CS 281, Spring 2022

Homework 1: Machine Bias

Due April 17, 6PM

Correspondence to: agataf@stanford.edu

Introduction

In 2016, ProPublica published an analysis of racial disparities in the performance of a commercial tool called COMPAS, which provides “risk assessments" of future recidivism of arrested individuals, and which is used by courthouses across the US for sentencing and pretrial release decisions. The article became the focal point of discussions around algorithmic fairness in the computer science community, and around the appropriateness of the use of automated tools in sentencing among social scientists and activists.
The computer science conversations largely focused around the tension between two incompatible definitions of fairness - separation and sufficiency. ProPublica argued that differences in error rates (the violation of separation) between white and Black defendants was evidence of unfairness. Northpointe - the commercial entity behind COMPAS - responded with their own analysis, which showed similar rates of calibration (sufficiency) between the two groups. These two perspectives of fairness extend beyond mathematics - they each have their own social and political interpretation. In this assignment, we’ll explore both the mathematical and the socio-political aspects of the debate.

You will be asked to read the original ProPublica article, and answer some comprehension questions (Problem 1). Then, in Problem 2, you’ll have a chance to explore the dataset that ProPublica used in their analysis (note that while this dataset contains real COMPAS scores, this is not the dataset used to *derive* them - the actual training data and model were never released). In Problem 3, you will consider the meaning of different fairness metrics introduced in class in the context given by the article, calculate them for the ProPublica dataset, and interpret the meaning of the discrepancies between groups. In Problem 4, you’ll be asked to step back from the algorithmic framework and consider some of the questions posed by social scientists and activists in light of COMPAS.

Note that the intuition behind the incompatibility of the two notions of fairness is subtle, and may not be immediate - it was elegantly described by Sam Corbett-Davies and Sharad Goel in their 2018 article “The Measure and Mismeasure of Fairness”.

A. Machine Bias (6 points)

Before we jump into fairness metrics, let’s understand the original ProPublica analysis. Carefully read Machine Bias by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner and answer the following questions. Keep your answers succinct (1-2 sentences).

1. (2 points) What was the purpose of the model? What intervention was it associated with?
2. (2 points) What assumptions (about the data, the outcome variable, the way the model will be used etc.) did Northpointe make in the construction of the model?
3. (2 points) How did ProPublica analyze the model? What were their primary findings?

B. The Broward County dataset (17 points)

The dataset used in the ProPublica analysis describes ex-offenders screened by COMPAS in 2013 and 2014. There are 53 columns in the original data describing length of jail stays, type of charges, the degree of crimes committed, and criminal history.

In the provided starter code (https://github.com/stanfordaiethics/hw1_public), the COMPAS dataset is loaded for you in a variable named raw_data which is a `pandas.DataFrame` object. (Learn more about pandas library https://bit.ly/3wTKBbQ). Each row of `raw_data` contains various feature values for a single person. The name of the features can be found in `raw_data.keys()`. Some examples of these features include `name`, `age`, `days_b_screening_arrest`, and etc. See here for a data dictionary derived from ProPublica's analysis.

1. (2 points) **Pre-processing**: Following ProPublica's analysis, we will start with some data pre-processing to remove entries which don't correspond to individuals with proper COMPAS scores. From `raw_data` remove the rows with the following feature values:
    - If the charge date, i.e `days_b_screening_arrest`, of a defendant's COMPAS scored crime was not within ±30 days from when the person was arrested, we assume the score is not associated with the listed offense. Thus only rows with `|days_b_screening_arrest| ≤ 30` should be kept.
    - ~~The rows with the recidivist `flag is_recid == -1` corresponds to rows where we could not find a COMPAS case and should be discarded.~~
    - ~~In a similar vein, ordinary traffic offenses, i.e rows with a `c_charge_degree` value of `'O'`, will not result in jail time and should be removed.~~
    - ~~Remove rows with no proper `score_text`, i.e N/A~~

    The filtered dataset should be stored in a variable named df and still be a `pandas.DataFrame` object.
    - Report the number of remaining rows in your filtered dataset df.

2. (2 points) **Demographics**: Count the number of individuals across racial identities, age categories and binary sex categories. Report your numbers.

3. (7 points) **The distribution of COMPAS scores**. The scores generated by COMPAS software are contained in the `v_decile_score` variable. Plot the distribution of scores as a histogram:
    a. (2 points) For the entire population
    b. (2 points) For 4 groups: Black women, white women, Black men, white men
    c. (2 points) Repeat the plots, this time using stacked bars: color-coding the proportion of each bar corresponding to individuals who did vs did not recidivate (`two_year_recid == 0` in one color, `two_year_recid == 1` in another).
    Hint: find pointers for creating stacked bars here
    https://www.pythoncharts.com/python/stacked-bar-charts/
    d. (1 point) What do you notice?

4. (6 points) **Reflection point.** Before we jump into data analysis, let us reflect on the composition of the dataset. Keep your answers succinct (2-3 sentences).
    a. (1 point) What do you notice about the demographic composition of the dataset?
    b. (1 point) What are the differences between the distribution of COMPAS scores between groups?
    c. (2 points) Think of the process of data generation. Who's overrepresented in the dataset? Who is not in this dataset? What factors (beyond who commits crime) might be impacting what you see in the dataset?
    d. (2 points) If this data was used to *train* a model that predicts recidivism within 2 years, how could these properties of the dataset be impacting model predictions?

Our analysis, going forward, will focus primarily on Black and white individuals - since those are the two groups most represented in the data.

C. Fairness metrics of COMPAS scores (28 points)

In this section, you will analyze the scores provided by COMPAS using the three types of fairness metrics introduced in class - *independence*, *separation* and *sufficiency*. First, you will be asked to consider the meaning of these metrics in the context of the article, and then you will plot and interpret the meaning of the discrepancies between groups.

Before we dive in, there are a couple of things to note: first, we're depending on recorded scores, and do not have access to the model that was used to generate them (which is proprietary). Second, while COMPAS is a "risk assessment tool", we do not have access to probability values - only decile scores. Finally, the outcomes can be grouped and analyzed in at least three different ways: we can compare individuals

- within each decimal score category,
- within each risk category or
- within each decision category (if we assume a uniform decision threshold that would be used to binarize scores into two recommendation categories: release/do not release).

All will reveal slightly different properties, and our assumptions about how scores are interpreted and used by decision-makers will matter a lot in interpreting the fairness of algorithms.

For consistency with ProPublica's analysis, we will assume that the decision threshold is at COMPAS risk 5 - in other words, that all individuals in the low risk category (1-4) are recommended for release (`d=0`), but all individuals in the medium (5-7) and high (8-10) risk categories are not (`d=1`).

Hint: add a binary column to your data frame that corresponds to this binary decision.

1. (6 points) **Intuition behind fairness metrics**
    a. (4 points) For each fairness metric, list a statistical property it represents (using the symbols *d*, *y*, *c* as introduced in class), and an interpretation of the desirable property it represents in the context where the COMPAS risk score is used for pretrial risk assessment.

Metric name	Statistical property	If this metric is satisfied, we can be sure that:
Independence		
Separation - FNR		
Separation - FPR		
Sufficiency - PPR		

    b. (2 points) The ProPublica article prioritized separation definitions in their analysis. In their rebuttal, Northpointe argued that sufficiency would be a more appropriate metric to evaluate fairness in this context. Which do you believe to be more relevant in this context and why? (1-3 sentences)

2. (12 points) **Calculations**. For each of the following fairness metrics, calculate the appropriate statistic for 5 variants of the data: (1) all individuals, (2) Black women, (3) white women, (4) Black men, (5) white men. Assume the decision threshold of 5 in all cases except for sufficiency (using decile scores).
    a. (4 points) Independence.
    b. (4 points) Separation
        i. false negative rate
        ii. false positive rate
    c. (4 points) Sufficiency (positive predictive value)
        i. using decile scores
        ii. using binary decisions

3. (10 points) **Interpretation.** Given your interpretation of the fairness metrics in the previous section, what do the results reveal about the potential consequences of using the algorithm for the considered groups? Give 1-2 sentences for each fairness metric, focusing on groups whose results differ most from the metrics calculated for the overall population.
    a. (2 points) Independence.
    b. (4 points) Separation
        i. false negative rate
        ii. false positive rate
    c. (4 points) Sufficiency (positive predictive value)
        i. using decile scores
        ii. using binary decisions

D. Final reflection point (9 points)

Please respond to the following questions (2-3 sentences or bullet points each).

1. (3 points) Do you think it is desirable to use the COMPAS model in pretrial release hearings, given the analysis? Why or why not? Which fairness metrics, if any, contribute to your opinion?
2. (3 points) Ben Green argues even “fair” algorithms can reinforce discrimination. What are some ways in which that could happen?
3. (3 points) One proposal for addressing concerns about the use of the COMPAS dataset in the criminal justice system is to make the algorithm and its data more transparent and accessible to the public. What are some potential benefits and drawbacks of this approach? How might increased transparency help to address concerns about bias and accuracy in the use of the COMPAS dataset?

References

1. Angwin, Julia, et al. "Machine bias." *Ethics of data and analytics*. Auerbach Publications, 2016. 254-264.
2. Corbett-Davies, Sam, et al. "A computer program used for bail and sentencing decisions was labeled biased against blacks. It’s actually not that clear." *Washington Post* 17 (2016).
3. Urban Spatial. "People-Based Machine Learning Models and Algorithmic Fairness." Urban Spatial, 22 Feb. 2021, urbanspatial.github.io/PublicPolicyAnalytics/people-based-ml-models-algorithmic-fairness.html.
4. Corbett-Davies, Sam, and Sharad Goel. "The measure and mismeasure of fairness: A critical review of fair machine learning." *arXiv preprint arXiv:1808.00023* (2018).
5. Green, Ben. "The false promise of risk assessments: epistemic reform and the limits of fairness." *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020.

6. Jacobs, Abigail Z., and Hanna Wallach. "Measurement and fairness." *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021.