Project Proposal

Jonathan Rittmo Sara Thiringer

2020-10-16

Introduction

In this project we aim to evaluate methods that can account for discrimination bias in predictive models. This can for example happen when a dataset on which the model is based is unbalanced with regards to a certain group, i.e. observations are fewer. Bias can also be introduced in a model because the data reflect true discrimination bias in the population. Take an employee wage data set as an example. 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 taking into account but different wages when gender is introduced there is evidence of true bias.

How should this be handled? 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. But how do different modelling methods fare when they are being trained on such data?

The second problem is more difficult. In the example given above the easy way to deal with such a problem is just to remove the biased variable for model training. However, when the problematic variable is correlated with other predictors in the model the bias can persist even after removal.

Aims

- Identify and evaluate methods for dealing with persisting discrimination bias.
  - Filter observations if not represented?
  - Manipulate target data?
- Evaluate models using synthetic data.
- Identify a method for quantifying fairness and compare accuracy vs. fairness in different models and methods.

Example datasets

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

For continues wage data in the Glassdoor and Wages dataset the goal of the model would be to predict wages between gender and/or racial groups. This would be done using methods such as regression trees and neural nets. For the specific datasets mentioned no additional modification of the sets would be needed except for taking the log wage, due to the known skewness of such data. We would then compare methods such as averaging out differences before training the model or filtering non-representative observations (????).

For the Silicon Valley and ForeverAlone data sets classification models would be built using penalised logisitic regression and boosted classification trees. In the Silicon Valley set the type of job held by an observation would be predicted and in the ForeverAlone set depression or attempt to commit suicide could be predicted. In both these sets women are severely underrepresented and would thus constitute the basis for our analysis of how to deal with such bias.