Project Proposal

Jonathan Rittmo Sara Thiringer

2020-10-16

Introduction

As more data and consequently more data-driven decisions have entered the world, the problem with algorithms reinforcing discriminatory structures have received an increasing amount of attention. Several cases showcase how algorithms that were designed to be neutral decision-makers have made discriminatory predictions. Examples include facial recognition being less accurate for people with darker skin (1), adds for higher paid jobs being shown more frequently to men (2), healthcare predictions underestimating the illness of black people (3) as well as individual tech company scandals such as the Apple card seemingly granting men a higher credit limit than women (4) and Amazon's automated recruitment tool unrighteously favoring men (5). However unintentional, these examples show the need for an awareness of discrimination and fairness when collecting data, training models and using predictions for decision-making.

Aim

In this project we therefore aim to evaluate methods that can account for discrimination bias in predictive models. Such discriminatory consequences can have many sources, such as imbalance in the dataset, correlations between variables of protected characteristics[1] caused by differences in demographics between groups or by previous discrimination such that human bias of protected characteristics have guided the labelling of the data. As such, bias can be introduced in a model because the data reflect direct discrimination bias in the population. As an example, consider an employee wage dataset. If a model, e.g., predicts equal wages between two groups when all variables are available except for this specific grouping (say gender) have been taken into account but different wages when gender is introduced there is evidence of direct discrimination. If the model predicts unequal wages despite removal of variables with protected carachteristics, this can be a sign of the carachteristics being correlated to other variables and hence imply indirect discrimination.

Common sources of discriminatory bias in datasets are:

- Undersampled groups.
- Having skewed samples.
- Too few variables, i.e. a limitation of features for the model to train on.
- Human bias having guided the labeling process
- Uncontroversial predictors being correlated to protected carachteristics

[1] Protected carachteristics refers to human carachteristics by which poeple can be subject to discrimination. In Sweden, these are stated by the Discrimination Act (2008:567) and include sex, transgender identity or expression, ethnicity, religion or other belief, disability, sexual orientation or age.

Based on what causes the discriminatory bias, the problem can be dealt with in different ways. We aim to investigate a couple of solutions. As an example, the problem of imbalanced data could be tackled by using synthetic data, i.e. parametric bootstrapping of the original data set to even out imbalance. The questions remains how different models perform when being trained on such data.

Other sources of discriminatory bias can be more difficult to deal with. In case of direct discrimination, removing the biased variable for the model will most probably be enough. However, if the bias persists even after removal, we need to further inspect other variables and consider introducing penalties or removing correlation. There are several methods introduced in the course to do this, such as penalized regression or Random Forest-style bootstrapping and Principal Component Analysis in case of decorrelating variables.

In order to be able to evaluate these different methods, we need a measurable criterion for fairness. We could then build build algorithms in such a way that we optimise with regards to fairness.

Main question

- How can we quantify fairness in order to be able to evaluate algorithmic fairness?
- What methods are available to increase algorithmic fairness? In what type of situations do they apply? (i.e. In what kind of situations can we expect them to be successful?)

Example dataset

Since our aim is to evaluate methodology rather than to analyse a specific dataset the choice of data matters most in that we want sets where some bias is present. Examples include but are not limited to:

- Glassdoor gender pay gap. A dataset containing wage data and demographic from Glassdoor, along with education, field, seniority etc. Source.

- Silicon valley diversity data. Diversity of the workforce in Silicon Valley. Source.

- Wages (not ISLR). Containing both gender and race. Source.

- The Demographic /r/ForeverAlone Dataset (very imbalanced between sexes). Source.

Example analyses

Since applying the non-discrimiation criteria will require extensive work for each dataset what is given here is solely an example of how we might tackle this issue for the /r/ForeverAlone dataset. This dataset consists of the variables of interest (I'm just listing all of them now but perhaps we should remove a few?):

- `gender`
- `sexuality`
- `age`
- `income`
- `race`
- `bodyweight`
- `virgin`

- `prostitution_legal`
- `pay_for_sex`
- `friends`
- `social_fear`
- `depressed`
- `what_help_from_others`
- `attempt_suicide`
- `employment`
- `job_title`
- `edu_level`
- `improve_yourself_how`

In this analysis we want a model able to predict the risk of an individual attempting to commit suicide where `gender` is our protected variable. First and foremost we would need to create synthetic data to even out the imbalance between gender groups. For simplicity, observations with other genders than male or female will be discarded in this example. This would be done by parametric bootstrapping of variables where distributions are estimated by the female sample in the dataset. Since most variables are nominal with only a few levels most can be estimated with a binomial distribution.

We would then need to look at the joint distributions of the variables of interest