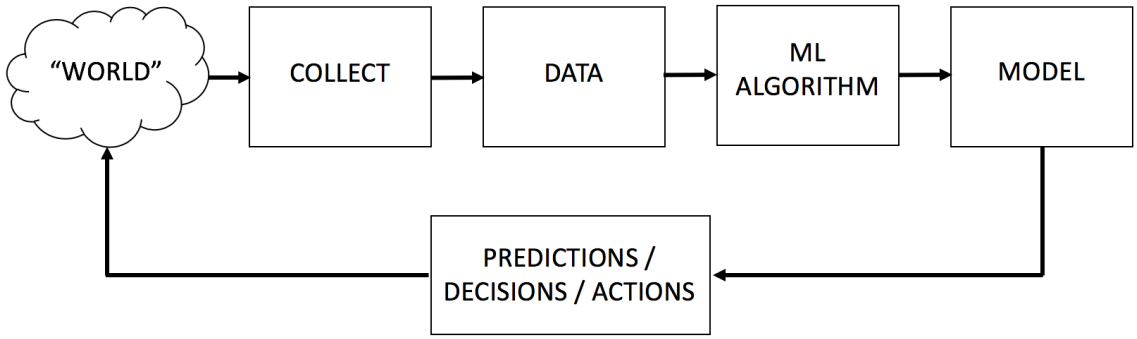
CIS399: The Science of Data Ethics	Lecture 7: Feb 7, 2019
Instructor: Michael Kearns & Ani Nenkova	Scribes: Ananya Chandra
-	and Allie Grey

1 Brief Recap

We spent the first few lectures talking about the basics of Machine Learning and the mathematical foundations that underlie basic models.



We discussed the math required to fit a model to specific data, and how this applies to the real world. We needed some probabilistic foundations to start the course. Then, we talked about the paper on sexual orientation detection which touches on the issue of machine learning applications in society. How do we feel about this ethically? Are there privacy violations involved? Are there methodological issues involved? There are various issues related to different steps of the machine learning process from data collection to data interpretation and model development. It is important to critique and analyze the studies and models that are there.

For the next two lectures, we will become a little bit more precise, and talk about fairness in machine learning. Within some limited part of this pipeline, it is possible to be precise in what we mean about an algorithm being unfair to a particular part of society. The headline story: it is not just possible to give quantitative notions of fairness/unfairness - there are a lot of definitions you can give. If you want one type of fairness, you may have to explicitly give up another type of fairness.

We will mainly be focusing on the part of the pipeline where we already have data (ML algorithm, model, and predictions/decisions/actions). We will be looking at this as an x/y pair, where x is what we know about individuals and y is what we are trying to predict about those individuals.

2 What is Fairness?

Discussion: What does it mean for a model to be fair?

- We can ask about the behavior of the model?
- What should the model depend on? (the x value)?

Some ideas:

- The model should behave similarly for similar individuals.

- The model should have the same number of false positive decisions across various racial/gender groups. (But is there more harm caused by false positives or false negatives? It depends on the situation.)

- The model should perform similarly across similar demographics.

- Difference between mutable and immutable factors. Do factors like income, credit, etc. come from circumstance or from choice?

- Do we take history into account? Not just who you are now, but how you got there. (Example: someone who had a lot of opportunities vs. someone who did not, but both end up in the same place).

Is the purpose of these protections to protect them now, to right past injustices, or something else? Think about a level playing field. All things being equal, we want all people with similar abilities should have an equal outcome in society. But this is not guaranteed.

There are thornier questions that are maybe more difficult to answer that we wanted to at least bring up.

2.1 Group Fairness Notions for Classification

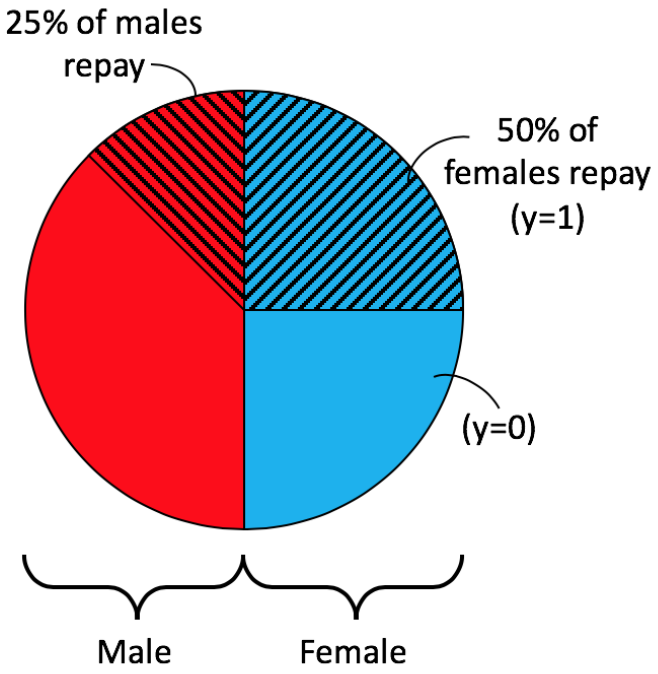
Recall our framework: where sample S = $\{< \bar{x_1}, y_1 >, ..., < \bar{x_m}, y_m >\}$. We choose $h \in H$ to minimize error on S. In this lecture we will assume that y is some discrete value (ie. $y \in 0, 1$ or $y \in < -1, +1 >$), which represents decisions such as "give or don't give a loan" or "admit or don't admit a student." Right now, we're not making a distinction between these students, but now we want to introduce the idea that different people in our sample may be from different demographic groups. Now imagine that each individual $\bar{x}$ may belong to various groups (ex: race, gender, age, disability status, sexual orientation). These definitions require us to decide, in advance, what discrimination we are worried about. We can worry about multiple, but we still need to state up front what kind of discrimination we are trying to protect against.

Example: for each individual, $\bar{x}$, we can have a gender, $\bar{x} \in \{$female, male$\}$. We cannot, from the start, indicate that this vector necessarily contains your group type.

Statistical Parity: The fundamental question: is a given $h \in H$ "fair" in P (our world) and gender. Let's suppose that when our model $h(\bar{x}) = 1$ predicts repayment, and when $y = 1$, this indicates actual repayment. Statistical parity generally says if you have a limited amount of something to give away, you have to distribute your resource equally across different groups. This can be written as: statistical parity would demand that $\Pr[h(\bar{x}) = 1 \mid gender(\bar{x}) = male] = \Pr[h(\bar{x}) = 1 \mid gender(\bar{x}) = \text{female}]$.

Running with this definition, demanding that if you give loans to 27 percent of men and 27 percent of women. But you can change these conditions to allow there to be a possibility of relaxation. This possibility of relaxation is called γ. When $\gamma = 0.05$, you are asking that the difference is less than 0.05. By having γ, you have a trade-off between statistical parity and predictive accuracy. There is always going to be a trade-off between fairness and accuracy.

We are picking our models from some class that we have committed to. But what if there is no model in this class, there is no way to get to 0 fairness, or there is no trivial way to achieve the definition? We want to assume that our model has something that allows us to meet some definition of fairness - even if it is not the most accurate.



2.2 Incorporating "Merit"

If there are certain people who are more likely to get to the outcome we are trying to achieve, we can consider this 'merit'. The first part of this is equality of error: don't distribute loans ($h(\bar{x}) = 1$) evenly, but distribute mistakes evenly. Pr[$h(\bar{x}) \neq y \mid gender(\bar{x}) = male$] - Pr[$h(\bar{x}) \neq y \mid gender(\bar{x})$ = female] $\leq \gamma$.

The overall error: $E(h) = \Pr[male]E(h \mid male) + \Pr[female]E(h \mid female)$. Minimizing our overall error (E(h)), may not make our conditional errors $E(h \mid male) = E(h \mid female)$. Why? The following are reasons for this:

1) Group imbalance, where one group is just a larger fraction of the population of the other group. (Example: if the world was 90 percent male and 10 percent female). Then, the distribution would be skewed towards the majority population.

2) We cannot predict one class or group as well as another.

For the equality of false negatives, we change our $y = 1$ and $h(\bar{x}) = 0$. For the equality of false positives, $y = 0$ and $h(\bar{x}) = 1$.