Lab 2: Data Science Ethics

Draven Schilling

CS4981

9-18-19

Using data science to explore large sets of data can lead to profound findings, but such findings are susceptible to ethical dilemmas. In this report I will be analyzing one such dilemma; algorithm fairness and how societal biases can impact the data models created.

So what is "algorithm fairness" and why is it important? It refers to the idea that methods of data analysis are susceptible to modes of societal bias, that is the results can be conveyed in a manner that illustrates a preference towards a specific group of people based on ideas such as race, religion, or sex either intentionally or unintentionally. By doing so, the models would thereby suggest a particular group be viewed in a negative light. Though sometimes grounded in biological facts such as the idea that men are more suited for physical labor then women, it's important for the data scientist to be careful when applying methods of analysis as to not generate conclusions that target a specific group or mislead an audience.

A modern example of algorithm fairness being manipulated by societal bias could be the Microsoft twitter AI "Tay" which was released in 2016 as a way of showing the public the growing potential of AI [1]. Tay was meant to converse with other Twitter uses and learn from them to develop a distinct identity. Unfortunately, after users found out her AI could be influenced by social biases, she was quickly manipulated into developing heightened discriminatory biases. Microsoft was therefore quickly forced to take the bot offline. After tweaking the AI, Tay was unintentionally released again a few months later only to fall victim to the same biases. Microsoft promises they will one day release Tay again after more rigorous testing, but almost three years of silence alludes to the challenges of addressing this problem.

Tay is an extreme example of an AI taking on societal biases as a result of the data fed into it. Tay was unable to distinguish and ignore the intentional biased information and as a result came up with a personality that reflected the same biases. In the case of Tay, the consequences were minimal. It seems as if for most general twitter users found the situation to be a good laugh and ultimately Microsoft as a company suffered very little. What Tay did do was bring more attention to how human based social biases were far from being absent in AI. Given these results it seems as if the risk was worth the reward because the damage was relatively harmless while the knowledge and social awareness gained was significant.

Another recent example of algorithm un-fairness was found in an effort to improve education quality in Washington D.C. Education reformers believed schools were underperforming because teachers were not doing a good job of teaching students [2]. To remedy this, they implemented a teacher assessment tool "IMPACT" which was meant to identify the worst teachers based on their students performance who would then be replaced. Unfortunately, this tool brought to life particular social biases which prevented it from being effective at solving the district's root problems. The tool failed to factor in relative location and the relative wealth of the surrounding school area and as a result, teachers in poor areas were being specifically targeted because their students were less successful then students in richer areas despite many 'poor' performing teachers having high metrics in other areas.

The tool illustrates that wealth gap is another factor which distorts algorithm fairness as it is much harder to attract and retain great teachers in poor neighborhoods and have their effectiveness show compared to equivalent teachers and students in wealthier areas. Ultimately this tool exacerbated the districts problems and only served to reaffirm social beliefs. Based on this evidence, the tool created more problems than it solved and was not worth the risk.

A final example of algorithm fairness manifests itself in the form of a predictive study using machine learning for diagnosis and treatment of medical patients [3]. This machine learning model would use existing medical record databases for reference of diagnosing and treating new patients. The goals were to provide objective and accurate diagnosis and treatment options to healthcare patients. Unfortunately, under this model it was quickly noticed that potential social biases could exist. One researcher predicted recommended treatments may be influenced based on wealth, status, and sample representation. For instance, the algorithm learns to treat patients of low socioeconomical status according to lower standards of care. Existing observational studies enforce this belief with evidence suggesting that practitioners have innate social biases that often influence their patient recommendations. Researchers concluded there was a very high probability the algorithms learns and reflects the same biases.

Given the goals of the project were to provide accurate diagnosis and treatment, I believe the risk is not worth the benefit. After researchers concluded there was a high likelihood of the algorithm taking on such biases, the initial goals should be considered a failure because the results from the experiment would not give researchers what they were looking for. Despite this, they could still execute the experiment and adapt with new research questions in mind. That is, the data could still be probed to learn interesting information. Ultimately in either case, I'm not sure using this model to prescribe actual recommendations to real patients is a good idea until extensive testing is done.

Overall, in most cases it seems like risking algorithm fairness is not worth the benefit. Generally, the goals of performing data science are to reduce bias and create objective models that can be interpreted and lead to new findings. If an algorithm is unfair in its analysis the results can often be misleading. With this in mind, it should be up to the data scientist to ensure that whichever algorithm they use, it best captures the goals of the project and leads to meaningful findings whether that is meant to capture social biases or not.

[3]Gianfrancesco, Milena A, et al. “Potential Biases in Machine Learning Algorithms Using Electronic Health Record Data.” *JAMA Internal Medicine*, U.S. National Library of Medicine, 1 Nov. 2018, www.ncbi.nlm.nih.gov/pmc/articles/PMC6347576/.

[2]“Impact of Algorithmic Bias on Society.” *Data Science W231 Behind the Data Humans and Values*, 11 Dec. 2018, blogs.ischool.berkeley.edu/w231/2018/12/11/impact-of-algorithmic-bias-on-society/.

[1] Murray, John. “Racist Data? Human Bias Is Infecting AI Development.” *Medium*, Towards Data Science, 8 May 2019, towardsdatascience.com/racist-data-human-bias-is-infecting-ai-development-8110c1ec50c.