

CS 260: Foundations of Data Science

Prof. Thao Nguyen

Fall 2025



HAVERFORD
COLLEGE

Admin

- **Sit somewhere new**
- **Lab 5** due Tuesday after Fall break (Oct 21)

Feedback form (thank you!)

General workload/difficulty/course pace

1

2 xxx

3 xxxxxxxxxxxxxxxxxxxx

4 xxxx

5 x

Feedback form (thank you!)

- What you understand well:
 - Linear regression, RSS
 - Classification, evaluation metrics
- What needs review:
 - SGD
 - Confusion matrix
 - Matrix operations

Feedback form (thank you!)

- Things that are working/helpful:
 - Handouts
 - Review, mini-quizzes
 - Labs (pair-programming is now optional)
 - Board work, examples (visuals)
 - Office/TA hours

Feedback form (thank you!)

- Not finishing handouts in class?
- Coding exercises: I will add
- Participation is useful & expected
- Lab instructions

Outline for today

- Naïve Bayes
- Intro to Algorithmic Bias
- Disparate Impact
- Ethics discussion: admissions at Haverford

Outline for today

- Naïve Bayes
- Intro to Algorithmic Bias
- Disparate Impact
- Ethics discussion: admissions at Haverford

Naïve Bayes

- Single example: $\vec{x} = [x_1, x_2, \dots, x_p]^T$
- Multi-class label: $y \in \{1, 2, \dots, K\}$
- Goal: Classification $\hat{y} = \operatorname{argmax}_{k=1, \dots, K} p(y = k | \vec{x})$

Bayesian Model

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

can ignore



**THE
EVIDENCE
 $P(B)$**

**THE
REST OF
THE MODEL**

**NAÏVE
BAYES**

Naïve Bayes

$$p(\vec{x}|y = k) = p(\underbrace{x_1}_A, \underbrace{x_2, x_3, \dots, x_p}_B | y = k)$$

$$P(A,B)=P(B)P(A|B)$$

$$= p(\underbrace{x_2, x_3, \dots, x_p}_{\substack{C \\ D}} | y = k) p(\underbrace{x_1}_A | \underbrace{x_2, \dots, x_p}_B, y = k)$$

$$= p(\underbrace{x_3, \dots, x_p}_D | y = k) p(\underbrace{x_2}_C | \underbrace{x_3, \dots, x_p}_D, y = k) \\ p(x_1 | x_2, \dots, x_p, y = k)$$

Naïve Bayes assumption

Conditional Independence: “feature j is independent from all other features given label k”

$$p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | x_1, y) \leftarrow \text{Bayes rule}$$

x_1 = 4 legs

x_2 = fur

y = cat

assume $p(x_2 | x_1, y) = p(x_2 | y)$

$$\Rightarrow p(x_1, x_2 | y) = p(x_1 | y) p(x_2 | y)$$

Naïve Bayes

$$\begin{aligned} p(\vec{x}|y = k) &= p(x_p|y = k)p(x_{p-1}|y = k) \dots p(x_2|y = k) p(x_1|y = k) \\ &= \prod_{j=1}^p p(x_j|y = k) \end{aligned}$$

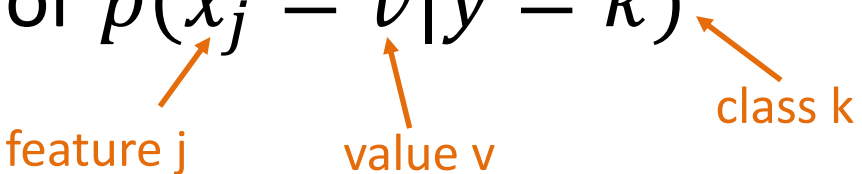
Naïve Bayes Model

$$p(y = k|\vec{x}) \propto p(y = k) \prod_{j=1}^p p(x_j|y = k)$$

proportional to

Obtaining $p(y = k)$ & $p(x_j | y = k)$


Estimate based on training data

- θ_k = estimate for $p(y = k)$
 - $\theta_{k,j,v}$ = estimate for $p(x_j = v | y = k)$
- 

Let N_k = # of examples with label k , we could define $\theta_k = \frac{N_k}{n}$

What happens if $N_k = 0$?

Laplace smoothing

- Technique to handle zero probability
- $\theta_k = \frac{N_k + 1}{n + K}$; $\sum \theta_k = \sum \frac{N_k + 1}{n + K} = \frac{1}{n + K} (n + K) = 1$
 # of classes for y
- Similarly, let $N_{k,j,v}$ = # of examples with feature j = value v and class label k

$$\theta_{k,j,v} = \frac{N_{k,j,v} + 1}{N_k + |f_j|}$$



of feature values
for feature j

Handout 11, page 1

Handout 11, page 1

\vec{x}	f_1	f_2	y
\vec{x}_1	pos	neg	1
\vec{x}_2	pos	pos	2
\vec{x}_3	pos	neg	2
\vec{x}_4	neg	neg	1
\vec{x}_5	pos	neg	2
\vec{x}_6	neg	neg	1
\vec{x}_7	neg	pos	2

$$\Theta_1 = \frac{3 + 1}{7 + 2}$$
$$= \boxed{\frac{4}{9}}$$
$$\Theta_2 = \frac{4 + 1}{7 + 2}$$
$$= \boxed{\frac{5}{9}}$$

Handout 11, page 1

②

$y=1$	pos	neg
f_1	$\frac{1+1}{3+2}$	$\frac{2+1}{3+2}$
f_2	$\frac{0+1}{3+2}$	$\frac{3+1}{3+2}$

$y=2$	pos	neg
f_1	$\frac{4}{6}$	$\frac{2}{6}$
f_2	$\frac{3}{6}$	$\frac{3}{6}$

f_1	f_2	y
pos	neg	1
pos	pos	2
pos	neg	2
neg	neg	1
pos	neg	2
neg	neg	1
neg	pos	2

purpose of Laplace!

no zero probabilities
(because multiplication)

Outline for today

- Naïve Bayes
- **Intro to Algorithmic Bias**
- Disparate Impact
- Ethics discussion: admissions at Haverford

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

```
3 model = initialization(...)
4 n_epochs = ...
5 train_data = ...
6 for i in n_epochs:
7     train_data = shuffle(train_data)
8     X, y = split(train_data)
9     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](#)

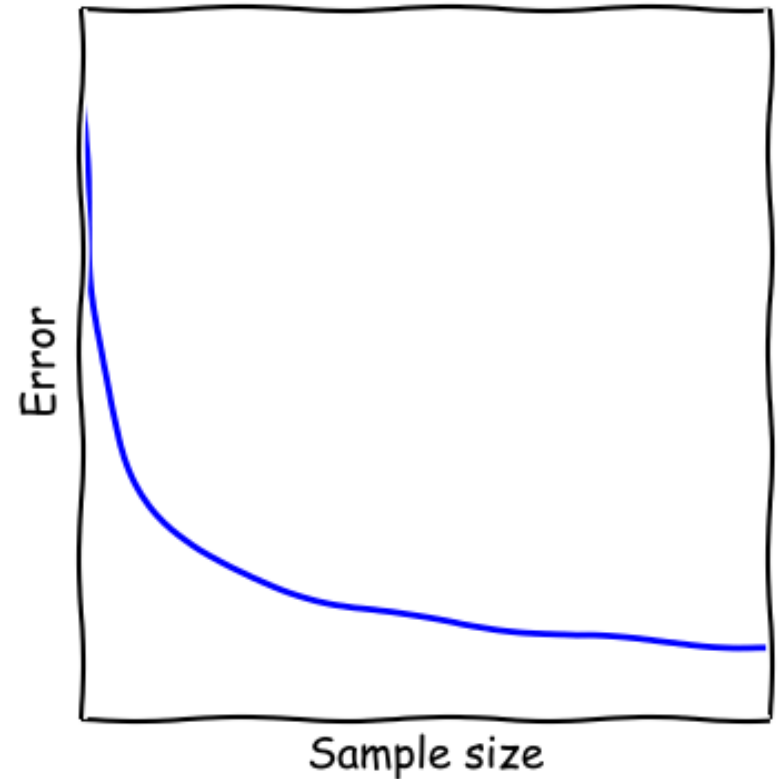
Are algorithms fair by default?

“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

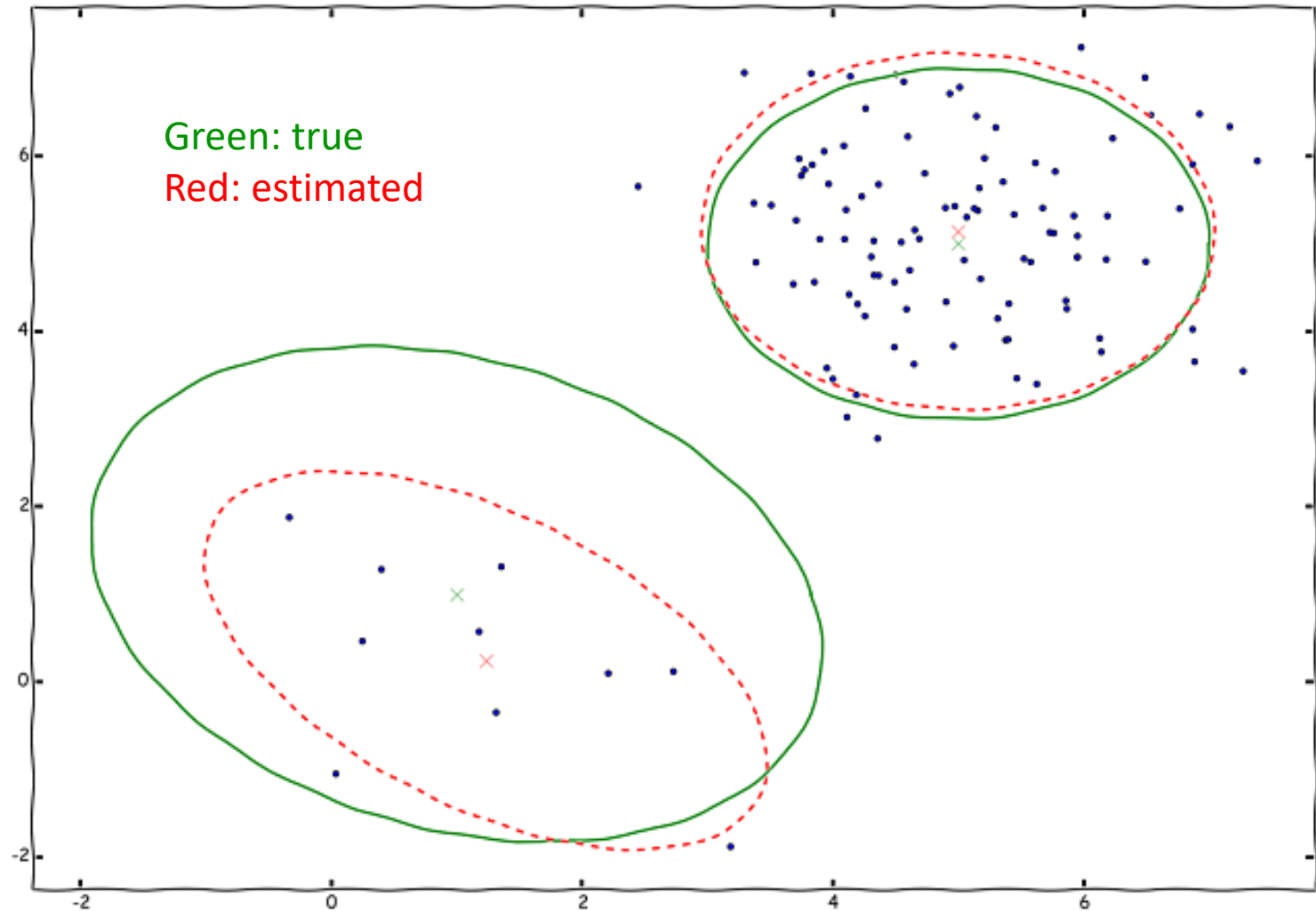
- More data from majority will make results more accurate for that group
- Less accurate for the minority



“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.”

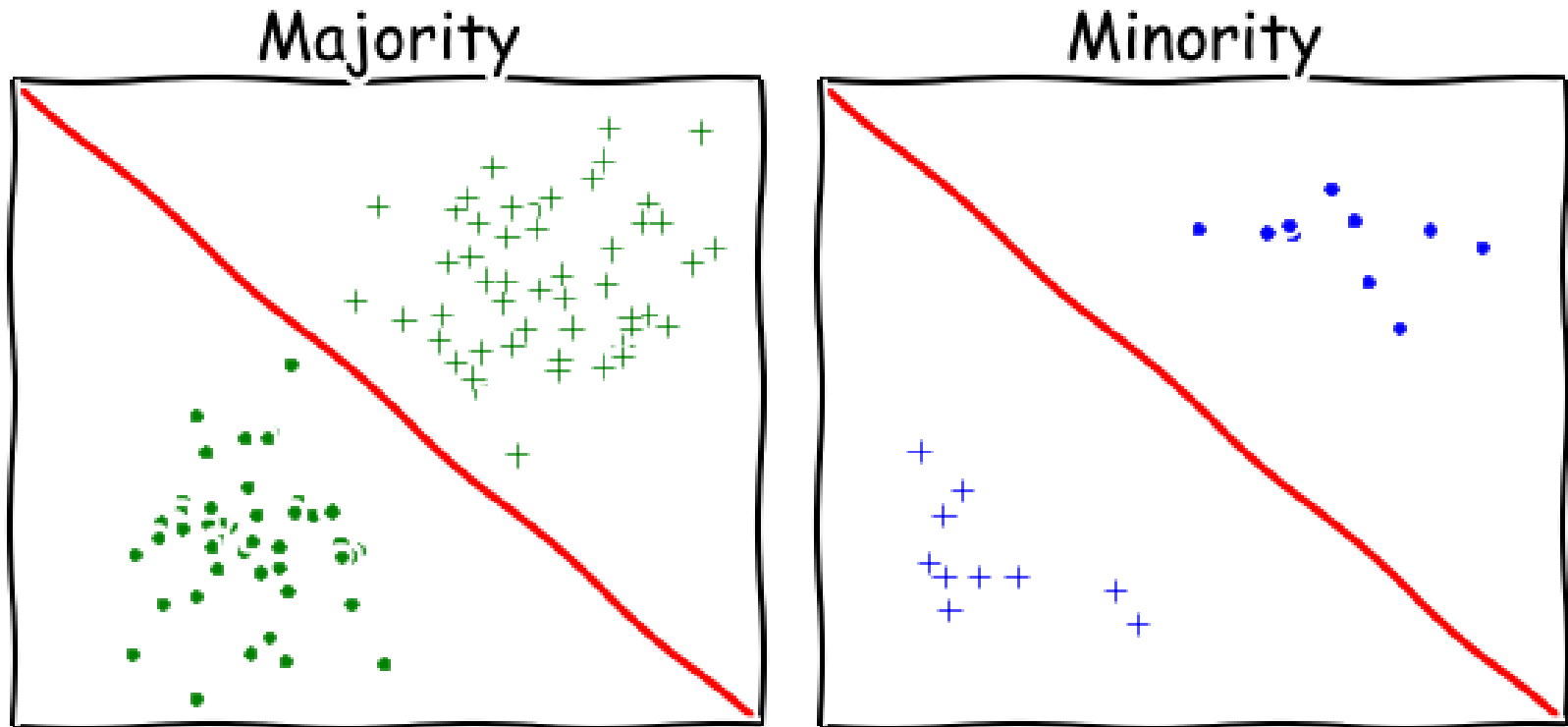
Image: Moritz Hardt

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



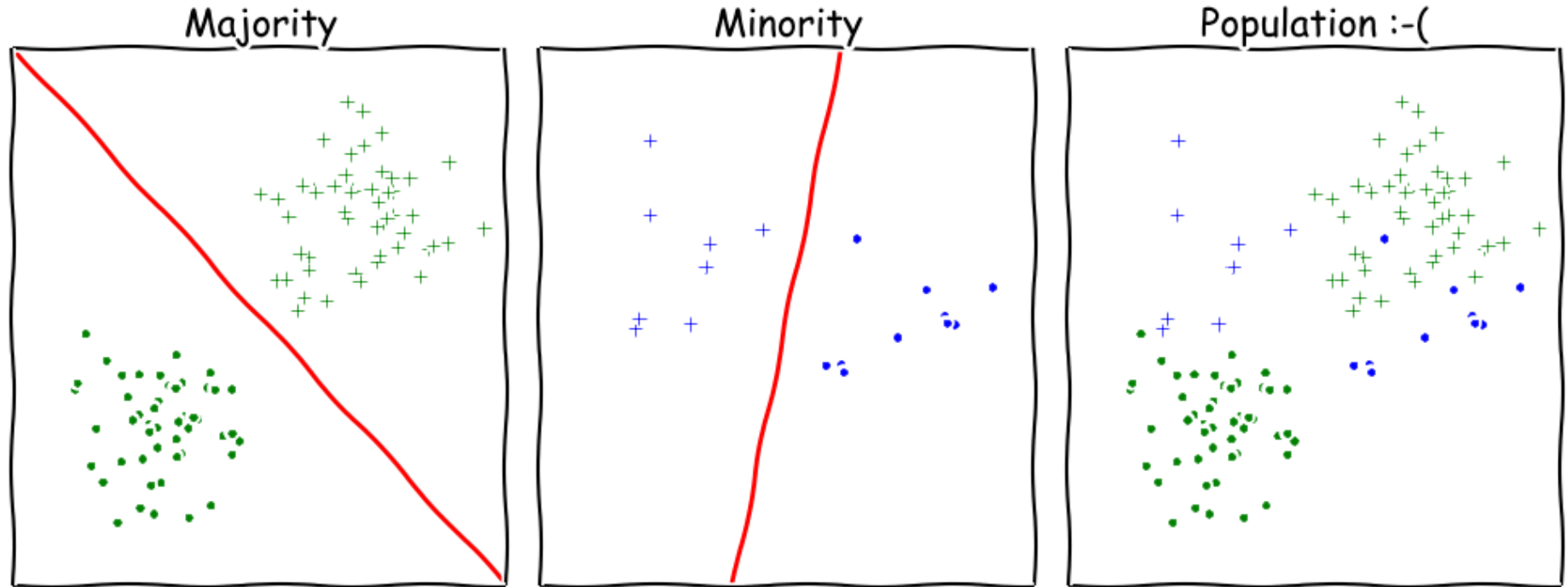
“Positively labeled examples are on opposite sides of the classifier for the two groups.” Image: Moritz Hardt

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 reflect social biases, algorithms will likely incorporate them
- “Protected” attributes (race, gender, religion, sexual orientation, etc.) are often redundantly encoded



Example: machine translation

Turkish - detected ▾

o bir aşçı
o bir mühendis
o bir doktor
o bir hemşire
o bir temizlikçi
o bir polis
o bir asker
o bir öğretmen



English ▾



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

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

Outline for today

- Naïve Bayes
- Intro to Algorithmic Bias
- **Disparate Impact**
- Ethics discussion: admissions at Haverford

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.)

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

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

Disparate Impact

features

- X: protected attributes
- Y: other attributes
- C: binary outcome $\in \{0,1\}$

0	minority group
1	majority group

not hired hired

Legal definition

If $P(C = 1|X = 0) < 0.8 * P(C = 1|X = 1)$

\Rightarrow disparate impact

Example: 40% of female applicants hired

30% of male applicants hired

Checking for Disparate Impact

- **Idea:** if we can predict X (protected attribute) from Y (other attributes), this could lead to disparate impact.

- **Metric:** Balanced Error Rate (BER) indicates confusion

$$BER = \frac{1}{2} (P(f(Y) = 0 | X = 1) + P(f(Y) = 1 | X = 0))$$

- 1) Train classifier $f(Y) \rightarrow X$
- 2) Calculate BER, low BER could imply disparate impact

Example of repair

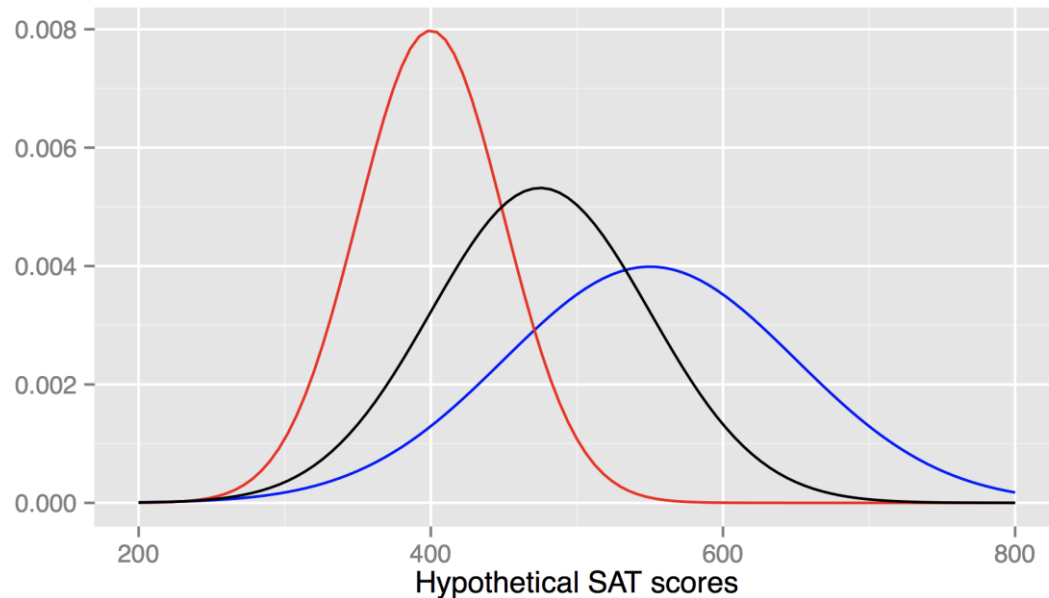


Figure 1: Consider the fake probability density functions shown here where the blue curve shows the distribution of SAT scores (Y) for $X = \text{female}$, with $\mu = 550, \sigma = 100$, while the red curve shows the distribution of SAT scores for $X = \text{male}$, with $\mu = 400, \sigma = 50$. The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

Handout 11, page 2

Handout 11, page 2

Handout 10

$$\vec{x} = [\text{neg}, \text{pos}]$$

$$\begin{aligned} \textcircled{1} \quad p(y=1|\vec{x}) &\approx p(y=1) p(f_1=\text{neg}|y=1) p(f_2=\text{pos}|y=1) \\ &\approx \frac{4}{9} \cdot \frac{3}{5} \cdot \frac{1}{5} = \frac{4}{75} \end{aligned}$$

$$\begin{aligned} p(y=2|\vec{x}) &\approx p(y=2) p(f_1=\text{neg}|y=2) p(f_2=\text{pos}|y=2) \\ &\approx \frac{5}{9} \cdot \frac{1}{3} \cdot \frac{1}{2} = \frac{5}{54} \end{aligned}$$

$$\frac{p(\vec{x})}{p(\vec{x})} \begin{matrix} \text{posterior} \\ \begin{matrix} 1 & 2 \\ \frac{4}{75} & \frac{5}{54} \end{matrix} \end{matrix}$$
$$\text{argmax}\left(\left[\frac{4}{75}, \frac{5}{54}\right]\right)$$

$$\star \boxed{\hat{y}=2} \star (\text{disease})$$

$$\Rightarrow \text{normalize}$$

$$\begin{matrix} \text{posterior} \\ \left[0.37, 0.63\right] \end{matrix}$$

Handout 11, page 2

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
x_1	Sunny	Hot	High	Weak	No
x_2	Sunny	Hot	High	Strong	No
x_3	Overcast	Hot	High	Weak	Yes
x_4	Rain	Mild	High	Weak	Yes
x_5	Rain	Cool	Normal	Weak	Yes
x_6	Rain	Cool	Normal	Strong	No
x_7	Overcast	Cool	Normal	Strong	Yes
x_8	Sunny	Mild	High	Weak	No
x_9	Sunny	Cool	Normal	Weak	Yes
x_{10}	Rain	Mild	Normal	Weak	Yes
x_{11}	Sunny	Mild	Normal	Strong	Yes
x_{12}	Overcast	Mild	High	Strong	Yes
x_{13}	Overcast	Hot	Normal	Weak	Yes
x_{14}	Rain	Mild	High	Strong	No

Handout 11, page 2

Condition on $y=\text{No}$

Day	Outlook	Temperature	Humidity	Wind	PlayTennis (y)
x_1	Sunny	Hot	High	Weak	No
x_2	Sunny	Hot	High	Strong	No
x_3	Overcast	Hot	High	Weak	Yes
x_4	Rain	Mild	High	Weak	Yes
x_5	Rain	Cool	Normal	Weak	Yes
x_6	Rain	Cool	Normal	Strong	No
x_7	Overcast	Cool	Normal	Strong	Yes
x_8	Sunny	Mild	High	Weak	No
x_9	Sunny	Cool	Normal	Weak	Yes
x_{10}	Rain	Mild	Normal	Weak	Yes
x_{11}	Sunny	Mild	Normal	Strong	Yes
x_{12}	Overcast	Mild	High	Strong	Yes
x_{13}	Overcast	Hot	Normal	Weak	Yes
x_{14}	Rain	Mild	High	Strong	No

Handout 11, page 2

y=No (0)

outlook	Sunny: 4	Overcast: 1	Rain: 3	/8
temperature	Cool: 2	Mild: 3	Hot: 3	/8
humidity	Normal: 2	High: 5		/7
wind	Weak: 3	Strong: 4		/7

y=Yes (1)

outlook	Sunny: 3	Overcast: 5	Rain: 4	/12
temperature	Cool: 4	Mild: 5	Hot: 3	/12
humidity	Normal: 7	High: 4		/11
wind	Weak: 7	Strong: 4		/11

Outline for today

- Naïve Bayes
- Intro to Algorithmic Bias
- Disparate Impact
- Ethics discussion: admissions at Haverford

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?