CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

Outline for May 3

- Go over common issues on Midterm 2
- Finish discussion of paper “Certifying and Removing Disparate Impact”
- Case study: Google Duplex
- Final thoughts

  - Last office hours TODAY! 1-3pm
  - Feel free to make a project appointment (up to one/group) any time before the presentation
  - Alternative time: May 13, 1-3pm
  - Main time: May 16, 9am-12pm

Outline for May 3

- Go over common issues on Midterm 2
- Finish discussion of paper “Certifying and Removing Disparate Impact”
- Case study: Google Duplex
- Final thoughts

Midterm Solutions

(not posted online)

Outline for May 3

- Go over common issues on Midterm 2
- Finish discussion of paper “Certifying and Removing Disparate Impact”
- Case study: Google Duplex
- Final thoughts

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y and outcome C

* X is protected (say X=0 is minority group, X=1 is majority)
* Y is unprotected (other features)
* C is outcome (say C=1 hired/admitted, C=0 not)

Indirect discrimination: C = f(Y)

* but strong correlation between X and Y

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y and outcome C

* X is protected (say X=0 is minority group, X=1 is majority)
* Y is unprotected (other features)
* C is outcome (say C=1 hired/admitted, C=0 not)

Indirect discrimination: $C = f(Y)$

* but strong correlation between X and Y

Disparate Impact (legal definition):

$$P(C=1 \mid X=0) < 0.8\, P(C=1 \mid X=1)$$

"Certifying and Removing Disparate Impact" Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

Certify (lack of) disparate impact

1) Train classifier f: Y -> X with the goal of minimizing the BER (balanced error rate)

Certify (lack of) disparate impact

1) Train classifier f: Y -> X with the goal of minimizing the BER (balanced error rate)
2) Calculate the BER of the optimal classifier, and call this value ε

Certify (lack of) disparate impact

1) Train classifier f: Y -> X with the goal of minimizing the BER (balanced error rate)
2) Calculate the BER of the optimal classifier, and call this value ε
3) Compute β, the fraction of minority applicants that were hired/admitted. Use β to compute ε', the optimal threshold

Certify (lack of) disparate impact

1) Train classifier f: Y -> X with the goal of minimizing the BER (balanced error rate)
2) Calculate the BER of the optimal classifier, and call this value ε
3) Compute β, the fraction of minority applicants that were hired/admitted. Use β to compute ε', the optimal threshold
4) If $\varepsilon' < \varepsilon$, no disparate impact

Repairing data to remove disparate impact

- Idea: change unprotected attribute(s) Y (assume just one attribute for now)

Repairing data to remove disparate impact

- Idea: change unprotected attribute(s) Y (assume just one attribute for now)

- Compute the distribution of Y with respect to each group x in X, call these values Y_x

Repairing data to remove disparate impact

- Idea: change unprotected attribute(s) Y (assume just one attribute for now)
- Compute the distribution of Y with respect to each group x in X, call these values Y_x
- Use "earth-movers" distance to shift each distribution of Y_x so that they have the same median

"Certifying and Removing Disparate Impact" Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, Suresh Venkatasubramanian

Example of repair

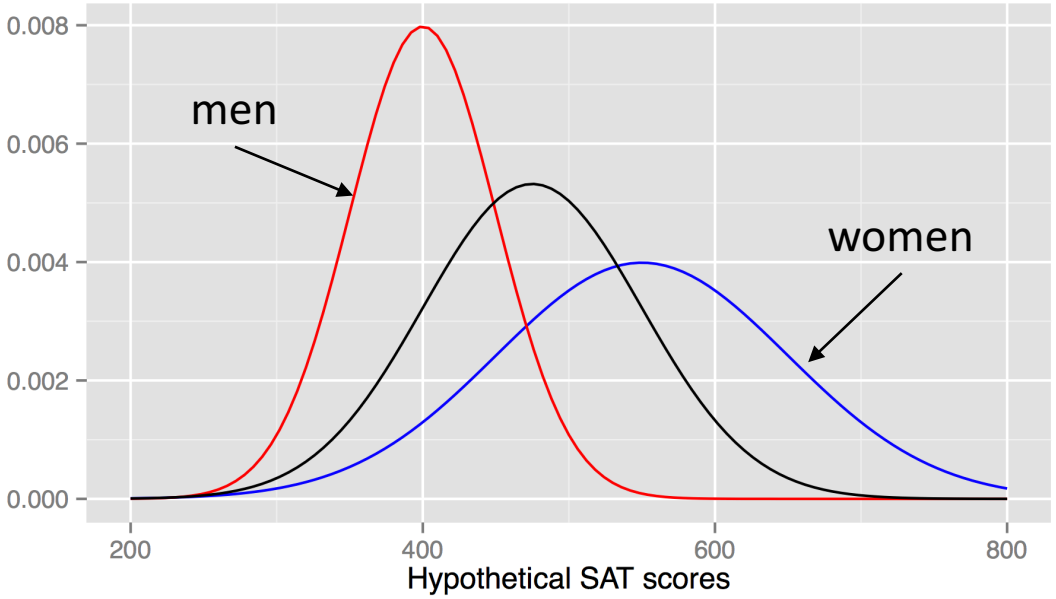

Figure 1: Consider the fake probability density functions shown here where the blue curve shows the distribution of SAT scores (Y) for $X = \texttt{female}$, with $\mu = 550, \sigma = 100$, while the red curve shows the distribution of SAT scores for $X = \texttt{male}$, with $\mu = 400, \sigma = 50$. The resulting fully repaired data is the distribution in black, with $\mu = 475, \sigma = 75$. Male students who originally had scores in the 95th percentile, i.e., had scores of 500, are given scores of 625 in the 95th percentile of the new distribution in $\bar{Y}$, while women with scores of 625 in $\bar{Y}$ originally had scores of 750.

Takeaways?

Outline for May 3

- Go over common issues on Midterm 2
- Finish discussion of paper “Certifying and Removing Disparate Impact”
- Case study: Google Duplex
- Final thoughts

Google Duplex

- Google Assistant can now make phone calls on behalf of users

https://youtu.be/lXUQ-DdSDoE?t=35

Discussion Questions (small groups)

1) How is Google inviting us to imagine its product being used?

2) What other possible uses can we imagine?

3) What are benefits/opportunities of this technology? (micro/macro)

4) What are costs/concerns of this technology? (micro/macro)

5) What choices are open to us if this is our product?

Questions adapted from Ronni Gura Sadovsky '08

Outline for May 3

- Go over common issues on Midterm 2
- Finish discussion of paper “Certifying and Removing Disparate Impact”
- Case study: Google Duplex
- Final thoughts

Discussion Questions

1) What are our responsibilities as engineers to ensure that our algorithms are fair?

2) How would you handle a situation where you felt you didn't have enough data (or the right data) necessary to build your algorithm?

3) How would you try to detect if your algorithm was making biased decisions during deployment?