HW3 - Fairness and Bias Analysis in Machine Learning

CSCI 535 - University of Southern California

Due Date: April 25, 2023

1 Goal

In this exercise, we analyze a given database to investigate possible biases in the data and learn about avoiding unintended biases for achieving a fair model. Bias in the data can affect the fairness of a machine learning model. Many of the current databases are influenced by the stereotypes in people's perceptions, during data selection and labelling.

When developing machine learning models, it is important to be aware of existing biases in your data so that you can come up with a fairer model. For example, in a health dataset, your data might have an age-related bias, noting that age has a negative correlation with health. This is a bias that might be useful in any downstream task that is identifying health issues, without discriminating against a certain age group. However, models for job qualifications should not discriminate by age, as age is a protected attribute in hiring decisions. In short, a dataset may be biased against individuals from a protected group (e.g., race, gender, religion) and these are biases you need to avoid in order to protect under-represented/protected groups from discrimination. The first step toward this goal is to simply become aware of any existing biases using statistical analysis.

2 Dataset

For this assignment, you have access to the labels from the 10k US Adult Faces Database [1]. This dataset contains face photographs and several measures for 2,222 of the faces including multiple psychological traits, e.g. intelligence, trustworthiness, sociability, etc. You can download the dataset here[1].

Additionally, the dataset provides annotations of demographics about the image including gender and race. All of these information are collected as perceived by a number of 15 annotators, and you are provided with the mean of the annotations for each image. There may be intrinsic associations of certain traits with specific ethnicities and genders, influenced by common biases and stereotypes.

You are required to investigate these biases, in this this assignment. For the following tasks you need to focus on the following set of attributes. To learn about the labels, please go over the README file from the database.

- **Psychological traits**: confident, egotistic, intelligent, kind, responsible, trustworthy, aggressive, caring, emotional, friendly, sociable.
- **Race**: White, Black, East Asian.
- **Gender**: Female, Male.

3 Your tasks

1. Among all the listed traits and demographics, investigate whether race is playing a role in the annotators' perception of traits. For example, is a certain race inherently perceived as more/less intelligent, aggressive, etc. You can compare the traits across races by looking at the mean, median, statistical tests (to check for

[1] https://drive.google.com/drive/folders/1F3BZVJAdcPjUAC9Vrr7wSC5Xbpnj4KyO?usp=sharing

significance) across the different groups. You can use a one-way ANOVA test as there are more than two groups. Identify and report the biases and explain your analyses.

2. Repeat the previous task, this time focusing on gender. Use t-test as there are two groups.
3. Rank the top 5 most significant biases you have found (from either race or gender) and report the corresponding p-values (or other adopted measure).
4. Plot the histograms for the most significant biases you found from the previous task and interpret the results.
5. In the last column of the file data.csv, you are provided with predictions of a classifier for whether or not each candidate is eligible for a competitive job position. The **four-fifth rule** prescribes that if a selection rate for any disadvantaged group is less than four-fifths of that for the group with the highest rate, there is an adverse impact on that group. Using this rule, identify any possible adversities toward certain groups.
6. What could be other contributing factors to bias in ML systems, aside from human labeler bias?
7. How can you mitigate such biases in machine learning? Please explain.

Note: You are only allowed to use the labels provided by us with this assignment. Please refrain from using the original data; assignments submitted with any other labels or data will not be graded.

4 Helpful Resource

- A Tutorial on Fairness in Machine Learning

5 Submission deadline

Please submit your report and source codes in a compressed folder named named as **lastname_firstname.zip** on Blackboard until **04.25.2023 at 23:59**. Late submissions will incur a 10% penalty per day.

References

[1] Bainbridge, W.A., Isola, P., Oliva, A.: The intrinsic memorability of face photographs. Journal of Experimental Psychology: General **142**(4), 1323 (2013)