CS 360: Machine Learning

Prof. Sara Mathieson

Fall 2019

HAVERFORD
COLLEGE

Admin

- Last office hours (in my office)
  - **Friday: 4-5pm**
  - **Tuesday Dec 17: 4-5pm**

- In lab today: **project meetings** with all groups

- Project presentations: **Wed, Dec 18, 1-4pm**
  - Room: KINSC S430
  - Everyone should ask one question!

CIFAR-10 competition results

- Takeaways:
  - Regularization
  - Batch normalization
  - Dropout
- Gareth: 91% (train), 79% (test)
- Jiaping: 92% (train), 73% (test)
- Honorable mention: Emily, Jocelyn

Outline for December 12

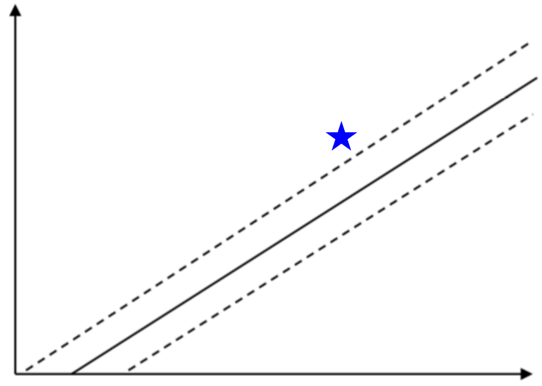
- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- Final thoughts

Outline for December 12

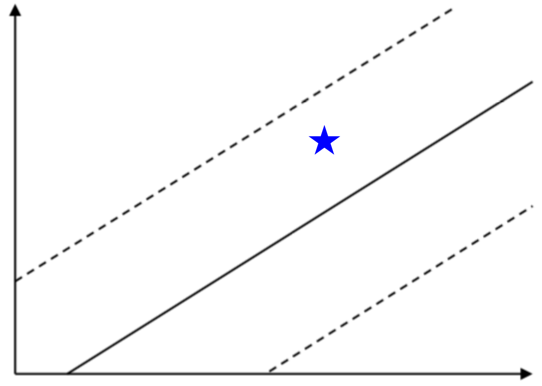
- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- Final thoughts

Midterm 2: in-lab

- Q12: *C* is the penalty on misclassified points



$C = 100$



$C = 1$

Midterm 2: take-home

- Q4(d)
  - Maybe people correctly identified the linear nature of the hidden layers
  - $S = W^{(3)} W^{(2)} W^{(1)} X$
  - Because we have SOFTMAX at the end, the model reduces to **multi-class logistic regression**

- Extra Credit: come talk to me!

Outline for December 12

- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- Final thoughts

Outline for December 12

- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- Final thoughts

Not posted online – let me know if you have questions

Outline for December 12

- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- Final thoughts

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 $\begin{cases} X = \text{protected attribute} \\ Y = \text{other attributes} \end{cases}$ $\Big\}$ $X=0$ minority group, $X=1$ majority group

label $\{ C = $ binary outcome

$C=1$ (hired)

$C=0$ (not)

Disparate Impact (legal definition)

$$\underbrace{P(C=1 \mid X=0)}_{\beta} \leq 0.8 \cdot P(C=1 \mid X=1)$$

example

40% women hired

60% men "

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

$$0.48$$

$\Rightarrow$ there is disparate impact

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

predictor: $\boxed{f : Y \to X}$

$\boxed{\text{Balanced Error Rate}}$ BER

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

want high! $\frac{1}{2}$

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

don't depend on f

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

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

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

β high? ☆
β low

Example of repair

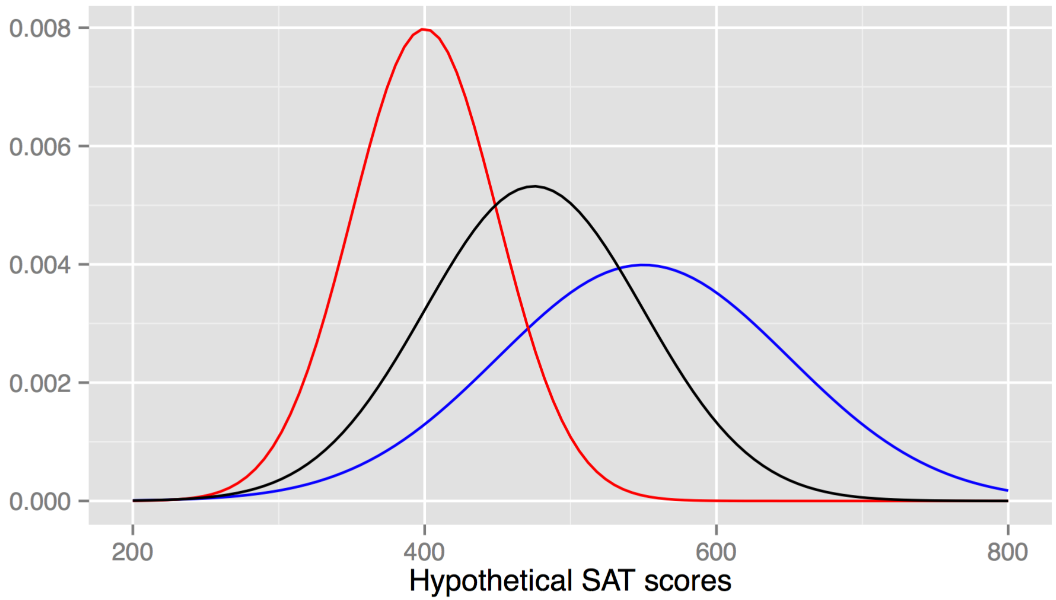


Figure 1: Consider the fake probability density functions shown here where the blue curve shows the distribution of SAT scores (Y) for $X = \texttt{female}$, with $\mu = 550, \sigma = 100$, while the red curve shows the distribution of SAT scores for $X = \texttt{male}$, with $\mu = 400, \sigma = 50$. 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 December 12

- Brief midterm followups
- Project Presentations:
  - Jocelyn & Lamiaa
  - Emile & Gareth
- Brief discussion of CNNs in genetics
- Certifying and removing disparate impact
- 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?