CSDS 440: Machine Learning

Soumya Ray (he/him, sray@case.edu)

Olin 516

Office hours T 11:15-11:45 or by appointment

Today

- Fairness and bias in automated decision making

Automated Decision Making

- Also “algorithmic” decision making
  - Some sort of automatic procedure is followed, with real-world effects (actions)
  - Has been around since forever
  - E.g. credit scores, vehicle safety ratings, etc

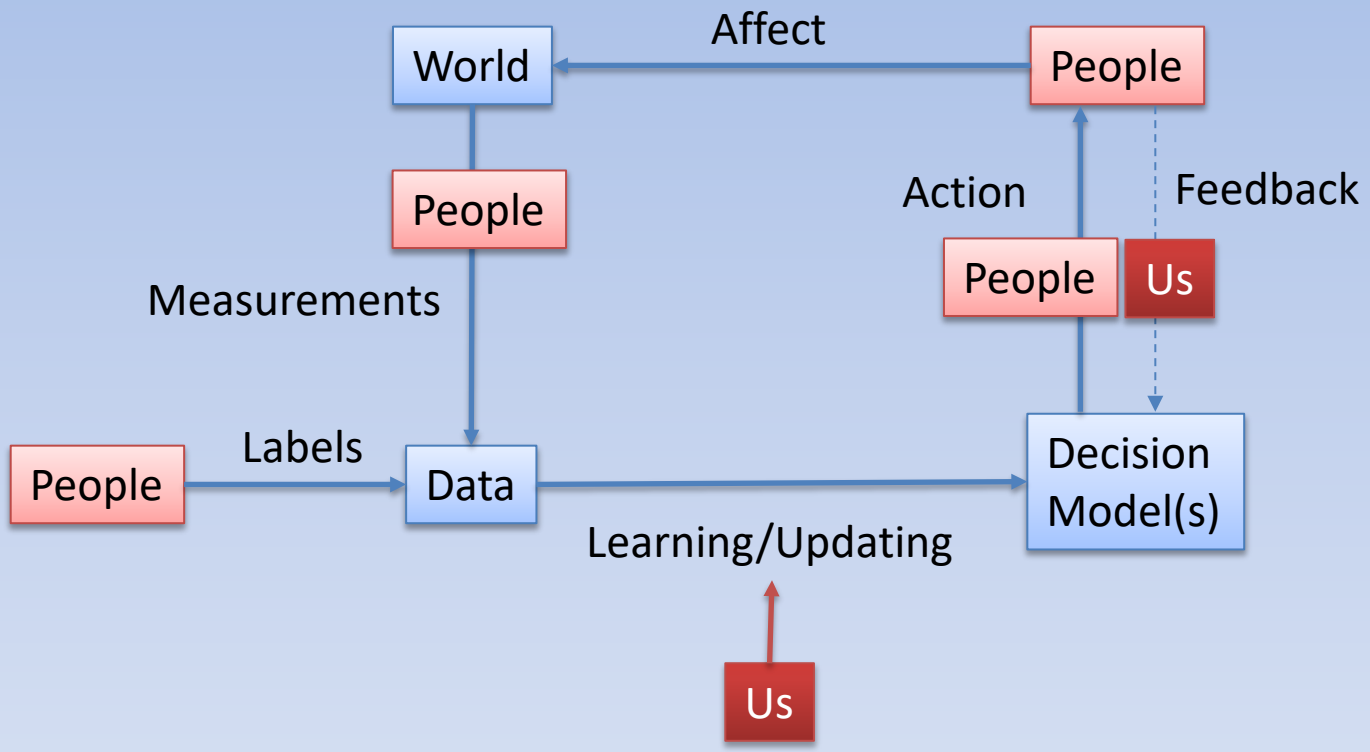
Bias and Fairness in Automated Decision Making

- So what's changed?
  - **Breadth** and **scale**
    - ML methods are now widely being deployed in practice, and being used to make decisions that affect people and lives at scale
    - Breadth and scale make errors in these systems, and the consequences of such errors unfortunate to devastating
  - **Opaqueness**
    - Mechanisms are not understandable even to their designers/deployers
    - Opaqueness makes errors hard to anticipate/debug/fix

Bias and Fairness in Automated Decision Making

- Of particular concern are *systematic* errors that are somehow *consequential for specific subgroups of the population*
- How can these happen?
- Can we quantify this kind of bias?
- How can we mitigate this?
- Before going further, we must acknowledge that *social inequities cannot be fully mitigated through technology*

The ML-Deployment Loop



1. It's a loop!
2. People are involved at multiple points.
3. Where is the power? And by extension, the responsibility?

Biases in Measurement and Labels

- If a decision making problem is about people, then the data involves characteristics/ products of people
  - These are often hard to quantify and may be biased
- The target to predict may also not be crisply defined
  - “Creditworthiness”
  - “Will Succeed at Position”

Learning

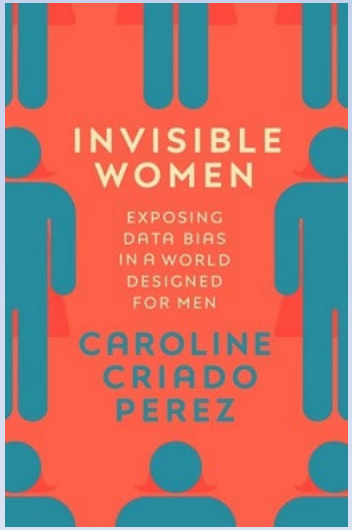
- If we learn models from biased data, we would *expect* that the model learns the biases
  - This is not surprising, we train models to do well on data that looks like the input!

- *So ML systems by default are designed to perpetuate bias*

When we build a statistical model of language, we should expect the gender associations of occupation words to roughly mirror real-world labor statistics. In addition, because of the male-as-norm bias (the use of male pronouns when the gender is unknown) we should expect translations to favor male pronouns. It turns out that when we repeat the experiment with dozens of occupation words, these two factors—labor statistics and the male-as norm bias—together almost perfectly predict which pronoun will be returned [by a language model]. (Caliskan, Bryson, and Narayanan, "Semantics Derived Automatically from Language Corpora Contain Human- Like Biases," Science 356, no. 6334 (2017): 183–86.)

Could we simply “clean” the data?

- Unfortunately, it’s often not that simple
  - Removing demographic/economic information does not imply the model will be blind to these (examples later)
  - Often you can’t tell how a feature is biased when it is used, the bias reveals itself over time in the pattern of predictions
  - Further, disparities often arise not due to explicit bias but due to *lack of data*



Predictions to Action

- In most ML models, predictions are based on *correlations* not *causation*
- This makes associated decisions sometimes arbitrary, and justifying those decisions problematic
- Second, remember the loop? Actions taken on the basis of predictions *often invalidate the prediction!*
  - e.g. traffic congestion prediction

The Feedback Loop

- A general source of bias is that *future data depends on the ML system's predictions*
  - The world is not i.i.d
- Even if the ML system expects a feedback loop, the nature of the feedback could itself be biased

Google searches for Black-sounding names such as "Latanya Farrell" were much more likely to results in ads for arrest records ("Latanya Farrell, Arrested?") than searches for White-sounding names ("Kristen Haring"). One potential explanation is that users are more likely to click on ads that conform to stereotypes, and the advertising system is optimized for maximizing clicks. (Sweeney, "Discrimination in Online Ad Delivery," Queue 11, no. 3 (March 2013): 10:10–29)

The Feedback Loop

- Just as the loop can invalidate predictions, it can also amplify predictions (and bias if the predictions were biased)
  - Example: Predictive policing

A 2016 paper analyzed a predictive policing algorithm... By applying it to data derived from Oakland police records, they found that Black people would be targeted for predictive policing of drug crimes at roughly twice the rate of Whites, even though the two groups have roughly equal rates of drug use. [A] simulation showed that this initial bias would be amplified by a feedback loop, with policing increasingly concentrated on targeted areas. This is despite the fact that the PredPol algorithm does not explicitly take demographics into account. (Lum and Isaac, “To Predict and Serve?” Significance 13, no. 5 (2016): 14–19. Ensign et al., “Runaway Feedback Loops in Predictive Policing,” arXiv Preprint arXiv:*1706.09847*, 2017.)

How do we make ML models less biased?

- Some general principles:
  - Make code/data **transparent to inspection** (as much as possible)
  - Identify **stakeholders** and potential **costs**
    - Cathy O'Neil, Hanna Gunn (2020). Near-Term Artificial Intelligence and the Ethical Matrix . Ch 8, Ethics of Artificial Intelligence, S. Matthew Liao (ed.) https://doi.org/10.1093/oso/9780190905033.003.0009
  - Prioritize **predictability**
  - Provide **explainability**
  - Make algorithms **robust against manipulation**
  - Establish clear chains of **accountability**

Fairness Criteria

- How can we modify algorithms to mitigate bias?

- People have proposed incorporating additional fairness criteria into the learning process, which reduce the impact of biased data or verify the result is un/less biased

Key idea

- We can try to formalize “fairness” through *probabilistic relationships* that should hold between the true and predicted labels and attributes defining subpopulations

Notation

- Let A denote a set of "sensitive attributes": different values for these attributes identify protected groups
- Let Y denote the target variable
- Let R denote the classifier's output

Independence

- $R \perp A$
  - The predicted output is independent of the sensitive attribute
  - Alternatively, $\mathrm{P}(R=1|A=\mathrm{a})=\mathrm{P}(R=1|A=\mathrm{b})$
  - In general, could ask for these to be within ε of each other
  - Intuition: “traits relevant for target are independent of sensitive attributes”

Issues

- This is ideal, but in practice, independence is tricky to satisfy
- It mandates equal output probabilities, but what about the *input*?
- E.g., suppose we have a lot of data for A=a, but little for A=b

Separation

- $R \perp A|Y$
  - The predicted score is independent of the sensitive attribute, *given the class label*
  - Alternatively, $\mathrm{P}(R=1|Y=1, A=\mathrm{a})=\mathrm{P}(R=1|Y=1,A=\mathrm{b})$ and $\mathrm{P}(R=1|Y=0, A=\mathrm{a})=\mathrm{P}(R=1|Y=0,A=\mathrm{b})$
  - Intuition: “all groups experience the same false positive rates and same false negative rates”

Sufficiency

- $Y \perp A|R$
  - The class label is independent of the sensitive attribute, *given the predicted score*
    - "Inverse" of separation
  - Alternatively, $\mathrm{P}(Y=1|R=1, A=\mathrm{a})=\mathrm{P}(Y=1|R=1,A=\mathrm{b})$ and $\mathrm{P}(Y=1|R=0, A=\mathrm{a})=\mathrm{P}(Y=1|R=0,A=\mathrm{b})$
  - Intuition: "all groups experience the same precision and same *negative predictive value*"

Relationships between Criteria

- These criteria all embody different aspects of fairness
- Unfortunately, they are related and it can be shown that generally not all of them can be satisfied at once
- So once again, tradeoffs must be made in the "type" of fairness we ask for

Example

- Suppose that A and Y are not independent. Then sufficiency and independence cannot both hold.
  - Try proving this!

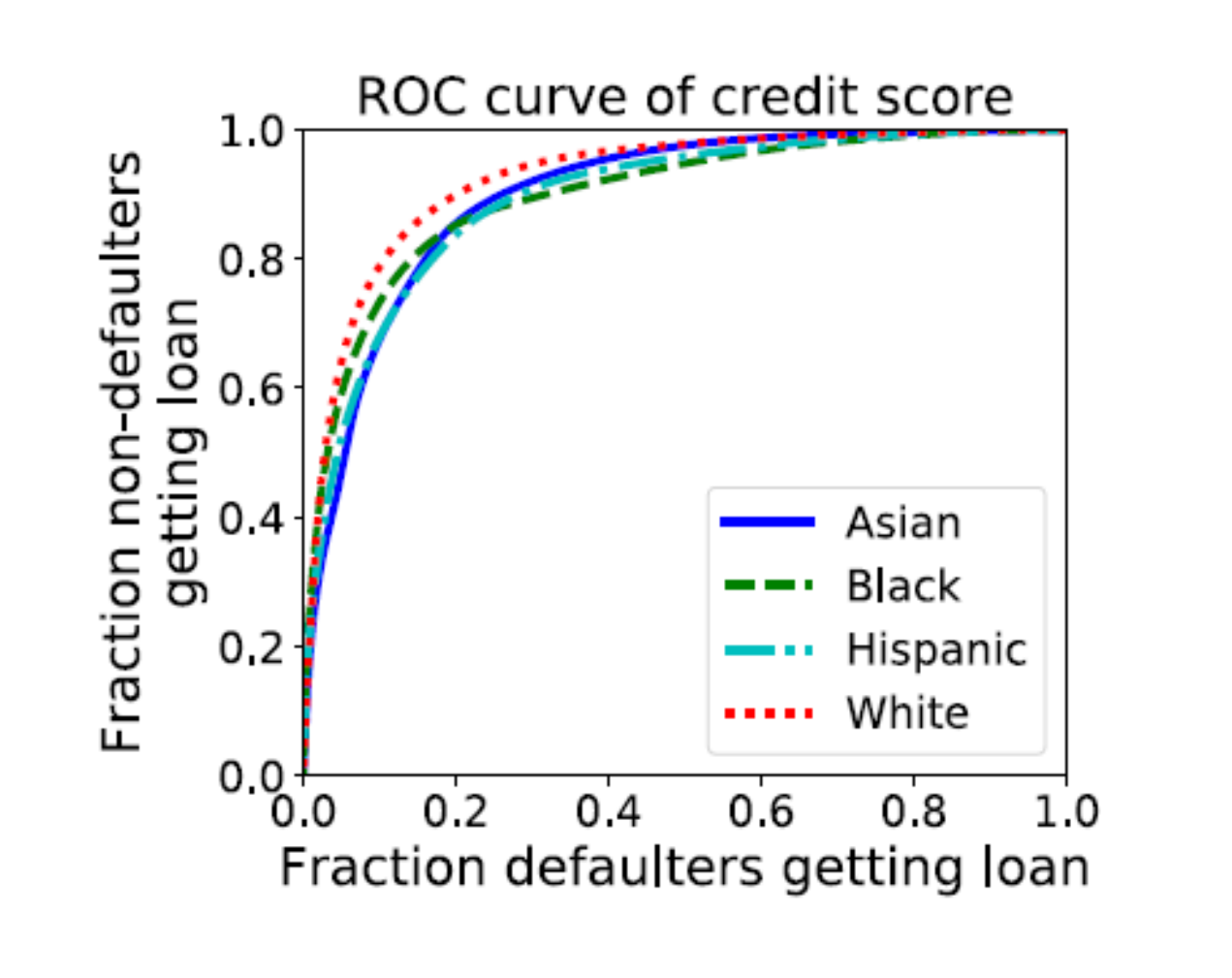
Satisfying Fairness Criteria

- One way to satisfy these criteria is through ROC analysis
- Once we have plotted and ROC graph for the classifier, we can choose operating points
- For most criteria, the operating points *will be different* by subgroup

Example: credit scoring (Barocas et al. 2021, Fairness and Machine Learning: Limitations and Opportunities)

- Predict “serious delinquency in at least one credit line of a certain time period”
- Using TransUnion credit scores normalized 0-100 (0 is “least creditworthy”)
- Authors obtained info on self-reported race from SSA

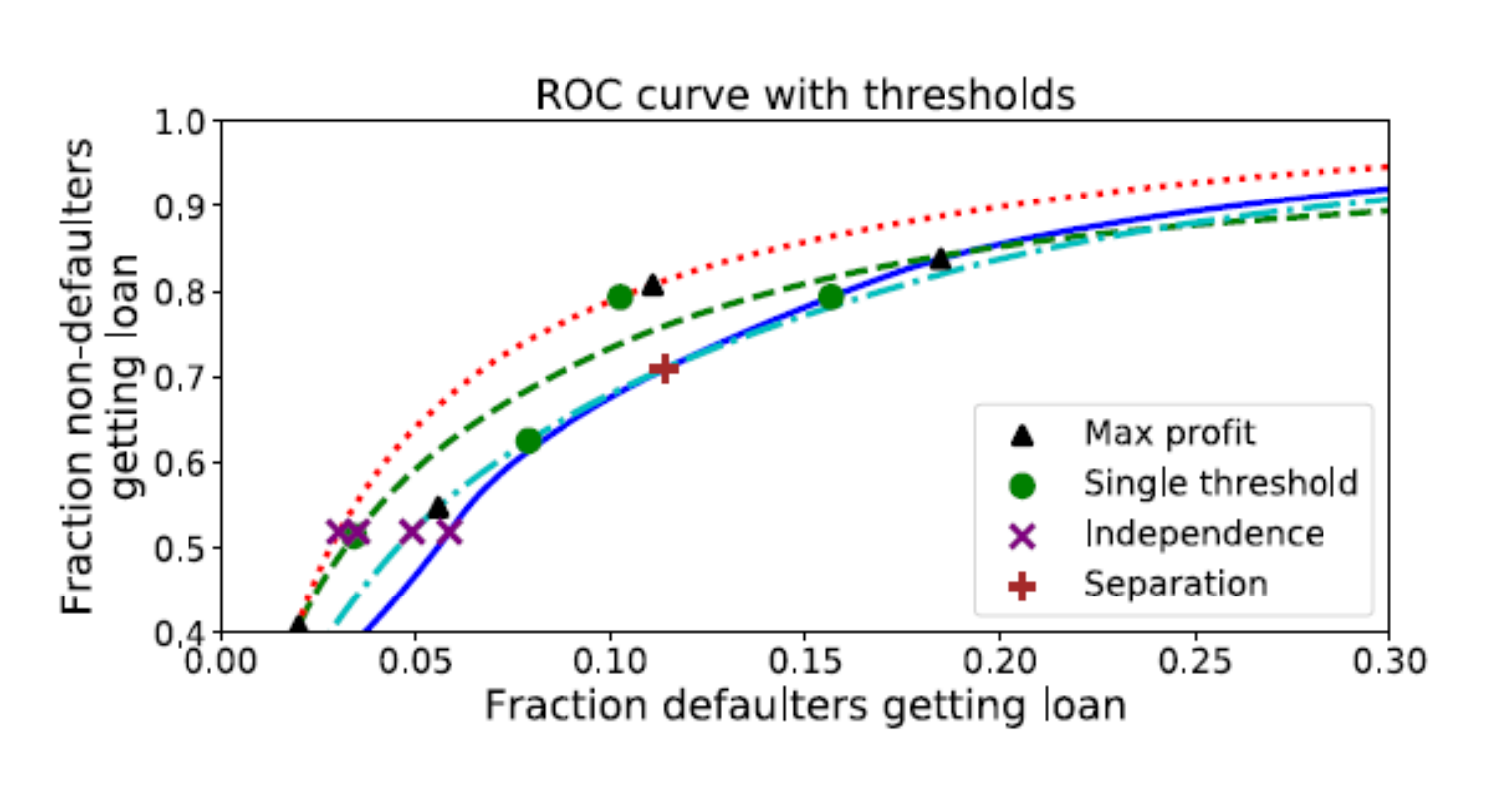
Example: credit scoring



Example: credit scoring

- Four different classification strategies:
  - Maximum profit: Pick possibly group-dependent score thresholds in a way that maximizes profit.
  - Single threshold: Pick a single uniform score threshold for all groups in a way that maximizes profit.
  - Separation: Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
  - Independence: Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.
- To “maximize profit”, assume the cost of a false positive is 6 times greater than the return on a true positive

Example: credit scoring



"'single threshold' achieves higher profit than 'separation,' which in turn achieves higher profit than 'independence.'" Hardt, Price, and Srebro, "Equality of Opportunity in Supervised Learning," in Proc. 29th Neur Info Proc Sys, 2016, 3315–23.

Transparency and Explainability

- As important as decisions themselves is the *process of decision making*
- It needs to be transparent and trust-enhancing
- To do this, it is important to develop explainable decision making systems, that can produce human-comprehensible explanations of their internal deliberations

Summary

- As ML is applied to more decision making tasks, being fair and unbiased---avoiding systematic errors that disadvantage specific subpopulations---is increasingly important
- The nature of the process makes it likely bias will creep in at many points, and will be perpetuated or even amplified
- We can attempt to make our systems robust to this, but there are many fairness criteria and tradeoffs exist between them
- Understanding these tradeoffs helps mitigate (certain kinds of) bias
- But, we must remember that it is unlikely there will ever by a fully technological solution to social inequity issues

Next Steps

- If you liked this class, here are some others to consider

Foundations

- Mathematical Programming (CSDS 477, MATH 327/427/433)
- Probability and Statistics (MATH 380, others)

- Potentially useful
  - Graph Theory (CSDS 455)
  - Mathematical Logic (CSDS 343, MATH 406)
  - Complexity (CSDS 343)
  - Numerical Analysis (ECSE 251)
  - Real and Functional Analysis (MATH 423, 424)

Related Courses

- Graphical Models (CSDS 491)
- Algorithmic Robotics (CSDS 499)
- Computer Vision (CSDS 531)
- Sequential Decision Making (CSDS 496)
- Natural Language Processing (CSDS 497)
- Computational Perception (CSDS 600)
- Information Theory (CSDS/MATH 394/494)
- Large Language Models
- Performant AI Systems
- Responsible AI Engineering
- ML on Graphs
- Causal Learning on Graphs

Thanks for attending!

- Please remember to turn in your Project writeups and code on/by 12/8 11:59pm(no extensions)
- Fill in course evaluations
  - Be as specific in your comments as you can
  - What was helpful? Wasn't helpful?
    - Lectures? Pdfs? Canvas? Books? Homeworks? Programming assignments? Project? Slack? Quizzes? Tests? Other?
- Contact me if interested in ML research and you liked and did well in this class
- Have a nice winter break and a healthy 2025! ☺