Week 8 Notes

Joseph Han, Wesley Lo, Sammy Grossman

Enhancing Fairness in ML

- What are some reasons of unfairness in ML ?
- How do we decide what fairness means?
- How do we incorporate these notions of fairness in ML algorithms?

ML is Unfair because Data is Unfair

- Earlier we used to think that data is fair
- It isn't - often times, data are biased
    - I disagree with the notion that 'big data" itself is unfair, in fact i don't think it's even accurate to label it as fair or unfair. It simply is big data, or the collection of massive amounts of information from our environment. What is unfair though, is the way machine learning models are using this information, which has our own human biases woven into it, to make decisions that are people without a conscience is making the decision.
    - There is nothing intrinsically unfair about data itself. But you data collection procedure might be unfair towards certain parts of the population. But that does not make the other collected part of data unfair.
- Bias in ML system caused by bias in training data
    - Source from which data is coming is biased
- Why?
    - ML labels are often given by biased human
    - Data often collected by humans in a secretive non-transparent manner

Example: stop and frisk program

- 4.4 million people stopped and frisked in NY from 2004 to 2012
- 83% firsked were African American or Hispanic
    - Wrong Interpretation: African American and Hispanic are just more likely to commit crime
    - Correct Interpretation: Decades of systemic racial discrimination has led to this bias
- End result: system found illegal in courts as a form of racial profiling
- Police officers has biases because of which they mislabeled or provide biases labels to data points (which were people)

Hurricane Sandy Tweet

- Data used was not good for the task
- Task: figure out which parts of New York are worst affected by the hurricane
- But the data showed completely different patterns
- Manhattan had most number of tweets-least affected
- Coney Island most affected - least number of tweets

StreetBump smartphone app

- Accelerometer data and GPS data from your phones to collect info on where are the potholes in cities

Student Example

- I actually had the chance to read and learn about racial discrimination in recruiting technology, in another class as well. I recall that certain companies would require a video to be submitted, where the interviewee verbally responds to prerecorded questions. The video would be put through a software that gave the interviewee a score based off of mannerisms and facial expressions. A particular interviewee was from China. He was extremely overqualified for the positions he was applying for, but he received extremely low scores on these video tests. The software was supposed to associate certain mannerisms and facial expressions with positive work qualities, but would give poor scores to certain races because of physical differences. It was sad to see big name companies in America using a technology that so blatantly discriminated against certain races.

ML is Unfair because Data is Unfair?

- Is this problem going to go away as we improve our penetration of smartphones in different cities and among different sections of society.
- NO!
- Technology is always differentially adopted
- Things that are being used to collect data today might not be used tomorrow, we might have different kinds of devices and then we would need to ensure equity in usage for that device
- With every dataset, important to ask the following questions:
    - Which people are excluded?
    - Which places are less visible?
    - What happens to people who are excluded (places which are less visible)?

Ways in which Data might be Unfair

- The data can reflect social biases of people who collect data
    - Stop and frisk
    - Predictive Policing (PredPol)

- The data might not reflect any social biases, but might not represent different categories differently
    - StreetBump is an example of different categories not being represented equally
    - Uncertainty bias
    - Where else did we see this?
        - Interpretability paper
        - The explanation that you give for the majority class might be completely different from the minority class

- Data might not reflect cultural differences sufficiently
- Example: classifier to degregate fake names from genuine names
- Step 1: Look at names from American website - FaceBook
- Step 2: Train your model on these names
- Step 3: what happens when you get indonesian names? Does the model still do as well
- What is happening?
    - American names: Walters, White, Jones, etc.
    - Thailand names? The law does not allow one to create any surname that is duplicated with any existing surnames
    - Under Thai law, only one family can create any given surname: any two people of the same surname must be related, and it is very rare for two people to share the same full name. 81% of family names were unique
- Key lesson: statistical patterns that apply to the majority be invalid within a minority group
    - Variable positivity correlated with target in the general population might be negatively correlated in minority population
    - Length of a name = variable is inversely correlated with genuineness of last name for American names
    - Is positively correlated when you look at Thai names

- Data might lead to undesirably complex models
- One idea
    - Maybe you train a different model for different
    - By luck, you get simpler models
    - How to combine these models is an open question

Be wary of Accuracy Estimate

- Accuracy estimates may not be able to give you a good picture about fairness
- What does 5% error actually mean?
- 100 samples
    - 90 majority sub-population (American names)
    - 10 minority sub-population (Thai names)
- Your classifier:
    - 100% accuracy on majority class
    - 50% accuracy on minority class
    - Net accuracy=95% (or 5% error)
    - But this is an unfair classifier

What is Fairness?

- Is it the same as unbiasedness?
- Some definitions of bias
    - Definition from legal community
        - Judgements based on preconceived notions and beliefs
    - Definition from statistics community
        - Systematic difference between sample and population

        - Definition from ML community
            - Bias variance tradeoff
                - High bias low variance - underfit
                - Low bias high variance - overfit

False Positives or False Negatives

- COMPAS
    - Predicts whether someone is high-risk or low-risk to commit crime
    - False positive - someone who was unfairly labeled as high risk and has to suffer incarceration (predominantly on the basis of race)
    - False negative
- AURA
    - An algorithmic tool used in Los Angeles to help identify victims of child abuse
    - False positive

Allocative Harms and Representational Harms

- Majority of literature understands Bisas as producing harms of allocation
- Allocative harm is when a system allocates or withholds certain groups an opportunity or resources
    - Who gets a mortgage, who gets a loan (if these decisions are biased, then this is an allocative harm)?
- Representational harms - systems reinforce the subordination of some groups along the lines of identity
    - Can happen regardless of whether resources are being withheld from certain groups
- Allocative Harm - More immediate harm - More immediate harm, readily quantifiable, transactional
- Representational Harm - More long-term harm, difficult to formalize, cultural
- Most work deals with allocation harm as it is an easier problem to solve technically, even though representational harm is at the root of all issues

Example of Representational Harm

- Example: Latanya Sweeney study
- Names associated with African Americans were yielding ads related to criminal background checks
- Employers doing searches on job applicants will see these results which may then lead to racial stereotypes and discrimination in hiring
    - Creation of this stereotype is also not what you want, even though there is no loss of jobs that African Americans face
- Stereotyping
- Man is to Computer Programming as Woman is to Homemaker? Debiasing Word Embeddings
- Denigration-using culturally inappropriate term
- Under-representation google image search for CEO
- Recognition – does a system recognize a face? Does it recognize respect, dignity?

- Facial recognition software cannot process darker skin tones
- Nikon's camera mischaracterized Asian features

Allocative Harm

- Scrub to neutral - remove the biased data
    - Who gets to decide what terms to remove? And why those in particular?
    - Whose idea of neutrality is at work? Is neutral what we have in the world? That might not be good
- Should you use demographic distributions?
    - Less than 8% women are CEOs...should we use this?
    - Women find it difficult to get into C-suite, how do you account for this and make the search results fairer?

Final Word of Advice

- ML scientist should take a page from social scientists, who have a long history of asking where the data they are working with comes from, what methods were used to gather and analyse it, and what cognitive biases they might bring to its interpretation

What is the COMPAS system

- Correctional offender management profiling for alternative sanction
- A case management and decision support tool
    - Developed by Northpointe (Now Equivalent)
    - Proprietary software
- Used by US courts to assess the likelihood of a defendant becoming a recidivist
    - recidivist - person repeating an undesirable behavior after they have either experienced negative
- Northpointe created risk scales for general and violent recidivism, and for pretrial misconduct
- Pretrial release risk scale
    - Measure of the potential for an individual to fall to appear and to commit new felonies while on release
        - Current charges
        - Pending charges
        - Prior arrest history
        - Previous pretrial failure
        - Residential stability
        - Employment status
        - Community ties
        - Substance abuse
- General Recidivism scale
    - Designed to predict new offenses upon release, and after the COMPAS assessment is given
        - Criminal history and associates
        - Drug involvement

        - Indications of juvenile delinquency
- Violent recidivism scale
    - Designed to predict new violent offenses upon release, and after the COMPAS assessment is given
        - History of violence
        - History of non-compliance
        - vocational /educational problem
        - The person's age at intake
        - The person's age at first arrest

- Violent recidivism risk score
    - = - (age * -w)
    - + (age-at-arrest * -w)
    - + (history of violence * w)
    - + (vocation education * w)
    - + (history of noncompliance * w)
    - *w* is a parameter that they learn from data

Why is COMPAS needed?

- The US locks up far more people than any other country, a disproportionate number of them African American.
- For more than two centuries, the key decisions in the legal process, from pretrial release to sentencing to parole, have been in the hands of human beings guided by their instincts and personal bias

- If computers could accurately predict which defendants were likely to commit new crimes, the criminal justice system could be fairer and more selective about who is incarcerated and for how long.

- If it's wrong in one direction, a dangerous criminal could go free. If it's wrong in another direction, it could result in someone unfairly receiving a harsher sentence or waiting longer for parole than is appropriate.

- Proponents of risk scores argue they can be used to reduce the rate of incarceration

How is the COMPAS system used?

- Rating a defendant's risk of future crime is often done in conjunction with an evaluation of a defendant's rehabilitation
- The Justice Department encourages the use of such combined assessments at every stage of the criminal justice process
- Reform bill currently pending in Congress would mandate the use of such assessment in federal prison

Brisha Bordon VS Vernon prater

- One possible conclusion: COMPAS is racially to be biased

- What is flawed in the argument in the initial part of their analysis?
    - Other factors might be there
        - Either in the dataset which might play a predictive
        - Or outside the dataset (which are confounding or hidden variables)
    - Machine error
    - Confirmation bias
    - Very small sample size x2
    - Examples show people who had race=AA, who were low risk but were predicted high risk (false positive)
    - People who had race = white, who were low risk but were predicted high risk (false positive)
        - False positive rate across races should have been measured
    - We can only say this confidently if system is able to prove causation
        - Standard ML models only establish correlation

What did Propublica do?

- Obtained the risk scores assigned to more than 7,000 people arrested in Broward County, Florida in 2013 and 2014
- Checked to see how many were charged with new crimes over the next two years
- Compared the recidivism risk categories predicted by the COMPAS tool to the actual recidivism rates of defendants in the two years after they scored

What did Probulica find?

- COMPAS score correctly predicted an offender's recidivism 61 percent of the time
- COMPAS score correct in its predictions of violent recidivism 20 percent of the time

What did Probulica find-Point 1

- African American defendants were often predicted to be at a higher risk of recidivism than they actually were
- Analysis found that AA defendants who did not recidivate over a two-year period were nearly twice as likely to be misclassified as higher risk compared to their white counterparts
- (45% vs. 23%)

What did Probulica find-Point 2

- Analysis found that white defendants who re-offended within the next two years were mistakenly labeled low risk almost twice as often as African American re-offenders
- (48 percent vs. 28 percent)

What did Probulica find-Point 3

- Even when controlling for prior crimes, future recidivism, age, and gender, African American defendants were 45 percent more likely to be assigned higher risk scores than white defendants.

- In violent crime category, African American defendants were 77 percent more likely to be assigned higher risk scores than white defendants.

What did Probulica find- Point 4

- African American defendants were also twice as likely as white defendants to be misclassified as being a higher risk of violent recidivism.

- White violent recidivists were 63 percent more likely to have been misclassified as a low risk of violent recidivism, compared with African American violent recidivists.

Propublica's results-violent crimes

All Defendants		
	Low	High
Survived	4121	1597
Recidivated	347	389
FP rate: 27.93		
FN rate: 47.15		
PPV: 0.20		
NPV: 0.92		
LR+: 1.89		
LR-: 0.65		

Black defendants		
	Low	High
Survived	1692	1043
Recidivated	170	273
FP rate: 38.14		
FN rate: 38.37		
PPV: 0.21		
NPV: 0.91		
LR+: 1.62		
LR-: 0.62		

White defendants		
	Low	High
Survived	1679	380
Recidivated	129	77
FP rate: 18.46		
FN rate: 62.62		
PPV: 0.17		
NPV: 0.93		
LR+: 2.03		
LR-: 0.77		

Method

- Create a logistic regression model
- Race
- Age
- Gender
- Criminal history
- Future recidivism
- Charge degree

The Secrecy behind Northpointe's model

- The company does not publicly disclose the calculations used to arrive at defendants' risk scores
- Not possible for either defendants or the public to see what might be driving the disparity.

Conclusion

- Logistic regression model might be an accurate explanation model since it correctly mimics the predictions of the original model. It would not be faithful to what the original model computes.

- Created a linear explanation model for COMPAS that depended on race, and then accused the black box COMPAS model of depending on race, conditioned on age and criminal history.
- COMPAS seems to be nonlinear, and it is entirely possible that COMPAS does not depend on race
- ProPublica's linear model was not truly an "explanation" for COMPAS, and they should not

have concluded that their explanation model uses the same important features as the black box it was approximating