CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2021

Admin

- **Midterm 1** handed out on Thursday (due the following Thursday – take in a 2 hour block)
- **Thursday**: review session in class/lab
- **TODAY**: make sure you have a midterm Study Guide and Handout 9 (or back of Handout 8)
- **Note-taker**: Joseph

Lab 4

- Lab 4 due TONIGHT
- Office hours today 3:30-5pm (H204)
- TA hours tonight 6:30-8:30pm (H110)

Midterm 1 Notes

- Handed out in class this Thursday, due the following Thursday.
- Timed exam: **2 hour limit**. DO NOT open the exam until you are ready to take it for 2 hours!
- You may use a one page (front and back) “study sheet”, handwritten, created by you
- You may also use a calculator
- Outside of your “study sheet” and calculator, **no other notes or resources**
- As per the Honor Code, all work must be your own

Outline for September 28

- Go over Lab 2
- Intro to Bayesian models
- Intro to algorithmic bias
- Redundantly encoded features and disparate impact

Outline for September 28

- Go over Lab 2
- Intro to Bayesian models
- Intro to algorithmic bias
- Redundantly encoded features and disparate impact

Not posted online

Outline for September 28

- Go over Lab 2
- Intro to Bayesian models
- Intro to algorithmic bias
- Redundantly encoded features and disparate impact

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Evidence**: this is the data (features) we actually observe, which we think will help us predict the outcome we're interested in

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Prior**: without seeing any evidence (data), what is our prior believe about each outcome (intuition: what is the outcome in the population as a whole?)

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Posterior**: this is the quantity we are actually interested in. **Given** the evidence, what is the probability of the outcome?

Components of a Bayesian Model

- Identify the evidence, prior, posterior, and likelihood in the equation below

$$p(y = k|\boldsymbol{x}) = \frac{p(y = k)p(\boldsymbol{x}|y = k)}{p(\boldsymbol{x})}$$

- **Likelihood**: given an outcome, what is the probability of observing this set of features?

Examples

- Last time: wanted to compute the probability an email message was **spam**, given the **words** of the email

- Another example: what is the probability of **Trisomy 21** (Down Syndrome), given the amount of sequencing of each chromosome?

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^{C}) \cdot \mathbb{P}(T_{21}^{C})}$$

Bayesian Model for Trisomy 21 (T_{21})

Input data are read counts for each chromosome (1,2,…,n):

$$q_1, q_2, \cdots, q_n = \vec{q}$$

Goal:

Prior probability of T_{21}

$$\mathbb{P}(T_{21}|\vec{q}\,) = \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,)}$$

$$= \frac{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21})}{\mathbb{P}(\vec{q}\,|T_{21}) \cdot \mathbb{P}(T_{21}) + \mathbb{P}(\vec{q}\,|T_{21}^{C}) \cdot \mathbb{P}(T_{21}^{C})}$$

Prior:

$P(T_{21})$

Maternal Age	Trisomy 21	All Trisomies
20	1 in 1,667	1 in 526
21	1 in 1,429	1 in 526
22	1 in 1,429	1 in 500
23	1 in 1,429	1 in 500
24	1 in 1,250	1 in 476
25	1 in 1,250	1 in 476
26	1 in 1,176	1 in 476
27	1 in 1,111	1 in 455
28	1 in 1,053	1 in 435
29	1 in 1,000	1 in 417
30	1 in 952	1 in 384
31	1 in 909	1 in 384
32	1 in 769	1 in 323
33	1 in 625	1 in 286
34	1 in 500	1 in 238
35	1 in 385	1 in 192
36	1 in 294	1 in 156
37	1 in 227	1 in 127
38	1 in 175	1 in 102
39	1 in 137	1 in 83
40	1 in 106	1 in 66
41	1 in 82	1 in 53
42	1 in 64	1 in 42
43	1 in 50	1 in 33
44	1 in 38	1 in 26
45	1 in 30	1 in 21
46	1 in 23	1 in 16
47	1 in 18	1 in 13
48	1 in 14	1 in 10
49	1 in 11	1 in 8

Handout 9 (back of Handout 8)

PLEASE LEAVE COMPUTERS ON

① $p(D \mid pos) = \frac{p(D)\,p(pos \mid D)}{p(pos)}$ ← normalizer

$p(D \mid pos)$: true positive rate

$= \frac{p(D)\,p(pos \mid D)}{p(pos \cdot H) + p(pos \cdot D)}$ (· and, + or)

$= \frac{p(D)\,p(pos \mid D)}{p(H)\,p(pos \mid H) + p(D)\,p(pos \mid D)}$

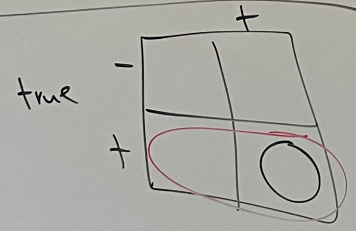
$= \frac{\frac{1}{100} \cdot \frac{9}{10}}{\frac{99}{100} \cdot \frac{1}{10} + \frac{1}{100} \cdot \frac{9}{10}}$

$= \frac{9}{99+9} = \frac{9}{108}$

$= \frac{1}{12} \approx 8.3\%$

$p(D \mid pos) + p(H \mid pos) = 1$

$p(D \mid neg) + p(D \mid pos) \neq 1$



Outline for September 28

- Go over Lab 2
- Intro to Bayesian models
- Intro to algorithmic bias
- Redundantly encoded features and disparate impact

What does it mean to claim that algorithms are biased (or racist or political…)?

```
model = initialization(...)
n_epochs = ...
train_data = ...
for i in n_epochs:
    train_data = shuffle(train_data)
    X, y = split(train_data)
    predictions = predict(X, model)
    error = calculate_error(y, predictions)
    model = update_model(model, error)
```

Pseudocode from A Gentle Introduction to Mini-Batch Gradient Descent and How to Configure Batch Size

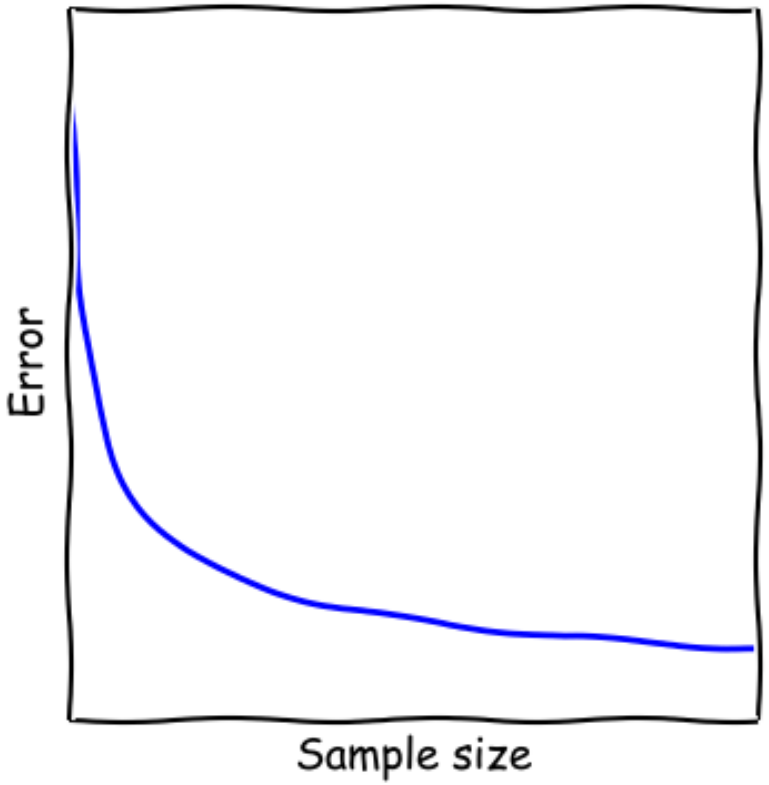
Slide: Ameet Soni

Are algorithms fair by default?

> *"After all, as the former CPD [Chicago Police Department] computer experts point out, the algorithms in themselves are neutral. 'This program had absolutely nothing to do with race… but multi-variable equations,' argues Goldstein. Meanwhile, the potential benefits of predictive policing are profound."*
>
> -Gilian Tett

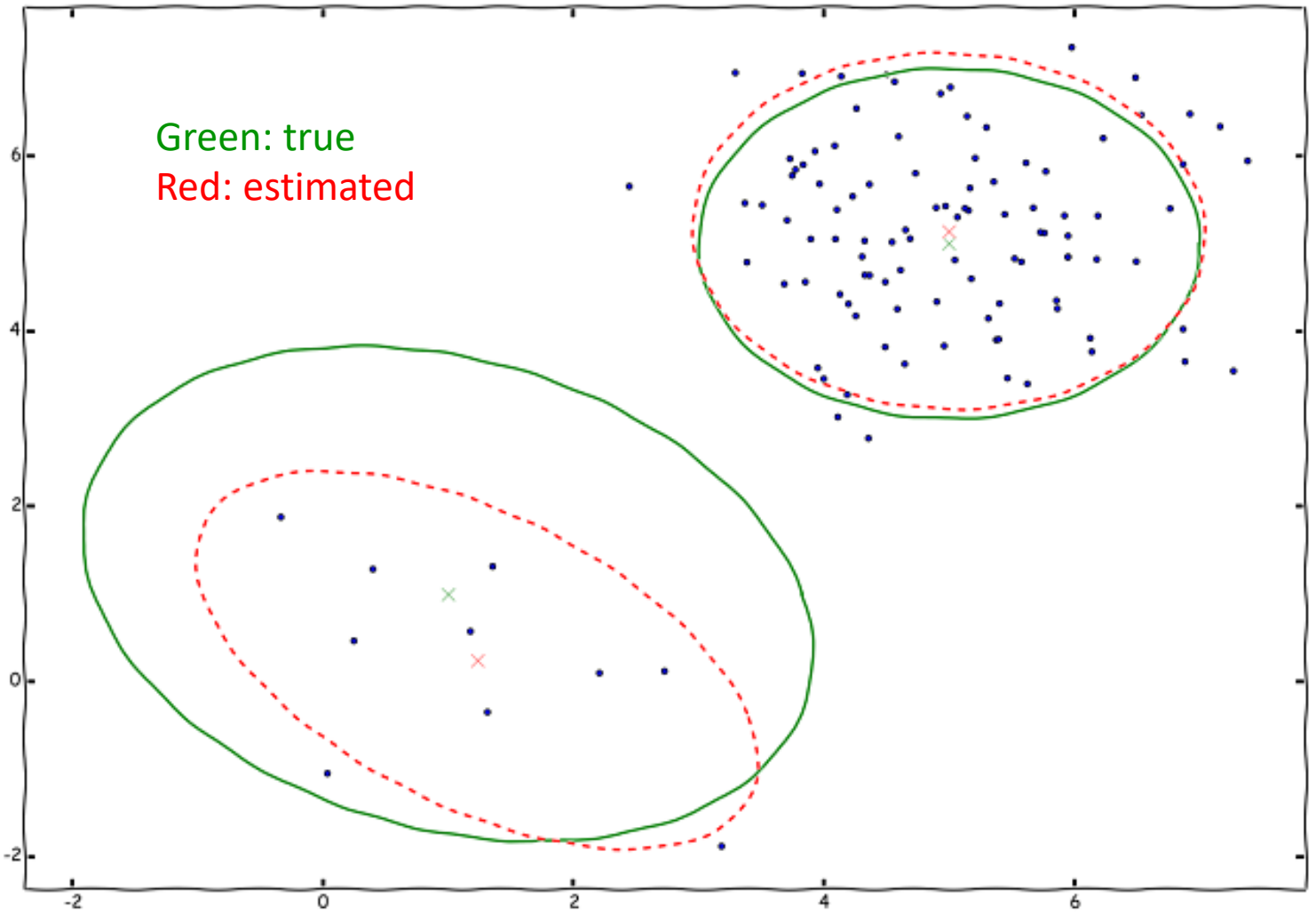
Sample size disparity

- More data from majority will make results more accurate for that group
- Less accurate for the minority



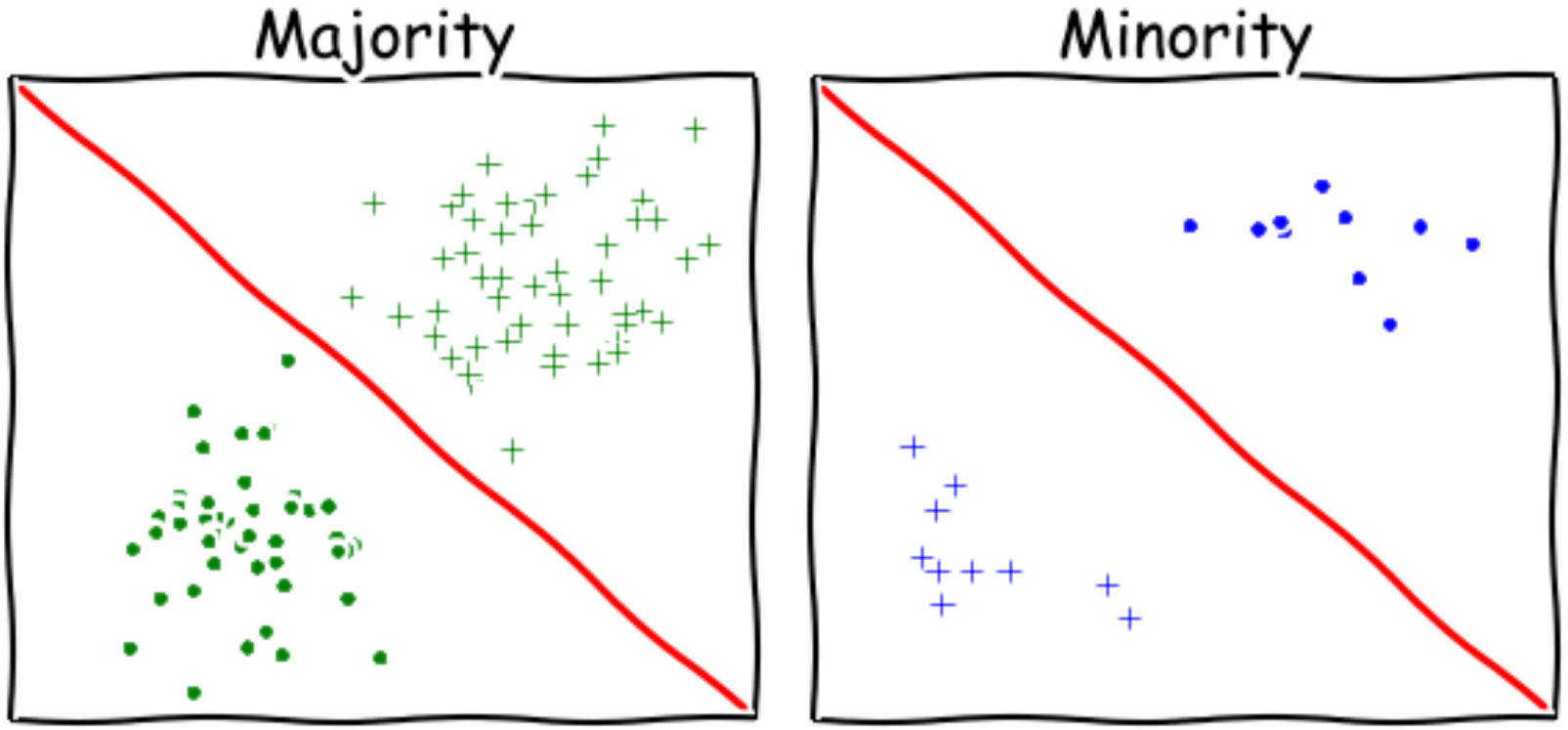
"The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate."
Image: Moritz Hardt

Sample size disparity



“Modeling a heterogeneous population as a gaussian mixture and learning its parameters using the EM algorithm. As expected, the estimates for the smaller group are significantly worse than for the larger. Dashed red ellipsoids describe the estimated covariance matrices. Solid green defines the correct covariance matrices. The green and red crosses indicate correct and estimated means, respectively.” Image: Moritz Hardt

Cultural Differences



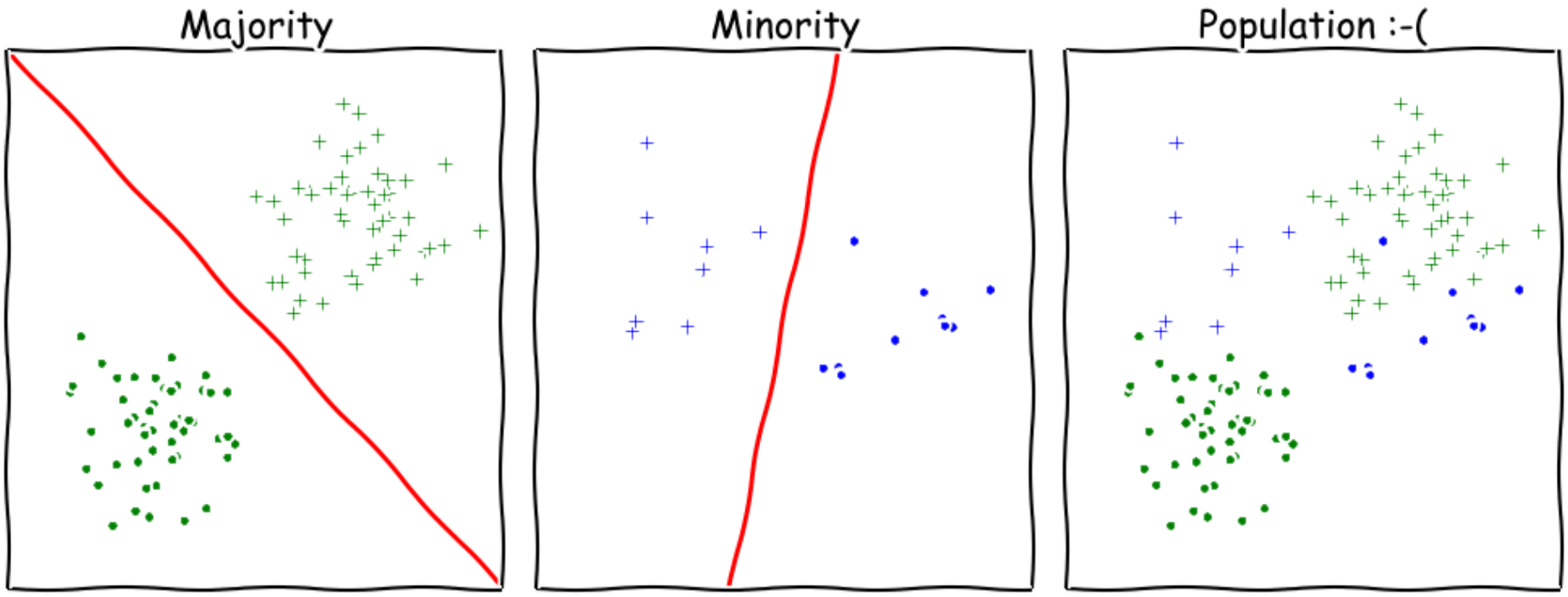
"Positively labeled examples are on opposite sides of the classifier for the two groups." Image: Moritz Hardt

Goal: determine if a user profile (on Facebook, Twitter, etc) is genuine

- positive: real profile
- negative: fake profile

Feature: length of name

Undesired Complexity

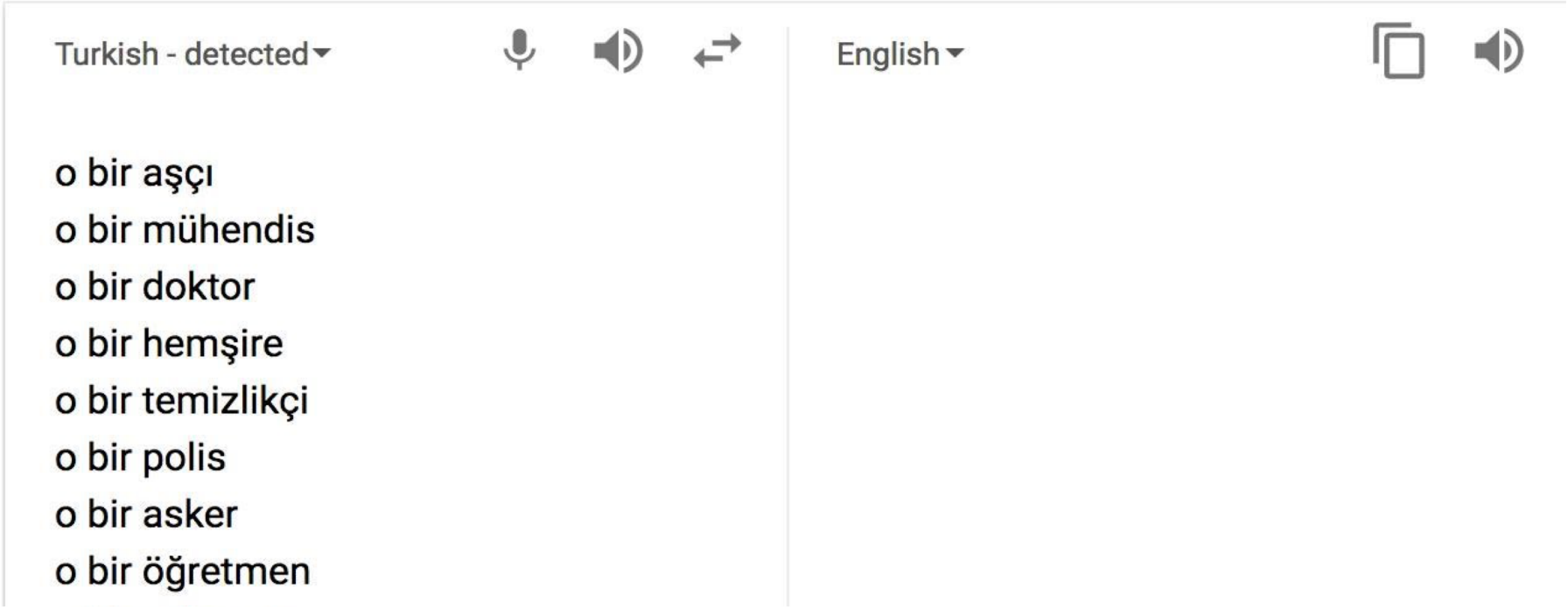


"Even if two groups of the population admit simple classifiers, the whole population may not."
Image: Moritz Hardt

“How big data is unfair” (takeaways)

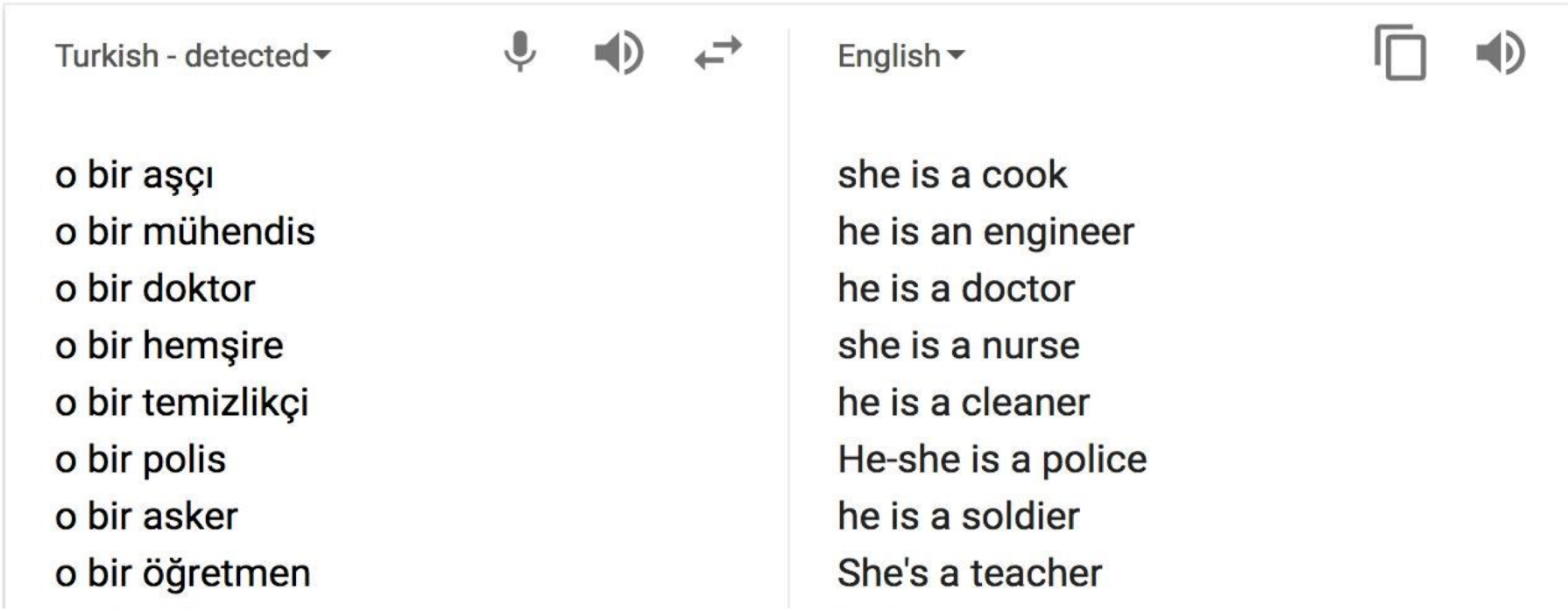
- ML is not fair by default, even though it relies on “neutral” multi-variable equations
- If training data reflects social biases, algorithm will likely incorporate them
- “Protected” attributes (race, gender, religion, sexual orientation, etc) often redundantly encoded

Example: machine translation



Slide: Ameet Soni

Example: machine translation



Slide: Ameet Soni

Challenges

Algorithms do not exist in a bubble

- Inherit the prejudices of their designers
- Reflect cultural biases
- Difficult to identify - can entrench/enhance issues
- Deny historically disadvantaged groups full participation

Slide: Ameet Soni

Outline for September 28

- Go over Lab 2
- Intro to Bayesian models
- Intro to algorithmic bias
- Redundantly encoded features and disparate impact

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

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

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Direct discrimination: C = f(X)

* Female instrumentalist not hired for orchestra
* Some ethnic groups not allowed to eat at a restaurant

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

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

* X is protected
* Y is unprotected (other features)

Goal: determine outcome C (hired, admitted, etc)

Indirect discrimination: C = f(Y)

* but strong correlation between X and Y
* Ex: housing loans
* Ex: programming experience

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

PLEASE LEAVE COMPUTERS ON

features (our X): X : protected attribute { X=0 minority group, X=1 majority group }

Y : other attributes ← modify Y

label (our y): C : binary outcome (hired, admitted), not: C=0

Disparate Impact: $P(C=1|X=0) \leq 0.8\, P(C=1|X=1)$

(legal definition) example: 40% women hired, 60% men hired

C=1, β

0.8 threshold for differences in hiring.

$0.4 \overset{?}{\leq} \underbrace{0.8(0.6)}_{0.48}$

Yes

⇒ is disparate impact.

Idea: if we can predict X from Y, could be disparate impact.

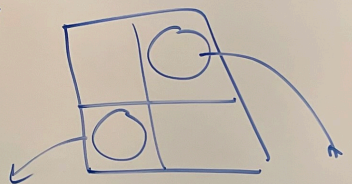
predictor / classifier (model): $f: Y \to X$

want to do our best

outcome	$X=0$	$X=1$
$C=0$	a	b
$C=1$	c	d

$\beta = \frac{c}{a+c}$

Balanced Error Rate (BER)



evaluation metric

$$\varepsilon = BER = \frac{1}{2}\left(P\left[f(Y)=0 \mid X=1 \right] + P\left[f(Y)=1 \mid X=0 \right] \right)$$

want high!

=> indicate confusion

$\varepsilon' = \frac{1}{2} - \frac{\beta}{8}$ threshold for BER

if $\varepsilon > \varepsilon'$ => no disparate impact

from us (f)

Example of repair

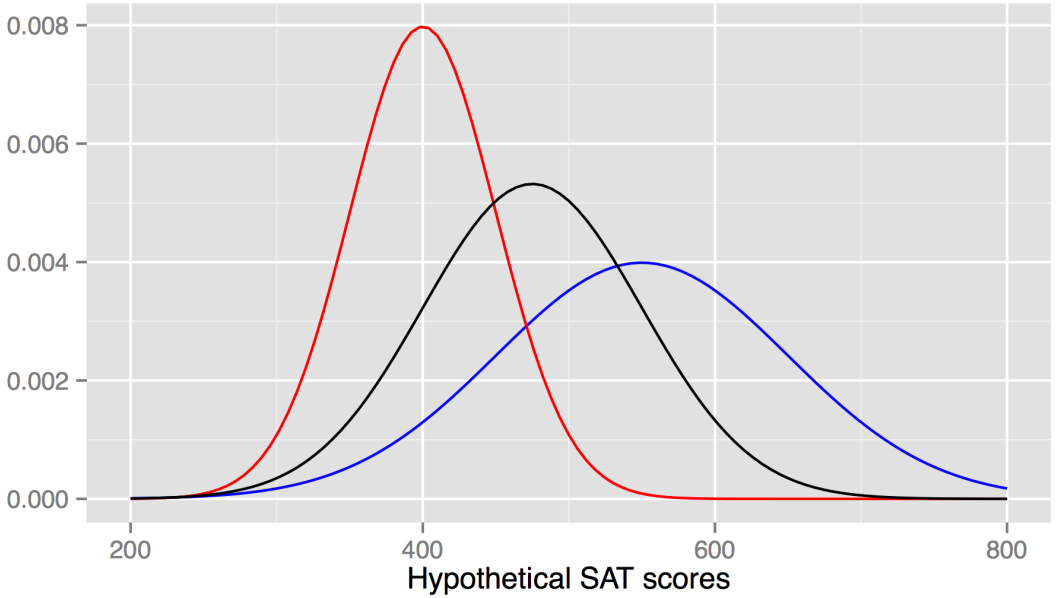


Figure 1: Consider the fake probability density functions shown here where the blue curve shows the distribution of SAT scores (Y) for $X = \texttt{female}$, with $\mu = 550, \sigma = 100$, while the red curve shows the distribution of SAT scores for $X = \texttt{male}$, with $\mu = 400, \sigma = 50$. 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.

Discussion: admissions at Haverford

- Haverford has suddenly started receiving 10x more applications than usual
- You are tasked with creating an algorithm to determine whether or not an applicant should be admitted
- Questions:
  - How would you encode features?
  - How would you use past admission data to train?
  - What loss function are you trying to optimize?