CS 260: Foundations of Data Science

Prof. Sara Mathieson

Fall 2023

HAVERFORD
COLLEGE

Admin

- **Midterm 1** due TODAY

- **Lab 5** due Wednesday after fall break
  - Naïve Bayes

- **Lab 3** grades posted

Outline for Oct 12

- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example
- Naïve Bayes implementation
- Handout 12, tennis example

Outline for Oct 12

- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example
- Naïve Bayes implementation
- Handout 12, tennis example

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

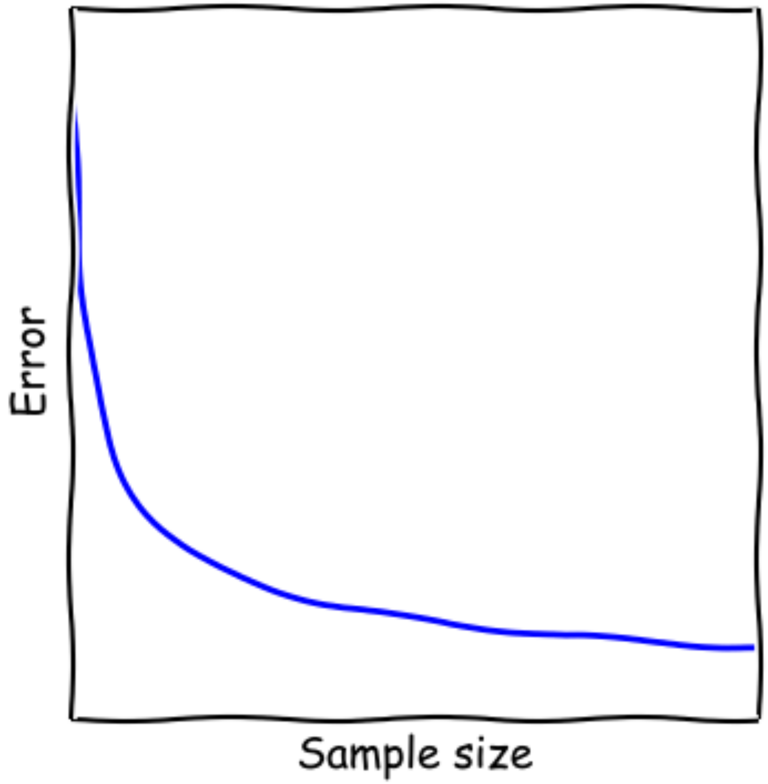
Are algorithms fair by default?

> *"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

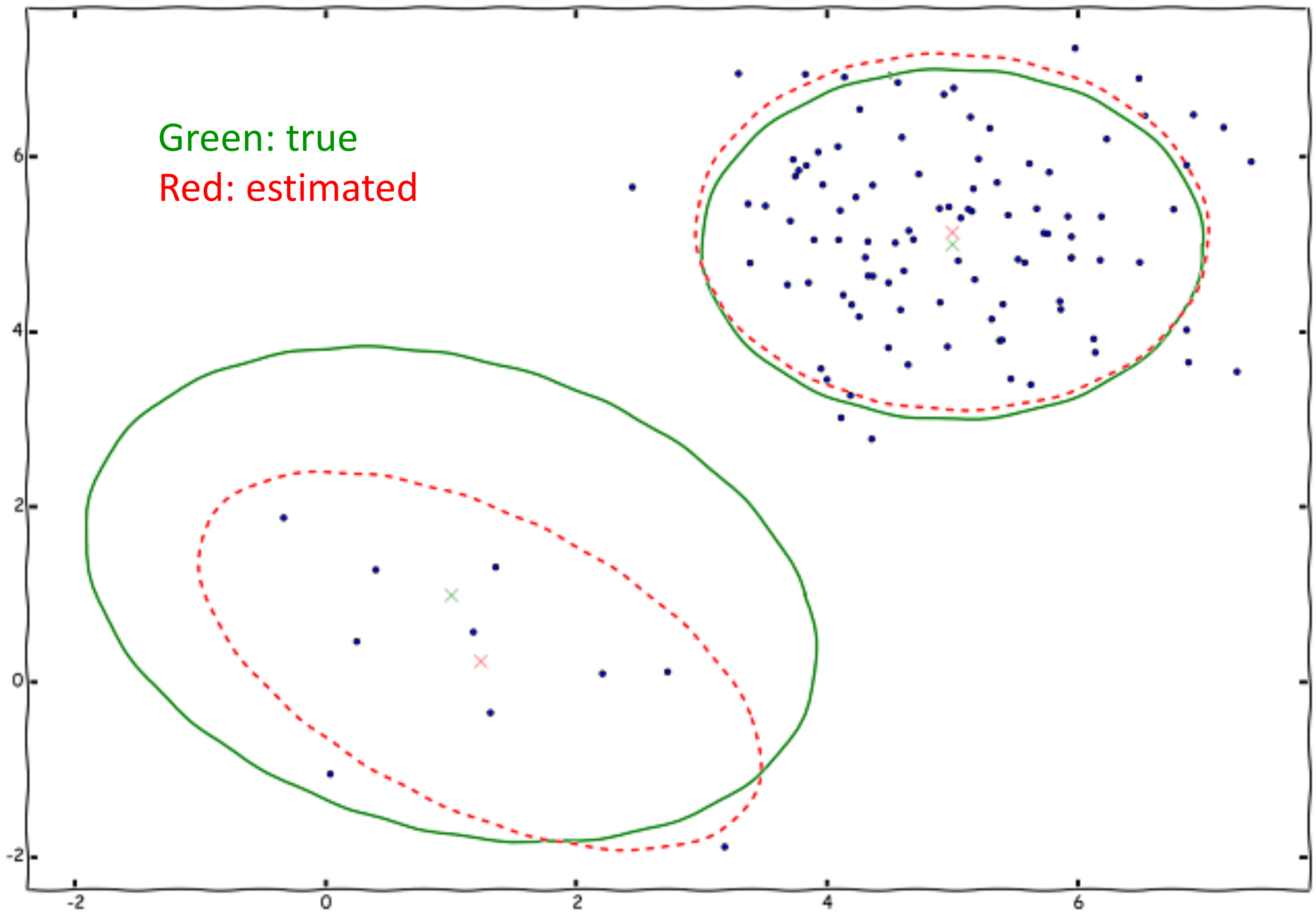
Sample size disparity

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



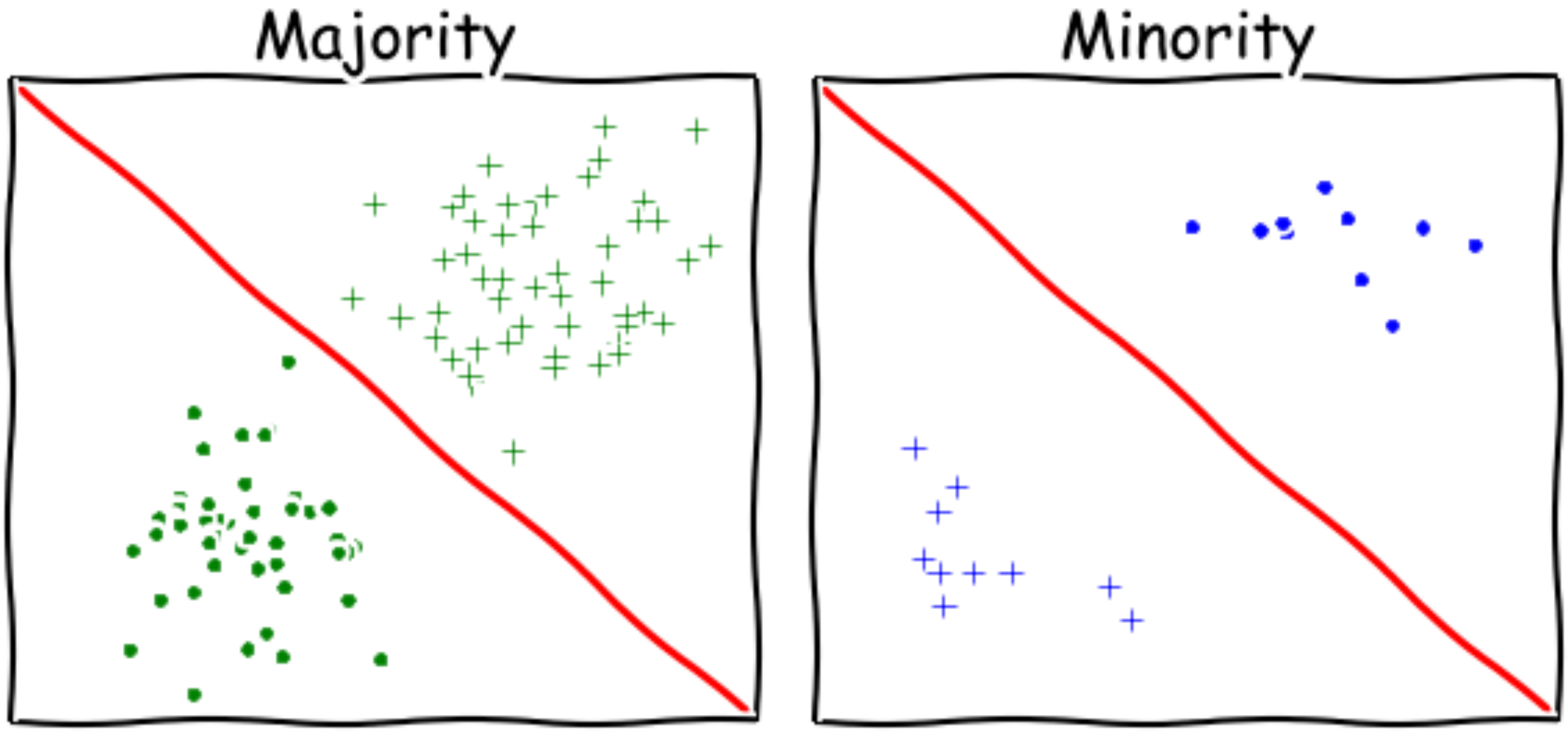
"The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate."
Image: Moritz Hardt

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



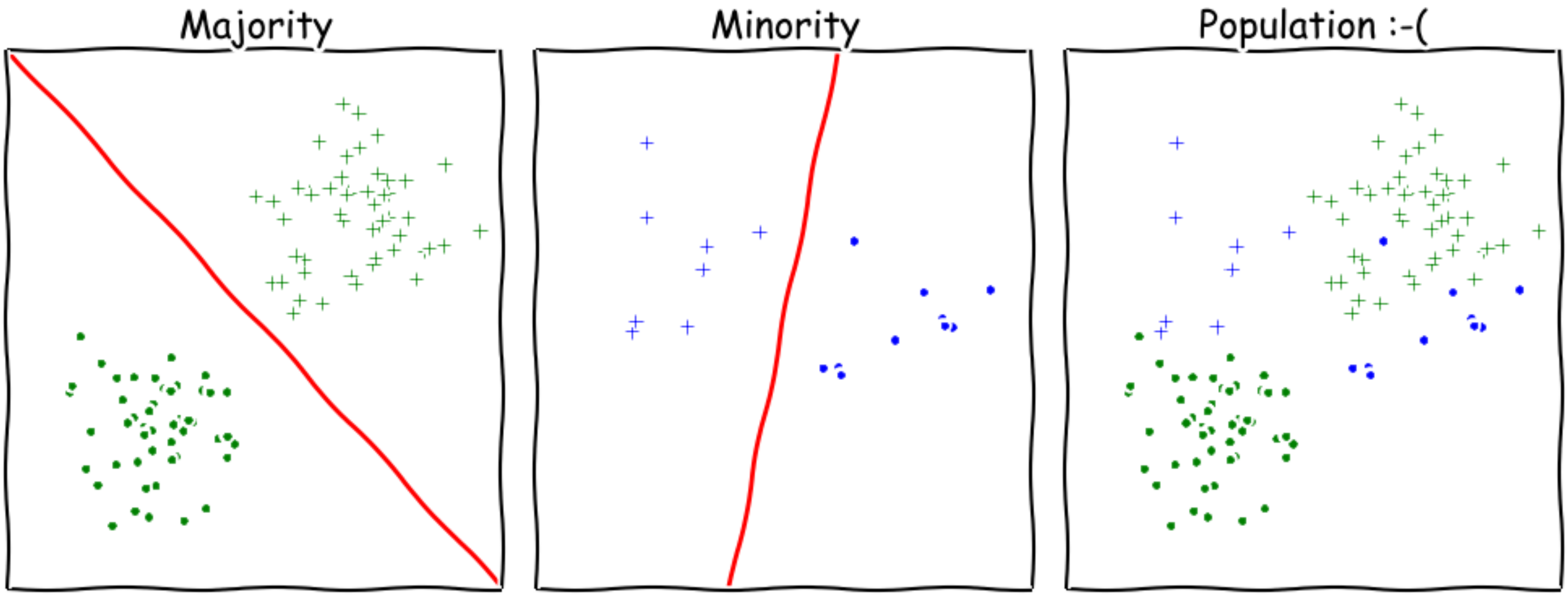
"Positively labeled examples are on opposite sides of the classifier for the two groups." Image: Moritz Hardt

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

- positive: real profile
- negative: fake profile

Feature: length of name

Undesired Complexity

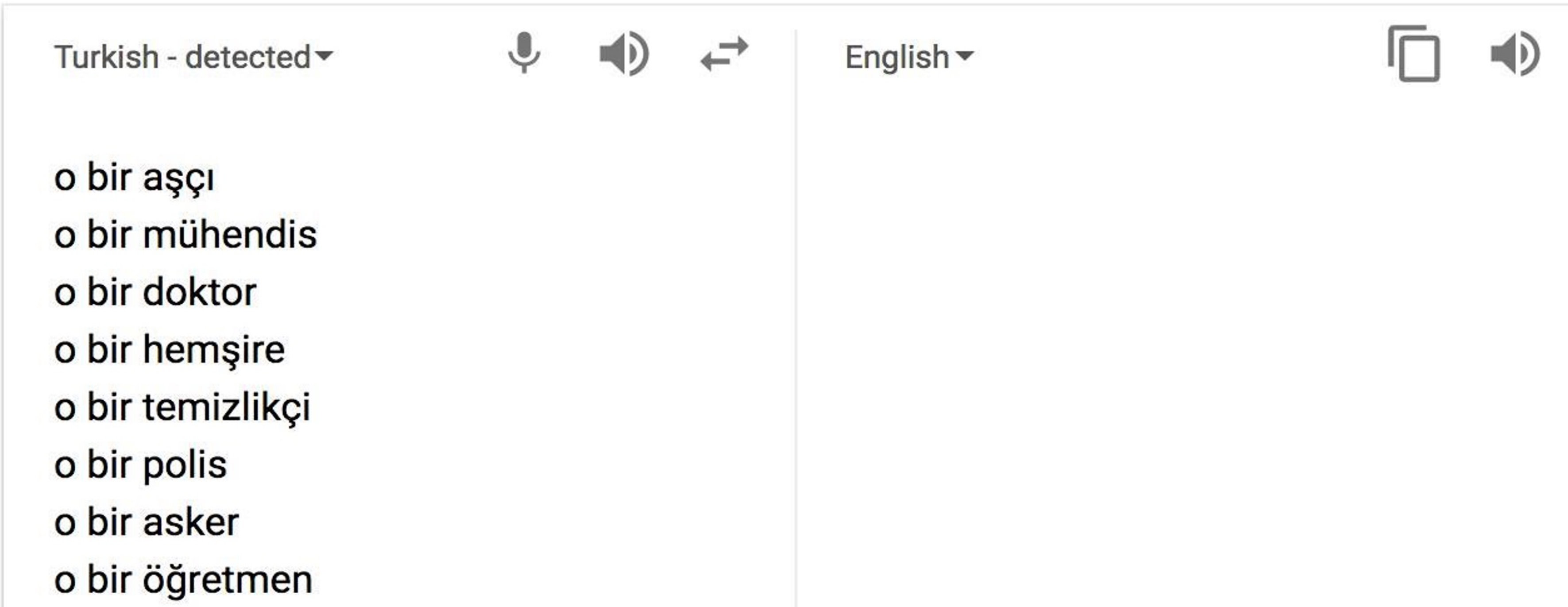


"Even if two groups of the population admit simple classifiers, the whole population may not."
Image: Moritz Hardt

"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

Turkish - detected	English
o bir aşçı	she is a cook
o bir mühendis	he is an engineer
o bir doktor	he is a doctor
o bir hemşire	she is a nurse
o bir temizlikçi	he is a cleaner
o bir polis	He-she is a police
o bir asker	he is a soldier
o bir öğretmen	She's a teacher

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 Oct 12

- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example
- Naïve Bayes implementation
- Handout 12, tennis example

How can we tell if an algorithm is biased?

D: dataset with attributes X, Y

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

“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)

“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

features

- X : protected attribute
  - 0 minority group
  - 1 majority group
- Y : other

C : binary outcome $\in \{0, 1\}$

- 0 → not hired
- 1 → hired

Disparate Impact

(legal definition)

if

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

$\Rightarrow$ disparate impact

ex

40% of women hired

30% of men hired

$0.4 \overset{?}{\not\leq} \underbrace{0.8(0.3)}_{0.24} \Rightarrow$ no

$0.4 \leq 0.8(0.6) \Rightarrow$ yes

Naive Bayes

Want high!

indicates confusion

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

predictor $f: Y \to X$

Balanced error rate BER

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

☆ ① train classifier $f(Y) \to X$

② if BER is low could be disparate impact

Outline for Oct 12

- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example

Handout 11 (#2), Handout 12 (#1)

- Naïve Bayes implementation
- Handout 12, tennis example

f_1	f_2	y
p	n	1
p	p	2
p	n	2
n	n	1
p	n	2
n	n	1
n	p	2

X

likelihood

$P(\bar{x} \mid y=1)$

feature values

names

y=1	p	n
f_1	$\frac{1+1}{3+2}$	$\frac{2+1}{3+2}$ ($\frac{3}{5}$)
f_2	$\frac{0+1}{3+2}$ ($\frac{1}{5}$)	$\frac{3+1}{3+2}$

y=2	p	n
f_1	$\frac{4}{6}$	$\frac{2}{6}$
f_2	$\frac{3}{6}$	$\frac{3}{6}$

Prior

$$\theta_1 = \frac{3+1}{7+2} = \frac{4}{9}$$

$$\theta_2 = \frac{4+1}{7+2} = \frac{5}{9}$$

add to 1

Handout 12

$\vec{x} = [\overset{f_1}{neg}, \overset{f_2}{pos}]$

① $p(\underbrace{y=1}_{\text{healthy}} \mid \vec{x}) \approx p(y=1)\, p(f_1=neg \mid y=1)\, p(f_2=pos \mid y=1)$

$\approx \frac{4}{9} \cdot \frac{3}{5} \cdot \frac{1}{5} = \frac{4}{75}$

$p(\underbrace{y=2}_{\text{disease}} \mid \vec{x}) \approx \frac{5}{9} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{54}$

max? $\frac{5}{54} \Rightarrow \boxed{\hat{y} = 2}$ ← argmax

$p(y=k$

0	1	2
$\frac{1}{3}$	$\frac{1}{5}$	$\frac{1}{2}$

2 ← argmax

$\frac{\left[\frac{4}{75}, \frac{5}{54}\right]}{\frac{4}{75} + \frac{5}{54}} \rightarrow [37\%, 63\%]$ ☆

Outline for Oct 12

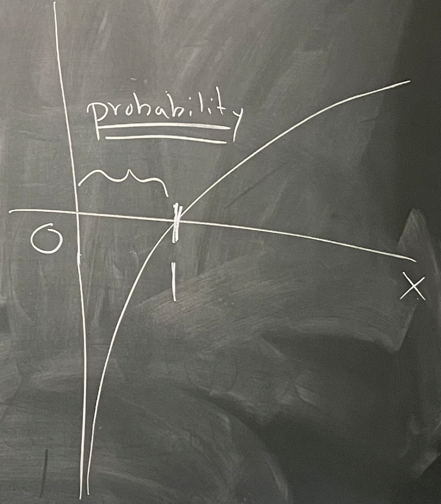
- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example
- Naïve Bayes implementation
- Handout 12, tennis example

$$\prod_{j=1}^{p} p(x_j = v \mid y = k) \approx p(y = k \mid \vec{x})$$

~~$p(\vec{x})$~~

$$P(A, B) = P(B)\,P(A \mid B)$$

$$\log\left(\frac{a}{b}\right) = \log a - \log b$$



issue is underflow!

$$\frac{1}{1000} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdot \frac{1}{10} \cdots \approx 0$$

$$\log(\Theta_k) = \log\left(\frac{N_k + 1}{n + K}\right) = \log(N_k + 1) - \log(n + K)$$

Outline for Oct 12

- Intro to Algorithmic Bias
- Disparate Impact
- Handout 11/12, clinical example
- Naïve Bayes implementation
- Handout 12, tennis example

Data Structure idea

(tennis example)

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
$\boldsymbol{x}_1$	Sunny	Hot	High	Weak	No
$\boldsymbol{x}_2$	Sunny	Hot	High	Strong	No
$\boldsymbol{x}_3$	Overcast	Hot	High	Weak	Yes
$\boldsymbol{x}_4$	Rain	Mild	High	Weak	Yes
$\boldsymbol{x}_5$	Rain	Cool	Normal	Weak	Yes
$\boldsymbol{x}_6$	Rain	Cool	Normal	Strong	No
$\boldsymbol{x}_7$	Overcast	Cool	Normal	Strong	Yes
$\boldsymbol{x}_8$	Sunny	Mild	High	Weak	No
$\boldsymbol{x}_9$	Sunny	Cool	Normal	Weak	Yes
$\boldsymbol{x}_{10}$	Rain	Mild	Normal	Weak	Yes
$\boldsymbol{x}_{11}$	Sunny	Mild	Normal	Strong	Yes
$\boldsymbol{x}_{12}$	Overcast	Mild	High	Strong	Yes
$\boldsymbol{x}_{13}$	Overcast	Hot	Normal	Weak	Yes
$\boldsymbol{x}_{14}$	Rain	Mild	High	Strong	No

Data Structure idea

(tennis example)

Condition on y=No

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
$\boldsymbol{x}_1$	Sunny	Hot	High	Weak	No
$\boldsymbol{x}_2$	Sunny	Hot	High	Strong	No
$\boldsymbol{x}_3$	Overcast	Hot	High	Weak	Yes
$\boldsymbol{x}_4$	Rain	Mild	High	Weak	Yes
$\boldsymbol{x}_5$	Rain	Cool	Normal	Weak	Yes
$\boldsymbol{x}_6$	Rain	Cool	Normal	Strong	No
$\boldsymbol{x}_7$	Overcast	Cool	Normal	Strong	Yes
$\boldsymbol{x}_8$	Sunny	Mild	High	Weak	No
$\boldsymbol{x}_9$	Sunny	Cool	Normal	Weak	Yes
$\boldsymbol{x}_{10}$	Rain	Mild	Normal	Weak	Yes
$\boldsymbol{x}_{11}$	Sunny	Mild	Normal	Strong	Yes
$\boldsymbol{x}_{12}$	Overcast	Mild	High	Strong	Yes
$\boldsymbol{x}_{13}$	Overcast	Hot	Normal	Weak	Yes
$\boldsymbol{x}_{14}$	Rain	Mild	High	Strong	No

Data Structure idea

(tennis example)

y=No (0)

outlook	**Sunny:**	**Overcast:**	**Rain:**
temperature	**Cool:**	**Mild:**	**Hot:**
humidity	**Normal:**	**High:**	
wind	**Weak:**	**Strong:**	

y=Yes (1)

outlook	**Sunny:**	**Overcast:**	**Rain:**
temperature	**Cool:**	**Mild:**	**Hot:**
humidity	**Normal:**	**High:**	
wind	**Weak:**	**Strong:**	

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?