Announcements

- Project Milestone 3 due **Friday, December 9 at 8pm**
  - 2 day extension

- Final exam is **Thursday, December 22 from 6-8pm**
  - Review sessions in the remaining classes
  - Example final exam has been released

Lecture 25: Ethics

CIS 4190/5190

Fall 2022

Ethics is Hard!

- **Ethical decision-making**
  - Challenging problem even without ML
  - Thousands of years of debate in philosophy, law, etc.
  - Changes over time with changing societal norms

- **Challenges with machine learning**
  - Data privacy issues
  - Internalize (and even amplifies) biases already present in data
  - New issues related to abuse of ML

ML Applications

- **Fairness/discrimination issues**
  - Policing/judicial decisions, financial decisions, etc.
  - Filtering resumes of job applicants
  - Global aid allocation based on satellite images
  - Echo chamber issues in news/video recommendations

- **Potentially problematic applications**
  - Dangers in safety-critical settings
  - Automating wide-scale surveillance based on facial recognition
  - Autonomous drones for military uses
  - Refugees turned away at the US border because an ML system assessed risk of terrorist activity based on Instagram posts

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Data Privacy Issues

- **Pima People Diabetes Dataset**
  - "Members of the tiny, isolated tribe had given DNA samples to university researchers starting in 1990, in the hope that they might provide genetic clues to the tribe's devastating rate of diabetes. But they learned that their blood samples had been used to study many other things, including mental illness and theories of the tribe's geographical origins that contradict their traditional stories."
  - **Data collection requires informed consent**

- Public data ≠ consent for research use

The New York Times

Subscribe now

Indian Tribe Wins Fight to Limit Research of Its DNA

321

Edmond Tilousi, 56, who can climb the eight miles to the rim of the Grand Canyon in three hours. Jim Wilson/The New York Times

By Amy Harmon

April 21, 2010

Discrimination in ML

- ML models may be biased against minorities

Discrimination in ML

- ML models may be biased against minorities

Gender stereotype *she-he* analogies.

sewing-carpentry	register-nurse-physician	housewife-shopkeeper
nurse-surgeon	interior designer-architect	softball-baseball
blond-burly	feminism-conservatism	cosmetics-pharmaceuticals
giggle-chuckle	vocalist-guitarist	petite-lanky
sassy-snappy	diva-superstar	charming-affable
volleyball-football	cupcakes-pizzas	hairdresser-barber

Gender appropriate *she-he* analogies.

queen-king	sister-brother	mother-father
waitress-waiter	ovarian cancer-prostate cancer	convent-monastery

Bolukbasi et al. 2016 : https://arxiv.org/abs/1607.06520
Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2veec/

Sources of Bias

- **Data representation:** Distribution of inputs $p(x)$
- **Tainted labels:** Distribution of label assignments $p(y \mid x)$
- **Sensitive features:** Selecting what features to include for each sample (e.g., whether to include sensitive attributes such as race and gender)

Data Representation

- Less data from minority groups → Higher error on minority groups
- **Example:** Many clinical trials historically recruited largely white males, leading to biases in understanding outcomes and side effects
- **Example:** Focus on easily accessible data (e.g. recent tweets, or easily measured features of people) can lead to biased datasets
- Need to be careful to gather representative datasets

Tainted Labels

- **Example:** Amazon hiring bias
  - Amazon's ML resume screening tool to predict hiring decisions based on 10 years of historical applicant data; but found it was biased against women
  - Labels tainted by historical bias
  - https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

- **Similar example**
  - Company filters hires by predicting how long they will stay at the company
  - But how long someone stays depends on how they were treated

Tainted Labels

- **Example:** Predictive policing
  - “PredPol” predictive policing system employed in some policy departments
  - Suppose that crime happens equally everywhere
  - Some areas more policed → More crime found in those areas
  - ML learns to predict crime in neighborhoods that were more policed

Tainted Labels

- Need to be careful that labels are unbiased

- However, can be very hard to unbias data!
  - “We should strive to avoid giving women lower salaries”
  - **ML model:** “women” = “lower salaries”

Sensitive Attributes as Features

- When should sensitive attributes be used as features?

- **Example:** Predicting diabetes risk
  - Race is a sensitive attribute that may not cause diabetes, but may be correlated with unrecorded features that cause diabetes
  - What if an insurance company decides that people of some races are at higher risk and should pay higher premium?

- Omitting sensitive attributes is not enough!
  - Other features such as current income may be correlated with race/gender

Data Collection Issues

- Need to gather representative sample
- Need to ensure labels are unbiased
- Need to think carefully about whether to include sensitive attributes

Datasheets for Datasets (Gebru et al.)

- Questions for dataset creators to think through and answer for users:
  - Motivation
  - Dataset Composition
  - Collection Process
  - Preprocessing
  - Uses
  - Distribution
  - Maintenance

- https://arxiv.org/abs/1803.09010

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Fairness and ML

- What does it mean to be fair?

Case Study: Criminal Justice

- Software by Northpointe to predict **recidivism** for defendants
  - I.e., risk of committing future crimes

- Used to help make bail, sentencing, and parole decisions

Case Study: Criminal Justice

- **Features:** 137 questions answered by defendants or criminal records:
  - “Was one of your parents ever sent to jail or prison?”
  - “How many of your friends/acquaintances are taking drugs illegally?”
  - “How often did you get in fights while at school?”
  - Agree or disagree? “A hungry person has a right to steal”
  - Agree or disagree? “If people make me angry or lose my temper, I can be dangerous.”

- Exact algorithm and model is a trade secret

Case Study: Criminal Justice

- Race is **not** a feature
- **Problem:** Correlated features
  - One of the developers of the system said it is difficult to construct a score that doesn't include items that can be correlated with race
  - E.g., poverty, joblessness and social marginalization
  - "If those are omitted from your risk assessment, accuracy goes down"
- Similar to Amazon hiring bias example

Case Study: Criminal Justice

MACHINE BIAS

Bias in Criminal Risk Scores Is Mathematically Inevitable, Researchers Say

ProPublica's analysis of bias against black defendants in criminal risk scores has prompted research showing that the disparity can be addressed — if the algorithms focus on the fairness of outcomes.

by Julia Angwin and Jeff Larson, Dec. 30, 2016, 4:44 p.m. EST

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Defining Fairness

- **Legally Protected Attributes**
  - Race, sex, color, religion, national origin (Civil Rights Act of 1964, Equal Pay Act of 1963)
  - Age (Discrimination in Employment Act of 1967)
  - Citizenship (Immigration Reform and Control Act)
  - Pregnancy (Pregnancy Discrimination Act)
  - Familial status (Civil Rights Act of 1968)
  - Disability (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
  - Veteran status (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
  - Genetic information (Genetic Information Nondiscrimination Act)

Defining Fairness

- **Potential definition:** Two individuals differing on sensitive attributes but otherwise identical should receive the same outcome

- **Issue:** What does it mean for two people to be “otherwise identical”?
  - What if just their accents differ?
  - What if just their attire differs?

- Also ignores historical discrimination encoded in features, which is even harder to address

Defining Fairness

- **Accuracy and fairness**
  - Low accuracy can result in unfairness
  - E.g., strong student scored as highly as weak one for college admissions
  - But highest accuracy model is not necessarily the most fair

- **Group fairness:** Account for performance on subgroups

$$\text{Fairness metric} = F\big(L(f; X_1), \ldots, L(f; X_k)\big)$$

Group Fairness

- **Problem setup**
  - Sensitive attribute A
  - ML model R mapping input features X to prediction $\hat{Y} = R(X)$
  - True outcome Y (typically binary, and $Y = 1$ is the "good" outcome)

- **Example:** Insurance risk prediction
  - $A = \text{age}$
  - $R = \text{predicted cost}$
  - $Y = \text{true cost}$

Group Fairness

- **Independence:** Risk score distribution should be equal across ages:

$$P(\text{risk score} \mid \text{age}) = P(\text{risk score})$$

  - E.g., equal proportion of low risk customers for young vs. old people
  - Often called demographic parity

- What if lower age groups in fact behave more riskily?

Group Fairness

- **Separation:** Risk score should be independent of age given outcome:

$$P(\text{risk score} \mid \text{age}, \text{true outcome}) = P(\text{risk score} \mid \text{true outcome})$$

  - Equivalent to saying the true positive rate and false positive rate are equal across subgroups

- **Example:** Both of the following hold:
  - Fraction of young, low-insurance-usage people correctly identified as low-risk = Fraction of old low-insurance-usage people correctly identified as low-risk
  - Fraction of young high-insurance-usage people wrongly identified as low-risk = Fraction of old high-insurance-usage people wrongly identified as low-risk

Group Fairness

- **Sufficiency:** Outcome should be independent of risk score given age:

$$P(\text{true outcome}, \text{age} \mid \text{risk score}) = P(\text{true outcome} \mid \text{risk score})$$

  - Intuitively, risk score tells us everything we need to know about the true outcome with respect to age

Group Fairness

Non-discrimination criteria		
Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Group Fairness

- Three notions are incompatible!

Proposition 2. *Assume that A and Y are not independent. Then sufficiency and independence cannot both hold.*

Proposition 3. *Assume Y is binary, A is not independent of Y, and R is not independent of Y. Then, independence and separation cannot both hold.*

Proposition 5. *Assume Y is not independent of A and assume $\hat{Y}$ is a binary classifier with nonzero false positive rate. Then, separation and sufficiency cannot both hold.*

- Thus, need carefully choose what kinds of fairness we ask for

Solon Barocas, Moritz Hardt, Arvind Narayanan, “Fairness and Machine Learning”

Algorithms for Ensuring Fairness

- Given a notion of fairness, there are a few ways of achieving it

- **Example:** Independence
  - **Pre-processing:** Adjust features to be uncorrelated with sensitive attribute
  - **Training constraints:** Impose the constraint during training
  - **Post-processing:** Adjust the learned classifier so its predictions are uncorrelated with the sensitive attribute

- **Goodhart's law:** "When a measure becomes a target, it ceases to be a good measure" – Marilyn Strathern
  - Do not blindly impose fairness, need to carefully examine predictions

Human-in-the-Loop Fairness

- **Potential solution:** Have domain experts weigh in on what performance metrics result in fair model selection/training

- **Challenges**
  - Experts may not understand limitations of ML models (e.g., does a judge using a system understand that it only has 60% accuracy?)
  - Potential for selective enforcement based on human biases

Human-in-the-Loop Fairness

- **Example:** In bail decision-making, judges selectively follow model
  - Less lenient against younger defendants, especially minorities
  - Younger defendants are actually more risky, but judges may have been lenient due to societal norms (e.g., “second chance”)
  - Judges followed algorithm less and less over time

https://www.washingtonpost.com/business/2019/11/19/algorithms-were-supposed-make-virginia-judges-more-fair-what-actually-happened-was-far-more-complicated/

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Misinformation about ML

6.1 The public predicts a 54% likelihood of high-level machine intelligence within 10 years

Respondents were asked to forecast when high-level machine intelligence will be developed. High-level machine intelligence was defined as the following:

> We have high-level machine intelligence when machines are able to perform almost all tasks that are economically relevant today better than the median human (today) at each task. These tasks include asking subtle common-sense questions such as those that travel agents would ask. For the following questions, you should ignore tasks that are legally or culturally restricted to humans, such as serving on a jury.[13]

Respondents were asked to predict the probability that high-level machine intelligence will be built in 10, 20, and 50 years.

Comparison: Experts predicts in the ~50-year (may be optimistic)

Example: Self-Driving Without LIDAR

Example: Resume Evaluation

How to persuade a robot that you should get the job

Do mere human beings stand a chance against software that claims to reveal what a real-life face-to-face chat can't?

Stephen Buranyi

Sat 3 Mar 2018 19.05 EST

Based on slides by Arvind Narayanan

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Feedback Loops in ML Systems

- ML models are often part of a larger system

- **Example:** Feedback loop in PredPol (used to predict crime)
  - This kind of approach is “especially nefarious” because police can say: “We’re not being biased, we’re just doing what the math tells us.” And the public perception might be that the algorithms are impartial. – Samuel Sinyangw

Feedback Loops in ML Systems

- **Recommender systems:** "A system for predicting the click through rate of news headlines on a website likely relies on user clicks as training labels, which in turn depend on previous predictions"

- **Potential for adversarial feedback**
  - Tricking a resume screening system by entering keywords like "Oxford"
  - **Anecdotal:** Computer vision systems to predict poverty and (semi-) automate global aid allocation decisions lead to people switching off their night lights and dressing up concrete roofs as thatched roofs

Satellite images used to predict poverty
By Paul Rincon
Science editor, BBC News website

Machine Learning: The High Interest Credit Card of Technical Debt

D. Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young
SE4ML: Software Engineering for Machine Learning (NIPS 2014 Workshop)

Extreme Example: “Future Features”

- **Scenario**
  - Build a highly complex classifier with 99% accuracy for a time-series problem
  - Later, build a new classifier with 98.5% accuracy, runs $1000\times$ faster
  - Catastrophic failure when deployed!

- **Problem**
  - Training data included classifier’s prediction from previous step as input
  - **New classifier:** “Recycles” the prediction from the previous step (i.e., just use that single feature as the prediction!)
  - Works fine when previous prediction was already accurate
  - No longer the case after deployment!

Potential Solution

- **DAGGER algorithm**
  - Originally designed for imitation learning (i.e., RL from expert data)
  - Continuously collect new labels and add to training set

- $Z \leftarrow$ Initial dataset
- For $t \in \{1,2, \ldots\}$:
  - Train f_β on D and use to make decisions on new examples X_t
  - Observe (or collect) ground truth labels Y_t for X_t
  - $Z \leftarrow Z \cup \{(X_t, Y_t)\}$

- Use multi-armed bandits when there is partial feedback

Drew Bagnell, Feedback in Machine Learning, https://www.youtube.com/watch?v=XRSvz4UOpo4

More Challenging Feedback Loops

- **Example:** Hiring ads
  - Women tend to click on job ad with second-highest salary
  - ML model learns that women do not click on highest salary job ad, so it stops recommending it
  - Second-highest salary job ad → Highest salary job ad
  - Women click on new second-highest salary job ad!

- No substitute for manual analysis of ML models in projection
  - You'll never be out of a job (at least for the foreseeable future)!

Lambrecht & Tucker. Algorithmic Bias? An Empirical Study of Apparent Gender-Based Discrimination in the Display of STEM Career Ads

Agenda

- Dataset issues
- Fairness/discrimination in ML models
- Misinformation about ML
- Feedback in ML systems
- Practical principles for ethical ML

Ethical Issues

- When you build ML models, you are responsible for how it is eventually deployed
  - Face classifier may be used by an authoritarian government to track people or target minority subgroups
  - Technology may be used in safety critical settings without sufficient validation

Best Practices for Ethical ML

- Human augmentation
- Bias evaluation
- Explainability and justification
- Displacement strategy

Human Augmentation

- Assess the impact of incorrect predictions and, when reasonable, design systems with human-in-the-loop review processes

- Especially important in domains with significant impact on human lives (e.g. justice, health, etc.)
  - All stakeholders' values and perspectives should be accounted for during algorithm design
  - Domain experts as human-in-the-loop reviewers of ML decisions

Bias Evaluation

- **Use tools to understand bias in ML models**
  - No standard strategy, need to careful consider potential sources of bias for the domain you are working in
  - Requires continuous monitoring, not one-time effort

Explainability and Justification

- **Use tools to explain ML predictions**
  - Even though accuracy may decrease, the explainability may be significant
  - Important for end users to be able to understand ML predictions
  - Especially important due to hype and misinformation about ML

- **Challenges**
  - Potential leaking of sensitive data
  - Easy to game, e.g., “adversarial feedback”
  - Loss of competitive advantage
  - Sometimes hard to interpret, even for experts

Samuele Lo Piano, “Ethical Principles …” 2020

Explainability and Justification

- **Legal considerations**
  - France’s Digital Republic Act gives the right to an explanation as regards decisions on an individual made by algorithms
  - How and to what extent the algorithm was used, which data was processed and its source, etc.
  - Other countries considering similar laws

Displacement Strategy

- Identify and document relevant information so that business change processes can be developed to mitigate the impact on workers being automated

- Ensure all stakeholders are brought on board and develop a change-management strategy before automation

- Often, the workers are asked to do labor (e.g., generating training data) that will help automate themselves. Are the appropriately compensated?

Accountability

- **Question:** Should a passenger in automated car be able to command it to go 80 MPH on a 55 MPH road?

- **Reasons for "No"**
  - It's illegal and can endanger others
  - Who is liable for accidents? Driver? Manufacturer? Insurance company?

- **Reasons for "Yes"**
  - Many exceptions!
  - Rushing someone to the hospital, escaping a tornado, etc.

Other Challenges

- The ethics of ML and AI systems is an urgent topic **now**, not because of speculative future scenarios
  - Open and active area of research, involves scholars from law, social sciences, etc., as well as domain experts
  - Law moves slowly, and legal frameworks have much to catch up to

- **Looking forward**
  - **AI safety**: How can we make AI without unintended negative consequences?
  - **AI alignment:** How can AI make decisions that align with our values?

Useful Tools

- IBM AI Fairness 360: https://aif360.mybluemix.net/
- Google ML Fairness Gym: https://github.com/google/ml-fairness-gym
- Facebook Fairness Flow: https://venturebeat.com/2021/03/31/ai-experts-warn-facebooks-anti-bias-tool-is-completely-insufficient/