CS 281, Spring 2025

Homework 1: Machine Bias

Due April 16, 1:00 PM

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 for arrested individuals and is used by courthouses across the US for sentencing and pretrial release decisions. The ProPublica 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 discussions around algorithmic fairness largely focused on the tension between two incompatible definitions of fairness - *separation* and *sufficiency (see Lecture 2)*. ProPublica argued that differences in COMPAS’ 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 (i.e., sufficiency) between the two groups. These two perspectives of fairness are highly nuanced, extending beyond mathematical differences to 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 of 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 was elegantly described by Sam Corbett-Davies and Sharad Goel in their 2018 article “The Measure and Mismeasure of Fairness”.

Deliverables: All starter code related to the first homework is in hw1.ipynb and instructions for setup and submission can be found in the README.md. Your code and short answers should directly go into this file. Please submit a single .pdf file of your notebook to GradeScope containing your code and responses to the questions. Please check that your final answer for all coding parts is clearly visible in the PDF you submit.

A. Machine Bias (6 points)

Before we jump into fairness metrics, let’s first understand the original ProPublica analysis of COMPAS. 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 COMPAS model? How did it impact real people, and through what decisions?
2. (2 points) Suppose you were creating the COMPAS model for Northpointe. Name one significant bias you would be concerned about in how the data was collected and one significant concern you would have regarding how the outcome variable is generated.
3. (2 points) What were the primary findings of ProPublica after analyzing the COMPAS model?

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 the length of jail stays, type of charges, the degree of crimes committed, and criminal history.

In the provided starter code, the COMPAS dataset is loaded for you in a variable named `raw_data` which is a `pandas.DataFrame` object (you can learn more about pandas library here: https://bit.ly/3wTKBbQ). Each row of `raw_data` contains various feature values for a single person. The names of the features can be found in `raw_data.columns()`. Some examples of these features include `name`, `age`, `days_b_screening_arrest`, 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 that do not 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 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.[1] 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 (as four separate histograms)
   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) Analyze the differences between each plot and report what 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) According to the US census, black individuals represent 13% in the county. Compare this number to demographic composition of the dataset. Does the dataset over-represent or under-represent Black individuals?

[1] While race, ethnicity, and sex are nuanced identities that are hardly reducible to categorical features, the standard set by algorithmic fairness literature typically assumes these more simplified categories.

b. (1 point) What are the differences between the distribution of COMPAS scores between the four groups?
c. (2 points) Reflect on how the Propublica analysis constructed the dataset and potential limitations and biases of the approach. What groups are overrepresented in the dataset? What groups are underrepresented from the dataset? Comment on how additional factors (e.g., policing practices) could have contributed to the observed crime rates in the dataset.
d. (2 points) Suppose we used this dataset to train a new model that predicts recidivism within 2 years. Explain one possible consequence of using the model that arises from properties of the dataset.

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 ProPublica 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, recall that while we have access to COMPAS's predicted scores, Northpointe never released the COMPAS model that generated the scores. Second, while COMPAS is a "risk assessment tool", we do not have access to probability values - only decile scores.

To compare individuals, the "recidivism risk" outcomes can be grouped and analyzed in at least three different ways:

- within each decile score category,
- within each risk category (these are coarser categories of decile scores) or
- based on each binary decision category (i.e., if we assume a single decision threshold like >0.5 that would be used to binarize scores into two recommendation categories: release/do not release).

Each outcome grouping will reveal slightly different perspectives on COMPAS fairnessIn fact, interpreting the fairness of COMPAS, and all algorithms in general, depends highly on the outcome variable, the assumptions on its meaning, and how it will be used by decision-makers.

Unless mentioned otherwise, we denote `d` as the binary release/do not release decision based on the outcome variable. For consistency with ProPublica's analysis, we will assume that the decision threshold is at COMPAS risk category 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 `d` 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 the 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)

Our analysis, going forward, will focus solely on comparing Black and white racial groups, as these identities are the groups most represented in the data. For the questions below, first filter the dataset to include only these groups.

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, and (5) white men. As mentioned above, assume the decision threshold of 5 in all cases except for sufficiency when 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 binary decisions
        ii. using decile scores (the `v_decile_score` variable)
            - Hint: `d` is no longer a binary variable here, and you should look at the distribution of your outcome y=1 at all possible levels of `d` for each group.

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 the 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 binary decisions
        ii. using decile scores

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) Suppose you are a policymaker working to oversee and regulate the use of AI models to ensure fairness. Name three things you would want to see from a future COMPAS v2.

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.