Rational, not Unfair : Incentives for Fair Machine Learning

Naman Goel

Machine learning is often criticized for being discriminatory towards different groups of our society or even individuals. The example scenarios range from simple ones like granting or denying bank loans to more complex ones such as racial/gender discrimination in predictive policing.

We know that humans make decisions based on their prior biases and the observations they make. For e.g., a bank may want to maximize its expected profit from loans and it is rational in using its “knowledge” to achieve this goal. Similarly, police is rational in using past experience to minimize the expected loss due to crimes. Such rational decision making can lead to unfair treatment with the involved people. While we anticipate more fairness from machines, the reality is that machine learning works based on the data it is fed and the algorithms that are written to achieve human defined goals such as empirical risk minimization. These goals don’t include fairness as part of their design. It can also be called a “continuous loop of injustice” because decisions taken by a ML system also affect the sample of data that it will see in future. Given the scale at which machine learning is (or will be) used to make decisions and the potential impact it may have on our society in long term, the issue needs attention.

While making machine learning fair is an open and complex problem with many ongoing research efforts [4] to tackle different theoretical and practical challenges, the goal of this project is to consider a simple scenario such as credit scoring and use a representative machine learning classification algorithm such as logistic regression to understand how a machine learning algorithm, that is rational for the decision maker, can also be made fair to the society ? In the first part of the project, student will understand the discrimination problem and find evidence that the problem exists not only in theory but also in practice through simulations as described in [1, 2, 3] on real datasets. The second part of the project will be focussed on developing new ideas for modifying the algorithms so that they are “fair” and measure how much loss (for e.g. misclassification on test data) does a decision maker suffer because of being fair and how much “fairer” the decisions become. In the last part, possibly provide justification for this loss by suggesting that there is a balancing incentive for being fair.
Required Skills : basic understanding of machine learning.
Preferred : Working knowledge of Python.
Duration : 3-4 months

References:

1. http://research.google.com/bigpicture/attacking-discrimination-in-ml/
2. http://blog.mrtz.org/2016/09/06/approaching-fairness.html
3. https://arxiv.org/abs/1610.02413 ,NIPS 2016
4. http://www.fatml.org/schedule/2016

Popular Media Items:

1. http://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html?_r=0
2. https://techcrunch.com/2015/08/02/machine-learning-and-human-bias-an-uneasy-pair/
3. http://www.sciencemag.org/news/2016/09/can-predictive-policing-prevent-crime-it-happens
4. http://www.law.nyu.edu/bernstein-institute/conference-2016
5. https://www.wired.com/insights/2013/08/predictive-policing-using-machine-learning-to-detect-patterns-of-crime/
6. https://research.googleblog.com/2016/10/equality-of-opportunity-in-machine.html