Real World Mini-Project 3: Fair ML

Version 2: Some folks in the UW community believed that the initial version of this assignment might be advocating for unfair machine learning. It was not. We have reworded a few sentences and added this explanation to make it unambiguous that the intent is to think about designing fair systems, and address historic inequity, not continue it.

One of the lessons of the assignment – that there are many definitions researchers have proposed for the meaning of "fair" and that what what is fair in one instance may be very unfair in another – is only apparent after successfully doing the assignment, so we have added that goal here to be explicit about the intended lesson as well.

These clarifications were made after the first due date for the assignment (before the late date) – for students who wish to see the version they were using when finishing the assignment, the old version is here.

Due Date: This assignment is due at Tuesday June 1 at 11:59 PM (Seattle time, i.e. GMT-7).
You will submit as a PDF to gradescope.

One of the most popular (and influential) applications of the tools of this course to the real world has been in machine learning. You have probably (hopefully!) heard that ML applications have had unintended effects in the real world, especially around fairness. In this assignment, we will use the tools of probability to get a taste of what fairness analysis looks like. Our specific goals are:

(a) See that probability and conditioning are useful formal representations of what it might mean to be fair.

(b) Realize that "fair" has more than one reasonable definition, and you can't necessarily have all of them!

(c) Get a sense that fairness is complicated!

This will **not** be a full introduction to fairness in ML; the topic could easiy fill an entire course. This assignment is intended to help you realize that the topic is important, requires careful attention to understand, and that it is a possible application of the tools you learned in this course.

The basic operation of machine learning is to take in a large set of data, find patterns in that data (using statistics), and make predictions about future data based on the given data. Bias can creep in at any part of that process.

1. Three Definitions of Fairness

ML Fairness is usually discussed in the context of disparate outcomes for historically underrepresented and marginalized groups (e.g. based on race and gender). To make the examples more concrete, imagine that our ML application is for a bank deciding on making loans. Our ML system will look at historical data of who repaid loans, and use that to predict whether a new person will repay their loan. Historical practices like redlining have long-lasting effects, which mean that any historical data would show differences in repayment rates across races – these differences would be a result of the historical practices (not the reliability of the applicants), so to use the data we must compensate for this misleading aspect of the data.

Suppose we choose a loan applicant at random: Let L be the indicator that our ML system says our applicant should get a loan, and let A be the the indicator that the applicant has the attribute we wish to ensure we treat fairly (in the loan context, the standard example due to longstanding inequities in banking is the attribute of race; A will therefore indicate whether the applicant is Black), and Y be the indicator that the person would truly successfully repay a loan.[1]

A "fairness criterion" is a possible definition of what constitutes "being fair." For example, "precision parity" is defined as $\mathbb{P}(Y = 1|L = 1, A = 1) = \mathbb{P}(Y = 1|L = 1, A = 0)$. That is the probability that the person will repay the loan, among people who our ML system will give loans, is equal regardless of whether the applicant has our protected attribute or not. Intuitively, the goal of this definition is to ensure that the ML system does not give more "benefit of the doubt" to one group than another (by ensuring that both groups have the same fraction of repaid loans).

[1] To make it possible to know Y (without using the system to decide on real loans before knowing whether it works) ML systems are usually evaluated on a "test set." A dataset gathered in advance where the true answers (the Y's) are known, but the data was not used to design the model.

(a) Another possible fairness criterion is "true positive parity" $\mathbb{P}(L = 1|Y = 1, A = 1) = \mathbb{P}(L = 1|Y = 1, A = 0)$. Give a (1-2 sentence) intuitive description of what this criterion is trying to achieve.

(b) A third possible fairness criterion is "false positive parity." This definition seeks to ensure that among those who do not repay their loan, the probability of our system wanting to give them the loan is equal whether they have the protected attribute or not. Intuitively, the goal is not to give more undeserved benefits (loans unlikely to be repaid) to people without the protected attribute. Note that as the definition is an equality, it also requires that we do not give more benefits to the group with the protected attribute either. Give the mathematical notation for false positive parity.

2. You Can't Have it All

A theorem[2] says that an ML system cannot achieve all of precision parity, true positive parity, and false positive parity unless the ML system never makes a misake (which never happens) or $\mathbb{P}(Y = 1|A = 0) = \mathbb{P}(Y = 1|A = 1)$, that is the actual repayment rate of loans is equal regardless of the protected attribute (which, since we're paying attention to the attribute because of historic disparate treatment, won't happen).

(a) For the loan context, suppose you could make the system meet exactly one of the three fairness criteria above. Which of the three definitions of fairness would you want to ensure your ML system meets, and why (any of the three can be chosen, but we want your answer to show an understanding of the definition and relate to the real-world) [3-4 sentences.]

(b) Choose one of the other fairness criteria (not the one you chose in the last part). Think of a scenario where you think that idea of fairness would be the most appropriate of the three. Explain why in that scenario, your choice is a good one. [3-4 sentences]

3. In the news

Find an news article about a paritcular instance of bias in a machine learning (or articifical intelligence) system. We expect the questions in this section can be answered in 1-2 sentences each.

(a) Tell us the title of the article, the source (e.g., "New York Times") and give us a link to it (don't worry about pretty formatting)

(b) What is the ML application trying to learn (e.g. "housing prices" or "recidivism rates")? What group does the article suggest is being treated unfairly?

(c) An ML application can be (roughly) grouped into 3 steps

(i) Data acquisition – collecting the data that will be the basis of the system

(ii) Model selection – choosing which "features" of the data to use in the system and how those data points will be combined (e.g. for a house pricing model, one might decide whether or not to include the number of rooms, square-footage, and time that it has been on the market without selling as 'features' to predict the selling price)

(iii) Use of prediction – once the model is trained, it will be used to make predictions (the model thinks this house that just went on the market will be worth Y dollars), those predictions will then be used in the real-world (e.g. a company might decide whether to invest in a property as a result of comparing the model's estimate of its value to its current price)

Does the article you read try to place the blame on any of these steps? If so which one(s)?

[2] proven in the last 5 years!

(d) Does the article state what the author thinks "fairness" would be? If not, is it implied? The definition might match the definitions in this assignment or be something different (there are other fairness criteria we havne't covered).

(e) For this application, what definition of fairness do you think is appropriate (this could be one of the three definitions above, or it could be another definition). Your definition here can be phrased intuitively or in the language of probability (you don't have to list both).