CS 66: Machine Learning

Prof. Sara Mathieson

Spring 2019

SWARTHMORE

Outline for May 1

- Introduction to bias in ML
- Thought experiment: admissions at Swarthmore
- Removing disparate impact
- Friday: big picture questions and discussion

  - Project check-in during lab today
  - Hand back exam on Friday

Outline for May 1

- Introduction to bias in ML
- Thought experiment: admissions at Swarthmore
- Removing disparate impact
- Friday: big picture questions and discussion

Article 1 takeaways

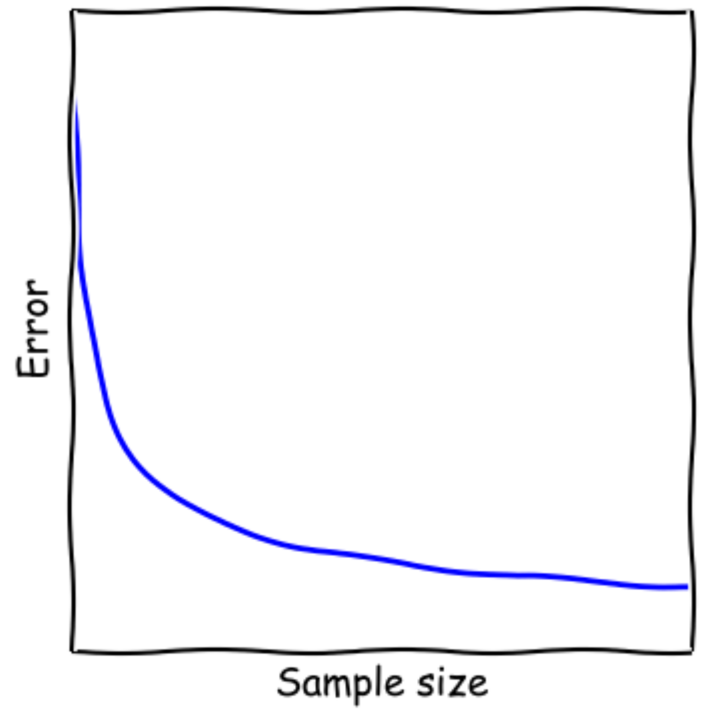
- ML is not fair by default, even though it relies on “neutral” multi-variable equations

- If training data reflects social biases, algorithm will likely incorporate them

- “Protected” attributes (race, gender, religion, sexual orientation, etc) often redundantly encoded

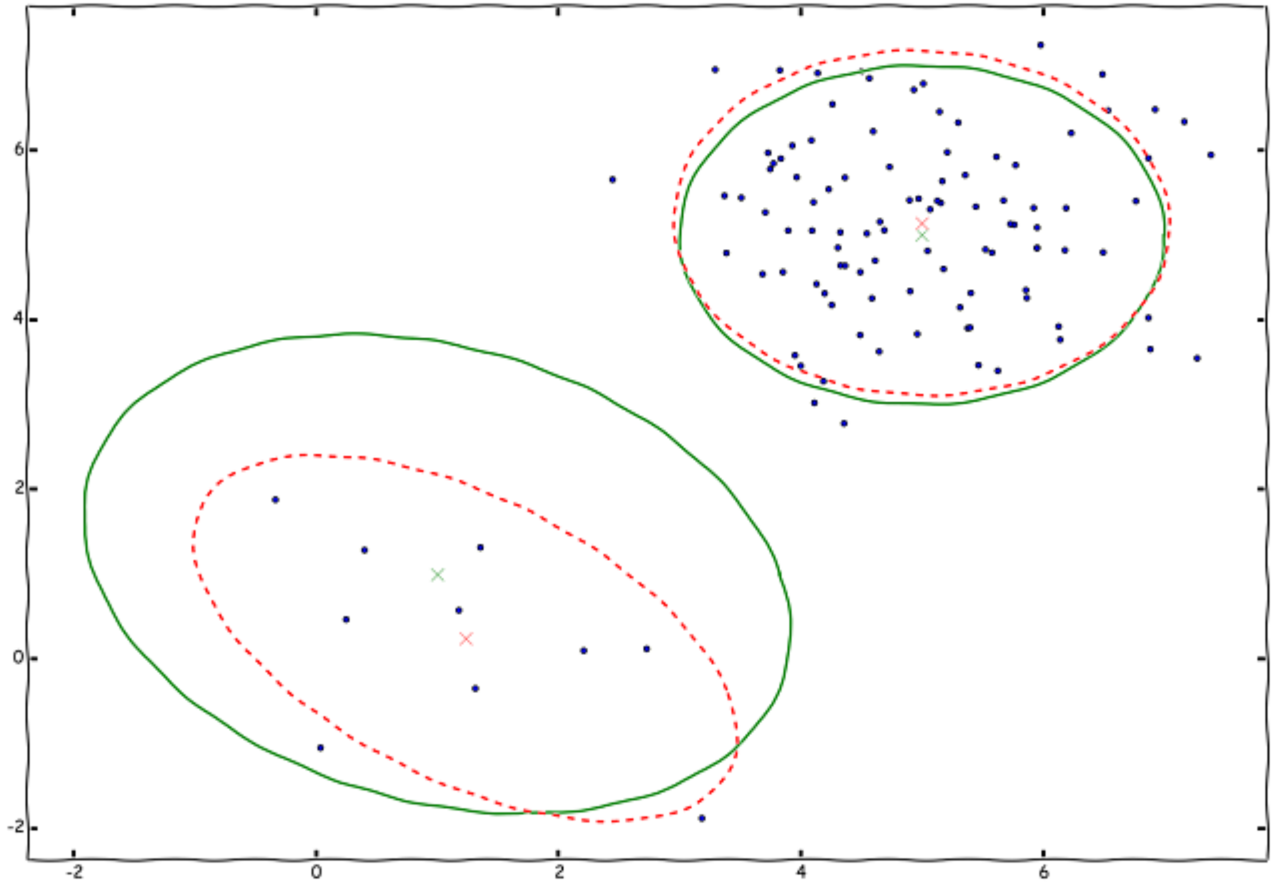
Sample size disparity

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



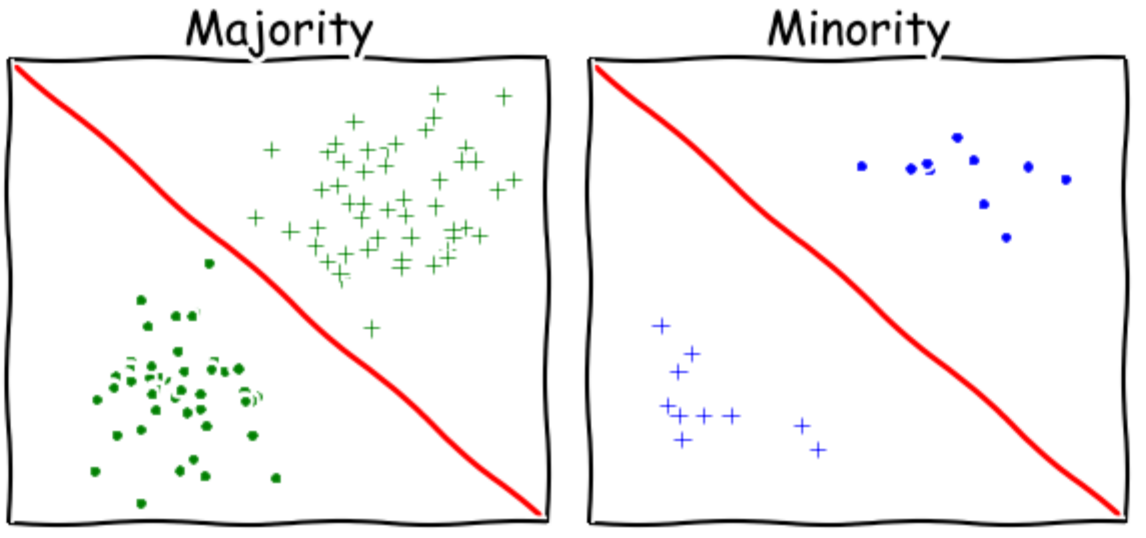
"The error of a classifier often decreases as the inverse square root of the sample size. Four times as many samples means halving the error rate."
Image: Moritz Hardt

Sample size disparity



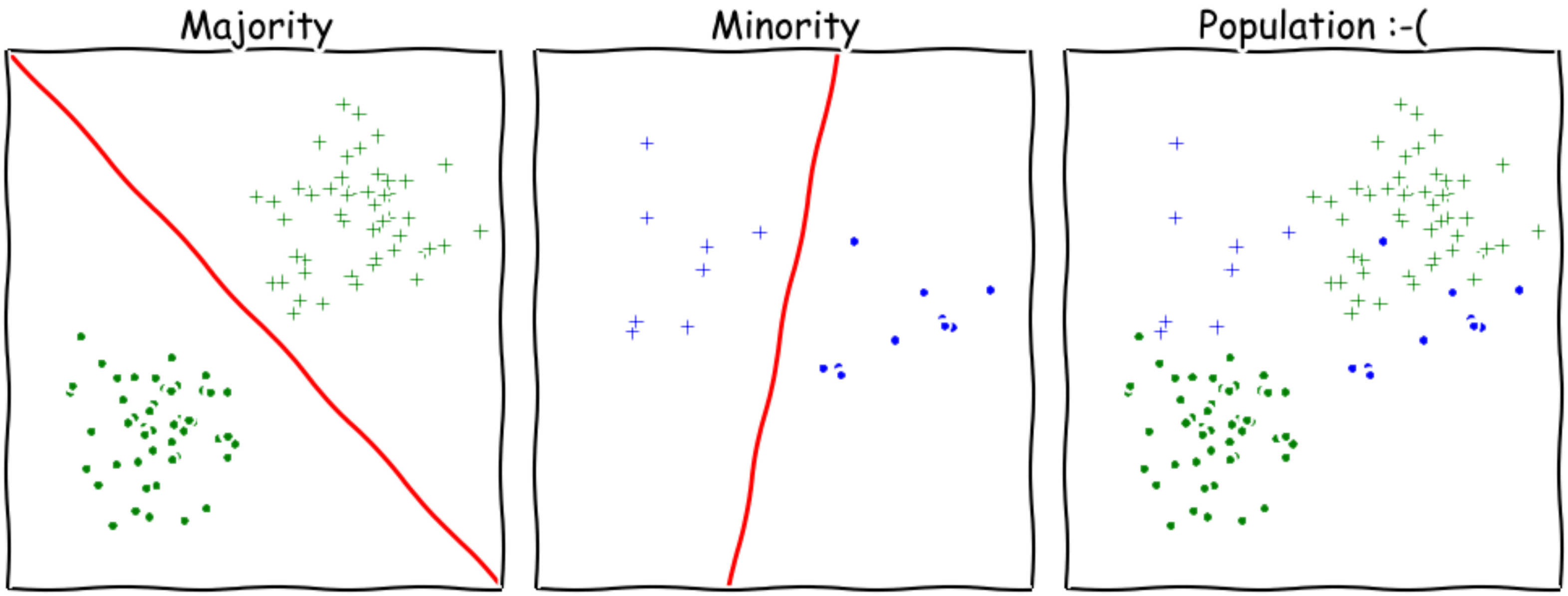
“Modeling a heterogeneous population as a gaussian mixture and learning its parameters using the EM algorithm. As expected, the estimates for the smaller group are significantly worse than for the larger. Dashed red ellipsoids describe the estimated covariance matrices. Solid green defines the correct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

Undesired complexity



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

Examples

- Many cameras and webcams have not been trained with racial diversity in mind

  http://content.time.com/time/business/article/0,8599,1954643,00.html

- Prestigious job ads automatically shown to men but not women

  https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/

- Housing loans (mortgages) given/denied automatically; correlate with neighborhoods and race

  https://www.brookings.edu/research/credit-denial-in-the-age-of-ai/

- Predictive policing

  https://www.theverge.com/2014/2/19/5419854/the-minority-report-this-computer-predicts-crime-but-is-it-racist

Examples modified from: Suresh Venkatasubramanian

Propublica, *Machine Bias*

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Example from: Ameet Soni

Word-embedding examples

Table 1. Summary of Word-Embedding Association Tests. We replicated eight well-known IAT findings using word embeddings (rows 1 to 3 and 6 to 10); we also help explain prejudiced human behavior concerning hiring in the same way (rows 4 and 5). Each result compares two sets of words from target concepts about which we are attempting to learn with two sets of attribute words. In each case, the first target is found compatible with the first attribute, and the second target with the second attribute. Throughout, we use word lists from the studies we seek to replicate. N, number of subjects; N_T, number of target words; N_A, number of attribute words. We report the effect sizes (d) and P values (P, rounded up) to emphasize that the statistical and substantive significance of both sets of results is uniformly high; we do not imply that our numbers are directly comparable with those of human studies. For the online IATs (rows 6, 7, and 10), P values were not reported but are known to be below the significance threshold of 10^{-2}. Rows 1 to 8 are discussed in the text; for completeness, this table also includes the two other IATs for which we were able to find suitable word lists (rows 9 and 10). We found similar results with word2vec, another algorithm for creating word embeddings, trained on a different corpus, Google News (see the supplementary materials).

Target words	Attribute words	Original finding				Our finding			
		Ref.	N	d	P	N_T	N_A	d	P
Flowers vs. insects	Pleasant vs. unpleasant	*(5)*	32	1.35	10^{-8}	25 × 2	25 × 2	1.50	10^{-7}
Instruments vs. weapons	Pleasant vs. unpleasant	*(5)*	32	1.66	10^{-10}	25 × 2	25 × 2	1.53	10^{-7}
European-American vs. African-American names	Pleasant vs. unpleasant	*(5)*	26	1.17	10^{-5}	32 × 2	25 × 2	1.41	10^{-8}
European-American vs. African-American names	Pleasant vs. unpleasant from *(5)*	*(7)*	Not applicable			16 × 2	25 × 2	1.50	10^{-4}
European-American vs. African-American names	Pleasant vs. unpleasant from *(9)*	*(7)*	Not applicable			16 × 2	8 × 2	1.28	10^{-3}
Male vs. female names	Career vs. family	*(9)*	39k	0.72	$<10^{-2}$	8 × 2	8 × 2	1.81	10^{-3}
Math vs. arts	Male vs. female terms	*(9)*	28k	0.82	$<10^{-2}$	8 × 2	8 × 2	1.06	.018
Science vs. arts	Male vs. female terms	*(10)*	91	1.47	10^{-24}	8 × 2	8 × 2	1.24	10^{-2}
Mental vs. physical disease	Temporary vs. permanent	*(23)*	135	1.01	10^{-3}	6 × 2	7 × 2	1.38	10^{-2}
Young vs. old people's names	Pleasant vs. unpleasant	*(9)*	43k	1.42	$<10^{-2}$	8 × 2	8 × 2	1.21	10^{-2}

Article 2: **Semantics derived automatically from language corpora contain human-like biases**

Aylin Caliskan,[1*] Joanna J. Bryson,[1,2*] Arvind Narayanan[1*]

Outline for May 1

- Introduction to bias in ML
- Thought experiment: admissions at Swarthmore
- Removing disparate impact
- Friday: big picture questions and discussion

Admissions at Swarthmore

- Swarthmore has suddenly started receiving 10x more applications than usual
- You are tasked with creating a Machine Learning 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?

features

- essay, activities
- SAT, GPA

training

- Decision Stump
- high school differences (geographic)

loss

- institutional needs
- not <u>all</u> history
- donors?
- order & choose less similar
- not supervised

Outline for May 1

- Introduction to bias in ML
- Thought experiment: admissions at Swarthmore
- Removing disparate impact
- Friday: big picture questions and discussion

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

X: protected $\begin{cases} X=0 & \text{minority} \\ X=1 & \text{majority} \end{cases}$

Y: unprotected

C: binary outcome $\begin{cases} C=0 & \text{not hired} \\ C=1 & \text{hired} \end{cases}$

<u>Disparate Impact</u> (legal definition)

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

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

① train classifier $f: Y \to X$ } do well

② Balanced error rate of $\underline{f}$

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

want high! but less than $\frac{1}{2}$

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

③ compu

compu

$\underline{f}$

$[f(Y)=1 \mid X=0]$

ε

③ compute $\beta = \dfrac{c}{c+a}$

compute $\varepsilon' = \dfrac{1}{2} - \dfrac{\beta}{8}$

error threshold

if $\varepsilon' < \varepsilon$

$\Rightarrow$ no <u>disparate impact</u>

Example of repair

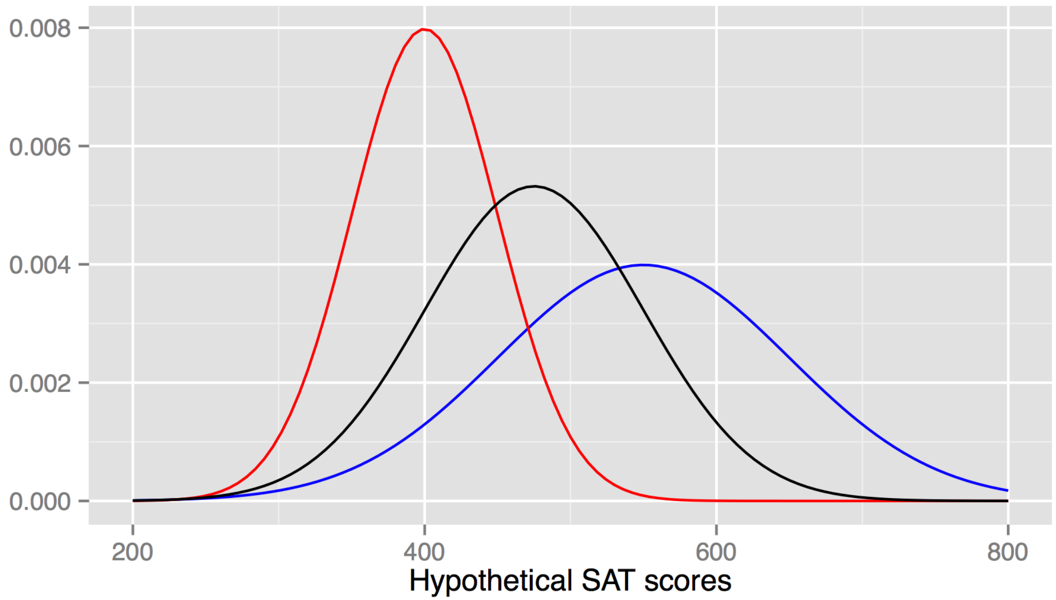


Figure 1: Consider the fake probability density functions shown here where the blue curve shows the distribution of SAT scores (Y) for $X = \texttt{female}$, with $\mu = 550, \sigma = 100$, while the red curve shows the distribution of SAT scores for $X = \texttt{male}$, with $\mu = 400, \sigma = 50$.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

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

Outline for May 1

- Introduction to bias in ML
- Thought experiment: admissions at Swarthmore
- Removing disparate impact
- Friday: big picture questions and discussion

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?