AI 539 | READING - 2

Pushpak Katkhede | Date – 01/23/2023

---

1. This paper includes many examples of bias occurring in decisions/predictions made by machine learning systems (e.g., section 2 and section 3). (a) Choose and describe one example that you find especially problematic or worrisome. (b) What are the possible negative impacts?

Ans: Find the observations below:

A) Out of the presented examples the most problematic was probably the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) with its downstream application in the Judiciary where judges would use this system to judge to decide whether to release or hold them in the prison. The motive is to predict whether an offender will commit the crime again or may prove as a harmful aspect to society and based on the software results their decision is highly influenced. The data used to train was from previous arrests and friend/family arrests. Results were found to be biased toward the African American people to be accounted more for the possibility of recommitting the crime.

B) Majorly, the effect of this bias would be life-changing for the people under influence of these decisions. For instance, consider subject A1, African American, and A2, another Caucasian offender. Now, the hypothesis is that A1 has learned its lesson that it will not commit any crime again, but A2 is still in a problematic mindset and will reattempt any crime. Hence, the algorithm might still favor A2 based on the results due to this bias, and they will be released. However, even though A1 will not commit any crime again there is still a higher chance of them being retained in the prison. This small bias can basically change the trajectory of a person's future life. I believe these kinds of sensitive decisions are fuzzier than the datasets used to train the models for them. And, as an individual, it will be wrong to be treated for such a big decision based on third parties' behavior. But I also believe that upon handling the presented representational bias, population bias and other bias present the model could be more and more accurate though they should only constitute only a partial part of the decision-making process.

2. This paper lists many kinds of bias in section 3.1. (a) Pick one that is new to you (or particularly interests you) and describe (define) it here. (b) Imagine and describe an application area not mentioned in the paper where that kind of bias could happen.

Ans:

A) I have never come across an application of Aggregation Bias **Simpson's Paradox** in my work. The idea explained in the paper is that observing a bias for a bigger sample may visualize a bias towards a specific class but when the dataset is separated and analyzed over smaller logical data record groups. The bias may reduce in the individual observations and can even be reversed against the previously favored class. The example discussed is of the UC Berkley admission data where mostly women are overall less seen to be admitted to the university. However, on splitting, the analysis department wises it is seen that women

tend to apply to programs with lesser acceptance rates and hence are mostly not favored for admission by the model. Also, they are found to be reversing the trend against men in subjects like biology, psychology, astronomy, and computational social sciences.

B) One application area in my mind is about assessing two doctors' prowess overall in their field with the data of previous surgeries they have performed. Initially, assessing the total number of successful surgeries might show doctor D1 better than the other doctor D2. But if split as per the types of surgeries it is highly possible that for a specific type of surgery, D2 can also be better than D1. This can affect the model trained as they will always favor D1 for any kind of surgery even if it's the one where D2 is better than D1.

3. What is one question you have after finishing the reading? (What wasn't clear? What needs more investigation?)

Ans: My question is still about aggregation bias and how we deal with it. How should we make our model mode robust about these sub-specialties of the doctor so that the doctor should be correctly recommended as per the specific expertise?

4. **Reflect on your work for the course over the past week.** What did you do that was effective and increased your knowledge? What could you do or change to increase what you gain from this course? Is there anything about this course you are anxious/worried about?

Ans: I would say that I am really enjoying working with you on this course. As per part of Assignment 1, I did rigorous digging on available examples on the internet about handling missing values with different techniques. And observed the examples keenly and experimented with the problem dataset. This worked very well for me and I was able to solve the assignment to at least a satisfactory level for myself. However, I still lacked the initial interpretation of the topics and should try to clear doubts and understand the ambiguity of the problems beforehand. I am not anxious or worried about anything now though I have a few questions about the trajectory of the project and would like to discuss this in person.