Project Description

Rasmus Bryld (s183898), Matilde de Place (s183960), and Sunniva Punsvik (s183924)

26. februar 2020

Fairness and bias are relatively new concepts in the realm of computer science, and over the last decade, the quality and accuracy of our different models (especially our machine and deep learning) have increased significantly. The outcome of this success has led us to introduce neural networks as part of various decision making tasks. In the US there has been a rise of using software to do risk assessments to help in court and every stage of the justice process, which has received heaps of criticism because this software, which is used in several states already, is biased against African Americans. This project will investigate what it means to have bias in data and how it may affect the algorithms we use to help with decision making, likewise to distinguish between good and bad bias. Additionally, we want to implement a classifier to be able to understand what exactly is going wrong when we use an algorithm to assess and generate risk scores, as well as validating. Lastly we want to have an in depth discussion and ethical considerations of using such technology and what it inherently means that we as a society sentence people to years in prison.

We are working with the COMPAS dataset and thus, we will investigate bias in this dataset and implement a classifier based on COMPAS. We will use statistical analysis to both do an analytical approach to search for bias in data and validate our classifier. After completing analysis, implementation, and validation; we want to have an immense focus on the ethical aspects of using this sort of technology. We want to provide different perspectives and views on this subject, as it easily may lead to become too subjective (and even a sensitive topic to some). However, that being said, we want to be thorough with the discussion because if we want to continuously use - or increase usage - of such technology, we ought to know its legal limitations, or rather, how we can assure that our algorithms are as fair and unbiased as possible.

Our success criteria for this project is to be able to train a classifier on the COMPAS dataset and do somewhat complex implementation and conduct detailed statistical analysis as a tool to produce results and reflect upon those results - and how this bias influence our classifier. We also want to be able to adjust our classifier to reduce bias. Additionally, we want to become more aware of how we use newly developed technology in an appropriate manner, knowing its strength and what consequences they entail, and how it is used in real-life and society.

We hope that the outcome for this project is that we are able to learn this new concepts that are introduced into our disciplines and what they mean in a technological and the real world setting. Conducting independent analysis and apply the methods and theories we have learned. We also hope to achieve knowledge on how to reduce discrimination in our technology.

3 Learning Objectives

1. Design and carry out a bigger project

2. Apply statistical tools to find bias in data and implement algorithm to alleviate negative biases
3. Discussing the ethical aspects of bias and fairness in computer science