Kyra Gray

Statistical Learning

Professor Kim

Due: 3/10/17

Podcast Response

Cathy O’Neil’s general argument outlines three components of a problematic algorithm and then gives real-world examples of how the consequences of these algorithms manifest themselves. She defines problematic algorithms as widespread, secretive, and destructive in someway. The first example she uses in the interview addresses the algorithmic models to predict recidivism risk (i.e. the risk of returning to jail after initial release). These models of recidivism in problematic, according to O’Neil, for a variety of reasons, starting with how the recidivism “scores” are calculated for each individual. The scores are calculated on based on two sets of data: first police interaction data which is inherently biased considering the areas in which interaction with police is more likely to occur and second survey data that has questions with proxies for race and class, without explicitly asking for those two identifiers. Then the recidivism score is given to judges during sentencing and often a high recidivism score can result in a longer sentencing, thus creating a self-fulfilling prophecy or a feedback loop. If high recidivism scores create longer sentences, then the individual becomes more isolated from their communities and therefore once released from jail more like to commit a crime again, validating the initial score.

The next example that O’Neil explores is a theoretical one of how tech companies hire engineers. It would make sense for the company to be motivated to use a machine-learning algorithm given the implicit costs of sifting through a large amount of applicant resumes and weeding out the ones who will ultimately be successful. In order to do this, the company would have to choose an algorithm, a data set, and define what “success” at the firm looks like. An obvious data set to choose would be current employees at the company and O’Neil propose that they define success as getting promoted at least twice in the employee’s first three years at the company. Now, they run the model and realize that no women get through the filter (perhaps a seemingly extreme example but not unheard of the world of tech), what would that mean? That women make bad engineers? The issue with the model here is it would not tell the company to check their culture and propose making it more welcoming to women; the algorithm doesn’t understand why female engineers don't have success at the company and therefore cannot make moral or ethical decisions, they can only pick up patterns that already exists. These patterns embody our historical practices, however only when our historical practices are perfect will the algorithm expose the truth. The tech company hiring process example illustrates model or data justified high stake decision making, without an actual check on fairness or meaningfulness.

The third example of O’Neill’s argument addresses the old myth of the American public schools that we can fix the educational system by just getting rid of bad teachers. There have been two generations of mechanism to assess teacher performance. The first generation of this assessment scoring was crude and flawed, was based on simply counting the number of students in each teacher’s class who were proficient in each given subject. However, it is important to realize that performance on standardized tests is high related to poverty, so this teacher assessment method was punishing the teacher of poor students and was not discerning enough at finding the actual “bad teachers.” One of the main issues with this initial assessment method was that it didn't control for the initial or incoming quality of the student at the beginning of the year. To account for this control, a derivative model was developed, based on a background model that estimated what an individual student should get on their standardized tests by the end of fourth grade given their performance at the end of third grade, and a couple other predictor variables. And thus the teacher’s score is just the collection of the difference between how each student actually performed versus how they were expected to perform. So essentially, teachers were being help accountable for the error term in already bad model that already has a lot of noise. This illustrates another example of a destructive model because firing decisions are made over this teacher assessment score, but these scores are simply not accurate enough to be the justification for that decision. O’Neil proposes a scoring system with feedback embedded into it, so that teacher will continue to get scored but they can then in turn get feedback on their score, feedback that a good teacher can reliably use to improve their score.