person should be approved or not, it will lead to the same true positive rate and same false positive rate across each group since the base rate is the same across each group, hence eliminating any differences that were present between different class groups.

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?

-> Under Rawl's Veil of Ignorance, people make decisions about the structure of society behind a "Veil" which prevents them from knowing characteristics about themselves: their race, social status, ethnicity, disability status, etc. Because people do not have this information, they could potentially be a part of any group. As a result, the idea is that they are more likely to set up a society that is more fair and just. However, because they also know that certain groups of people in society face unjust behavior and an inequitable lifestyle, they may make special arrangements or rules in order to protect such groups of people and ensure that society is fair to all. These groups of people who might be disadvantaged due to societal norms due to their race, gender, socioeconomic status, etc. are known as the protected class.

Even if we remove the protected variable, it can re enter our analysis in a few different way. One of the ways is through proxy variables. These can include the zip code of a neighborhood (similar communities in a region), name (infer the race or ethnicity), primary language spoken, etc. With all these variables, the protected variable can likely be inferred and reintroduced to the results. Another way this can happen is through patterns in correlated variables. An example of this could be income which can directly influence the neighborhood one can afford to live in. In this manner, one could reintroduce some sort of protected variable as the algorithm can pick up on this correlation between income and neighborhood and thus reintroduce some form of bias similar to if the protected variable was still present.

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 in its current state is not justifiable due to its inherent biases and fairness issues. From a utilitarian perspective, where one seeks to maximize the good of society overall, COMPAS is not a good tool as it is subject to racial bias and poor prediction accuracy at only about 65%[4]. Additionally, it violates the principle of equalized odds as it has different false positive rates for different racial groups, with Black people having a much higher chance of being incorrectly labeled as repeat offenders. Moreover, it violates Rawl's Veil of Ignorance principle, where any person deciding on societal rules would not be willing to choose an algorithm such as COMPAS which would then unfavorably treat an already disadvantaged group. I believe that the use of COMPAS, even if as a tool to supplement a judge's discretion, is not justifiable due to its current performance and its inability to fulfill the above measures of fairness.

---

1. https://link.springer.com/article/10.1007/s00146-023-01676-3 (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

↩

4. **https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm** (https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)↩