Nathan Beneke

Machine Learning in Society: Who Trains Whom?

May 15th, 2019

Reducing Bias in Humans and Machines

We often consider the dramatic consequences of bias being present in machine learning models. A facet of this conversation that is often missing, however, is that the consequences of bias being present in humans can be just as dramatic and that human bias is almost always the cause of bias in machine learning. Whether the data is biased because those that collected or produced are, the developers are biased, or simply existing societal divides that are the result of bias and prejudice make fair data collection difficult it is clear that the biggest challenge to making unbiased machines is correcting for human biases. With this in mind, it is plain to see that bias in machine learning models is strong evidence that the same bias is present in either the developers of the machine or society at large. Furthermore, machine learning can be used to identify and measure systemic bias in areas that would otherwise be cost-prohibitive to properly investigate. To see this, we first define bias in a human context and discuss how to identify and mitigate it. Second, we see an example of using Natural Language Processing (NLP) to detect and quantify gender bias in the courts of the Pacific Islands. Lastly we define bias in a statistical context and investigate methods to detect bias in machine learning models.

From a human perspective, bias is a very intuitive concept and individual bias is easily identifiable, as well as treatable, without the use of machine learning. Merriam-Webster defines bias as “an inclination of temperament or outlook, especially a personal and sometimes unreasoned judgement.” From this definition it is clear that humans need bias to operate and make decisions in our daily lives. Without bias, one would be forced to think in depth about each and every decision, this is simply infeasible so we rely on our biases to make quick judgement calls. Because we need many of our biases, the rest of this paper will focus on those biases that most consider to be systemically and morally problematic such as racism and sexism. With these biases being such a large and well known problem in society, it is somewhat surprising that there are simple and effective tests to detect them in individuals. The Implicit-Association Test (IAT) was developed in 1998 and is widely used today in research, therapy, and corporate contexts. Although it has its limitations and there are some doubts about its accuracy, most scholars agree that the IAT or similar tests are effective at identifying unconscious bias in humans across broad different categories. Many of these tests can be taken for free online and cover a broad range of categories such as race, gender, weight, and sexuality (Project Implicit). Because there is such a straightforward way to test bias in individuals, machine learning is simply unnecessary for the task.

Methods of mitigating individual unconscious bias are less effective than methods to detect it, however there are a number of options. Many organizations such as Facebook, Google, and the US Department of Justice have some of their members or employees undergo training to reduce their unconscious bias. One of the most common methods is *counterstereotyping*, being exposed to examples of individuals who go directly counter to some stereotype. For example, to counterstereotype gender bias one can read stories, or write essays about powerful women. It has been shown that real life examples, such as having women as professors, are particularly valuable for this (Dasgupta). Another method is *negation*, actively rejecting stereotypes when they are encountered. One particular method is to have participants press a button labelled “no” when exposed to an example of something conforming to a stereotype, and “yes” when exposed to a counterstereotypical example. The efficacy of this method is under dispute by scholars, with some believing the only positive effects come from its intersection with counterstereotyping. A third method is *perspective-taking*, or intentional and specific empathy. Typically this takes the form of participants being familiarized with a particular example of the effects of bias on an individual minority or a group and then being lead through exercises to empathize with the affected person or group. There is also some recognition for more traditional techniques being effective as well, such as loving-kindness meditation--a technique rooted in many religions such as Buddhism, Hinduism, and Jainism. (Unconscious Bias Training)

With all of these well developed techniques for detecting and reducing bias in humans, it seems there is no particular use for machine learning in this regard for individuals. If machine learning can be usefully applied to detecting human biases, it is only in systemic bias. Even in that area, there are simple statistical means to detect bias on large scales when data is available. It is when data is not available that machine learning can be useful in this way. HURIDOCS helped ICAAD--both international human rights organizations--collect and analyze quantified data about gender bias in the judicial systems of the Pacific Islands using machine learning. They used natural language processing to identify 908 sentencing procedures for crimes related to domestic violence, sexual assault, and gender-motivated murder from a larger pool of publicly available transcripts of court proceedings. The training data for this was a set of labelled court proceedings from an earlier ICAAD study. They then used another model to extract data from the transcripts, such as the ages of the perpetrator and victim, sentence length, and specifics about the case. Finally they used this data to estimate the effect of gender biases on the length of the final sentence. (Applying Machine Learning to Detect Judicial Bias in the Pacific Islands) They concluded that on average gender bias reduced sentencing lengths for domestic violence by 60%, sexual assault by 40%, and gender-motivated murder by 18%. They study is even able to break down these number further into more specific gender bias categories. (Analysis of Sentencing Practices in SGBV Cases in the PICs)

Without machine learning it would have been cost-prohibitive to collect and analyze this data, suggesting that machine learning can be vital to detecting and fighting systemic bias. It is worth noting that in this example, the judges were not attempting to hide their biases, their problematic views are expressed openly in transcripts. More data has to be collected to tell if a similar approach could be used in a system were bias is taboo and attempts are made to appear impartial--such as racism in the United States' judicial system. One factor that is also potentially problematic is the identification of cases. Because no machine learning model is perfect, it is certain that there are some cases the model falsely flagged as not having to do with gender violence. If there is some characteristic these cases share, that characteristic is completely left out of the analyzed data. Perhaps worse, there is no cost-effective way to know whether that has happened or not, and if so in what way. Even with these potential downsides, approaches like this show great promise and utility for human rights groups worldwide.

Detecting bias in machine learning models is, in many ways, much more straightforward than in humans. Importantly, bias in machine learning models can be much more formally defined than it can for humans. We use a statistical definition for this purpose. Here *bias* is the deviation of the expected value of a statistical estimate from the quantity it estimates. For example, if one estimated the mean age of a country's population by sampling the ages of residents in a nursing home, the estimator would be drastically biased towards a high age. Another important thing to note is that there is an inherent trade-off between bias and variance, that is the less biased a machine learning model is the less accurate it will be. (Brownlee) A completely unbiased model would not only be terribly inaccurate, but would also display exactly the biases its training data contains. Consider that the very biases we want to reduce are introduced into data sets by systemic biases, and that because of this an unbiased estimator for these data sets is ironically "biased" in the human sense of the word. As a more concrete example, consider a data set of résumés for software engineers that work at some specific company and a machine learning model to classify people the company should interview, based on their résumé. Suppose this company has a toxic work environment for women, and because women are underrepresented in computer science they have very few data points for women to begin with, that causes women to quit or underperform more often than the men in similar roles. An "unbiased" machine learning model would then correctly correlate being a woman with being likely to quit shortly after hire or underperform. As illustrated by this example, a statistically unbiased machine learning model is clearly undesirable. Instead, in order to create a fair and useful machine learning model we want a model which is as statistically unbiased as possible, while minimizing the correlation between factors such as gender or race and the outcome of any particular input.

In order to minimize these aforementioned correlations, one must first identify what factors are correlated with the outcome. There are a number of methods and tools to accomplish this, we focus on an open source tool called FairML, which allows the user to audit “black-box” classification machine learning models for highly correlated data features. FairML is a library for the programming language Python that integrates with other popular libraries, Pandas and NumPy. Together these libraries provide a framework for developing and training machine learning models that provides many tools which make the process of developing a machine learning model less complicated. A machine learning model built using these tools is output as a single function which can then be trained, tested, and deployed using any arbitrary dataset that is properly formatted. This model is provided to FairML, along with a set of data to train and test on, which then outputs a list of data features along with how correlated they are to the classifier. The precise mathematical details are beyond the scope of this paper, but an overview follows. FairML uses four different ranking algorithms to determine the highly correlated features: Iterative Orthogonal Feature Projection Algorithm (IOFP); minimum Redundancy, Maximum Relevance Feature Selection (mRMR); LASSO; and Random Forest Feature Selection. IOFP is the most highly weighted algorithm for the final ranking, and unfortunately very linear algebra intensive. The basic procedure for IOFP is:

1) Train the model as normal.
2) Use normal statistical methods to determine the most highly correlated feature, without regards to accuracy in the model’s classification.
3) Orthogonally transform the rest of the features into the previously identified feature. Essentially, this replaces the most highly correlated feature of the dataset with some summary of every other feature.

4) Retrain the model using this modified dataset, and compare the differences in classification.
5) Continue doing this until all features have been replaced.
6) The most correlated features are the ones with the most difference in classification outcome before and after the feature was replaced.

Tools such as FairML are highly useful for people to independently audit proprietary and closed-source machine learning models, as they do not require that the auditor has access to the source code for the model. Of course to be most accurate, FairML needs to be run with the actual data being used for the model, and the data is also often proprietary and closed-source. (Adebayoj)

In conclusion, bias in machine learning models reflect the biases of our society and a machine learning model being biased is strong evidence of the same bias existing in the population training data was taken from, or in those collecting said data. While machine learning is not useful to detect bias in individuals, it can be used as a useful tool to collect and analyze data about biases which would otherwise be difficult or impossible to come by. Similarly to detecting bias in humans, it is possible to detect bias in machine learning models without exposing the models source code or methods. While it is clear that there is no silver bullet to address bias in either machines or humans, by utilizing both detection and correction techniques humans and machines can work together to dramatically reduce bias in each other.

References

Adebayoj, Julius. "Adebayoj/Fairml." *GitHub*, Massachusetts Institute of Technology, 23 Mar. 2017, github.com/adebayoj/fairml.

"Analysis of Sentencing Practices in SGBV Cases in the PICs." *ICAAD*, 6 Apr. 2017

"Bias." *Merriam-Webster*, Merriam-Webster, www.merriam-webster.com/dictionary/bias.

Brownlee, Jason. “Gentle Introduction to the Bias-Variance Trade-Off in Machine Learning.” *Machine Learning Mastery*, 2 Feb. 2017, machinelearningmastery.com/gentle-introduction-to-the-bias-variance-trade-off-in-machine-learning/.

Dasgupta, Nilanjana, and Shaki Asgari. “Seeing Is Believing: Exposure to Counterstereotypic Women Leaders and Its Effect on the Malleability of Automatic Gender Stereotyping.” *Journal of Experimental Social Psychology*, Academic Press, 8 Apr. 2004.Guilhem, Manushak.

“Applying Machine Learning to Detect Judicial Bias in the Pacific Islands | HURIDOCS.” *Human Rights Information and Documentation Systems, International*, 2017, www.huridocs.org/2017/03/applying-machine-learning-to-find-judicial-bias-in-the-pacific-islands/.

“Project Implicit.” *About Us*, implicit.harvard.edu/implicit/aboutus.html.

“Unconscious Bias Training.” *Wikipedia*, Wikimedia Foundation, 8 May 2019, en.wikipedia.org/wiki/Unconscious_bias_training.