The Societal Challenge: Legal Perspectives on Discrimination

Robin Schmidt & Floyd Kretschmar
MSc Informatik
University of Tübingen
Matriculation number 4255055 and 4205979
[rob.schmidt|Marcel.Mustermann]@student.uni-tuebingen.de

Abstract

This template provides guidance on the structure for your 4 page report. Please feel free to deviate from the proposed structure if you feel that it is useful; but try to follow the spirit of the guidelines. In the abstract, summarize the topic of your report in about 5-8 lines of text. Do not cite your assigned papers here, but instead give a very concise overview over the insights you report on in this text.

1 Introduction

Description of Fairness in ML (target variable/class labels, training data, feature selection, Proxies, Masking)

Defining any kind of target variable or class labels is always a very subjective process, where the presented problem could unintentionally be parsed in a way which systematically disadvantages certain classes [1]. In addition to that, the training data could be biased by either considering cases in which prejudice has played a role or simply over- or underrepresenting a certain class [1]. If this data gets used, it would lead to a discriminatory model.

2 Relevant Work

2.1 Fairness and machine learning: Limitations and Opportunities

Discuss your first assigned paper [2]. Outline the main idea and key results. If suitable, reproduce key mathematical insights. Ideally, also provide critical comments of your own were suitable. But make sure to clearly delineate the ideas and experiments in the assigned paper from your personal opinion or analysis.

2.2 Fairness in Criminal Justice Risk Assessments: The State of the Art

Discuss your second assigned paper [3]. Outline the main idea and key results. If suitable, reproduce key mathematical insights. Ideally, also provide critical comments of your own were suitable. But make sure to clearly delineate the ideas and experiments in the assigned paper from your personal opinion or analysis.

2.3 Big Data's Disparate Impact

The whitepaper "Big Data's Disparate Impact" [1] by Solon Barocas and Andrew D. Selbst is separated into three main parts, which deal with slightly different topics regarding fairness in machine learning, in particular data mining. The first part focuses on the various ways through which data

mining can discriminitate certain classes, while the second and third part discuss the liability issue of discrimination in data mining for the american title VII (equal employment opportunity) [4] of the civil rights act and the difficulty for future legal reforms.

According to their studies, there are five main structures in data mining which can cause discrimination for certain classes. In particular, these are the "definition of the target variable and class labels" (I), "training data" (II), "feature selection" (III), "proxies" (IV) and "masking" (V) [1]. All of these topics have already been clarified and described regarding their extent in section 1 and therefore won't need special attention here.

In the american civil rights act, especially in title VII, there are two presented cases for discrimination, namely "disparate treatment" and "disparate impact", which also find usage in the presented whitepaper. While disparate treatment describes an uneven behavior towards a certain person due to a particular characteristic (e.g. gender, race or religion), disparate impact represents a neutral rule which treats everyone equally in form, but has a damaging effect on a subset of people with such a certain characteristic.

In their whitepaper Barocas and Selbst argue that formal liability in disparate treatment doesn't correspond to any special step within data mining and that using a protected class as an input for any classificatory model should be a legal harm, because this process corresponds to the employer classifying and differentiating potential hires according to exactly this protected class [1]. They also show that the disparate treatment either occurs at the decision to apply a biased predictive model or when the biased result gets used for the ultimate hiring decision and draw the conclusion that the disparate treatment doctrine doesn't regulate discriminatory data mining to a satisfying extent [1].

While considering the disparate impact doctrine, the authors state that in such a case the plaintiff must prove that "a particular facially neutral employment practice causes a disparate impact with respect to a protected class" [1] [4].[1] In response, the defendant-employer is then allowed to justify the challanged practice by showing the job relation and business necessity.[2] The plantiff then still has the chance to show that an alternative, less discriminatory employment practice could have been used instead [1]. For the case of data mining this means that liability regarding disparate impact can be caused by using a non job related target variable [1]. As soon as the target variable is shown to be job related, there are two questions which need to be answered. First, whether or not the model is predictive of the trait and secondly if the model with statistical significance predicts what it is supposed to predict [1]. Barocas and Selbst also explain that it is hard to know which features would make an existing model more or less discriminatory and therefore proving that a less discriminatory alternative would exist becomes a very hard task to solve [1].

The presented legal reform possibilities can be separated into *internal* data mining issues as well as politicial and constitutional *external* constraints [1].

2.4 Machine Bias

Discuss your forth assigned paper. Outline the main idea and key results. If suitable, reproduce key mathematical insights. Ideally, also provide critical comments of your own were suitable. But make sure to clearly delineate the ideas and experiments in the assigned paper from your personal opinion or analysis.

3 Discussion

In this section you can summarize and link your assigned reading. Try to distill an overall insight from the papers, not to make a laundry list of individual results. Did you come across open questions that were not answered in the papers? Are there hidden pitfalls or problems that, in your opinion, the papers do not solve or marginalize? Provide a critical but constructive reading without being dismissive. Ideally, try to do some literature research of your own to find follow-on papers or related works.

[1] 42 U.S.C.§2000e-2(k)(1)(A)
[2] *Id.*

4 Summary

Provide a concise summary of your findings, in about 3-10 lines of text.

5 Appendix: Possible Additions

Here we can provide all relevant additional information.

References

[1] S. Barocas and A. D. Selbst. Big data's disparate impact. *SSRN Electronic Journal*, 2016. doi: 10.2139/ssrn.2477899.

[2] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2018. `http://www.fairmlbook.org`.

[3] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. Fairness in criminal justice risk assessments. *Sociological Methods & Research*, 104(6), 2018. ISSN 0049-1241. doi: 10.1177/0049124118782533.

[4] U.S. Equal Employment Opportunity Commission. Title VII of the Civil Rights Act of 1964, 1964. URL `https://www.eeoc.gov/laws/statutes/titlevii.cfm`. Accessed: 2019-05-19.