Lecture 25: Fairness in ML
About 1,810,000,000 results (1.10 seconds)

**Shoes at Zappos.com**
- [www.zappos.com/Shoes](http://www.zappos.com/Shoes)
- 4.7 ★★★★★ rating for zappos.com
- Fast, Free Shipping & Free 365 Day Returns on Huge Selection of Shoes!
- Dicenstock · Nike · Converse · New Balance · Born · Frye
- Types: Sneakers, Slippers, Heels, Boots, Flats, Running Shoes

**Womens Shoes at Macy's - Save 40-60% on Cyber Monday - macy's.com**
- [www.macy's.com/Womens_Shoes/Holiday_Deals](http://www.macy's.com/Womens_Shoes/Holiday_Deals)
- 4.2 ★★★★★ rating for macy's.com
- Cyber Monday Ends 11/9/0, Deals Going Fast + Extra 20% Off!
- Styles: Boots, Wedges, Loafers, Flats, Slippers, Sneakers, Pumps, Zipper Pumps, Booties
- Ratings: Returns 9/10 · Product quality 9/10 · Shipping 8.5/10 · Service 8.5/10 · Selection 8.5/10

**Converse® Official Site - Converse.com**
- [www.converse.com/Shoes](http://www.converse.com/Shoes)
- Full Converse Collection, Shop Our New Designs for Men, Women & Kids.
- Free Shipping for Members · Free 60 Day Returns
- Types: Men's Sneakers, Women's Sneakers, Kid's Sneakers, Custom Sneakers

**Brand Name Shoes for Less - Up to 70% Off - zulily.com**
- [www.zulily.com](http://www.zulily.com)
- Exclusive Deals On All Shoes, Socks, & More at Zulily. Shop & Save Today!

**Men and Womens shoes, Shipped Free | Zappos.com**
- [www.zappos.com/shoes](http://www.zappos.com/shoes)
- Boots Sneakers & Athletic Heels Flats Sandals View All... Sneakers & Athletic Boots Oxford's Loafers
- Sandals View All... Sneakers and Athletic Boots Slippers Flats Sandals view all...
- Sandals · Womens Shoes · Sneakers & Athletic Shoes · Popular Men's Shoe Styles
What product would you like...
ML Dream

What news would you prefer to read…
ML Dream

Find the best job for you…

LinkedIn

8,715 Data Scientist jobs in United States

Data Scientist
LeadGenius
San Francisco Bay Area
We are looking for a seasoned Data Scientist/Machine Learning engineer to build the next generation mission critical data platform. Solid engineering and coding skills.

Data Scientist
Feedzai
Atlanta, Georgia
...client’s data feeds. Work with the the client to explore their data and better understand it.

Data Scientist
Jetlore
Sunnyvale, California
We are looking for an exceptional data scientist who is excited to work on challenging problems involving massive amount of data. Ping pong skills is a plus!

Data Scientist
Covestro
Greater Pittsburgh Area
Covestro is in search of a Data Scientist... and data analysts to help influence changes...
ML Dream

- For every user predict: Ads, products, news, ...
- Have tons of data to learn this task well
- Have right models that can learn from all this data
For every user predict: Ads, products, news, …
Have tons of data to learn this task well
Have right models that can learn from all this data

With Big Data comes Bigger Responsibilities …
Women less likely to be shown ads for high-paid jobs on Google, study shows

Automated testing and analysis of company’s advertising system reveals male job seekers are shown far more adverts for high-paying executive jobs.

Female job seekers are much less likely to be shown adverts on Google for highly paid jobs than men, researchers have found.
### Prediction Fails Differently for Black Defendants

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>African American</th>
</tr>
</thead>
<tbody>
<tr>
<td>Labeled Higher Risk, But Didn’t Re-Offend</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Labeled Lower Risk, Yet Did Re-Offend</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Overall, Northpointe’s assessment tool correctly predicts recidivism 61 percent of the time. But blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend. It makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower risk but go on to commit other crimes. *(Source: ProPublica analysis of data from Broward County, Fla.)*
Can we make ML Fair?

- These are machine learning algorithms that learn to predict automatically
- They are not designed to be unfair
- Why is this happening?
- How do we fix them?
Why Now?
Why Now?

Loads of data collected everywhere!
Why Now?

galvanize

Machine Learning

coursera

The Data Incubator

METIS

Startup.ML

datascience@berkeley
Why Now?

Raise in number of Data Scientists!
the algorithms in themselves are neutral. “This program had absolutely nothing to do with race... but multi-variable equations,”
Why is ML Unfair?

- Data collection, labeling etc. can have unintentional biases
  - We learn from past data, historic biases
- Data in itself nor algorithms explicitly know of social inequities
Ignore all protected attributes. 
Eg. Don’t look at race, gender etc.
Ignore all protected attributes.
Eg. Don’t look at race, gender etc.

Problem: You don’t need to look to be able to predict
Ignored all protected attributes.
Eg. Don’t look at race, gender etc.

Problem: You don’t need to look to be able to predict

Eg. User visits “www.artofmanliness.com”
... highly likely to be male
Biases are often not intentional …
Biases are often not intentional …

Most training examples standard white American names: James, John, Robert, Jennifer, Michael, …

Ethnic names often unique, much fewer training examples
Biases are often not intentional …

Most training examples standard white American names: James, John, Robert, Jennifer, Michael, …

Ethnic names often unique, much fewer training examples

Most ML models aim for accuracy for the majority at the expense of mistakes on the smaller protected class
Demographic Parity
Demographic Parity

Population
Demographic Parity

Population $T^c$ $T$
Demographic Parity

Population

$T$: Protected subset

$T^c$: Rest of the population
Demographic Parity

$T : \text{Protected subset}$

$T^c : \text{Rest of the population}$

\[ P(\text{Outcome}|T) \approx P(\text{Outcome}|T^c) \]
Demographic Parity

Population

Eg. Fraction of people shown high paying jobs in $T$ and in $T^c$ is equal
• Is this good enough?
• Is this good enough?

• Say there is this algorithm to select people to invite to apply for this exclusive, credit card with high annual fee
• Is this good enough?

• Say there is this algorithm to select people to invite to apply for this exclusive, credit card with high annual fee

• One way to satisfy the demographic parity:
• Is this good enough?

• Say there is this algorithm to select people to invite to apply for this exclusive, credit card with high annual fee

• One way to satisfy the demographic parity:

• Make offer to higher income people in the unprotected class
• Is this good enough?

• Say there is this algorithm to select people to invite to apply for this exclusive, credit card with high annual fee

• One way to satisfy the demographic parity:
  • Make offer to higher income people in the unprotected class
  • Make offer to lower income people in protected class (in same proportion)
FAIRNESS THROUGH AWARENESS

• Is this good enough?

• Say there is this algorithm to select people to invite to apply for this exclusive, credit card with high annual fee

• One way to satisfy the demographic parity:
  
  • Make offer to higher income people in the unprotected class

  • Make offer to lower income people in protected class (in same proportion)

  NOT REALLY FAIR!
FAIRNESS THROUGH AWARENESS
• Setup:
• Setup:

• Variable T: indicated protected class or not
Setup:

- Variable T: indicated protected class or not
- Output variable O: Indicates our prediction or outcome
Setup:

- Variable T: indicated protected class or not
- Output variable O: Indicates our prediction or outcome
- Variable Y: indicates true/target/desired outcome (eg. Individual capable/qualified, individual can afford etc.)
• Setup:

  • Variable T: indicated protected class or not
  • Output variable O: Indicates our prediction or outcome
  • Variable Y: indicates true/target/desired outcome (eg. Individual capable/qualified, individual can afford etc.)

**Demographic Parity**

\[ P(O=1|T=1) = P(O=1|T=0) \]
• Setup:
  
  • Variable $T$: indicated protected class or not
  
  • Output variable $O$: Indicates our prediction or outcome
  
  • Variable $Y$: indicates true/target.desired outcome (eg. Individual capable/qualified, individual can afford etc.)

**Demographic Parity**

$$P(O=1|T=1) = P(O=1|T=0)$$

Problem: when $T=0$, $O$ can correlate with $Y$ and if $T=1$, $O$ can be random
Equalized Odds
Equalized Odds

For all \( o, y \in \{0,1\} \)

\[
P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)
\]
Equalized Odds

For all $o, y$ in $\{0, 1\}$

$$P(O=o|Y=y, T=1) = P(O=o|Y=y, T=0)$$

- $O$ is independent of $T$ given $Y$
Equalized Odds

For all $o$, $y$ in $\{0,1\}$

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- $O$ is independent of $T$ given $Y$

Equalized Odds

For all $o, y$ in $\{0,1\}$

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- $O$ is independent of $T$ given $Y$


- Incentive to reduce error uniformly in all groups
Equalized Odds

For all $o, y$ in $\{0, 1\}$

$$P(O=o|Y=y,T=1) = P(O=o|Y=y,T=0)$$

- $O$ is independent of $T$ given $Y$


- Incentive to reduce error uniformly in all groups

Problem: Say in $T$, $2/100$ people qualify and outside $50/100$ qualify.
Company can make 26 offers: 25 to qualifying people in $T'$ and 1 in $T$
FaiRNESS THROuGH AWARENESS

Sufficiency or Predictive Rate Parity
Sufficiency or Predictive Rate Parity

For all o, y in {0,1}

\[ P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0) \]
Sufficiency or Predictive Rate Parity

For all $o, y$ in $\{0,1\}$

$$P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)$$

- $Y$ is independent of $T$ given $O$
Sufficiency or Predictive Rate Parity

For all \( o, y \) in \( \{0,1\} \)

\[
P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)
\]

- \( Y \) is independent of \( T \) given \( O \)
- Equal chance of success (\( Y=1 \)) given acceptance
Sufficiency or Predictive Rate Parity

For all $o$, $y$ in $\{0,1\}$

$$P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)$$

- $Y$ is independent of $T$ given $O$
- Equal chance of success ($Y=1$) given acceptance
Sufficiency or Predictive Rate Parity

For all \( o, y \) in \( \{0,1\} \)

\[
P(Y=y|O=o,T=1) = P(Y=y|O=o,T=0)
\]

- \( Y \) is independent of \( T \) given \( O \)
- Equal chance of success \( (Y=1) \) given acceptance

Problem: Same as equal odds
IMPOSSIBILITY RESULT
• Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive
• Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive

• Chouldechova, 2016 and Kleinberg et al. 2016
**Impossibility Result**

- Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive.
- Demographic parity Vs Sufficiency: if $T$ is dependent of $Y$. 
• Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive.


• Demographic parity Vs Sufficiency: if T is dependent of Y.

• Demographic parity Vs Equal Odds: T is dependent of Y and O is dependent of Y.
• Turns out that other than degenerate cases, any two of the three criterion are mutually exclusive.


• Demographic parity Vs Sufficiency: if $T$ is dependent of $Y$.

• Demographic parity Vs Equal Odds: $T$ is dependent of $Y$ and $O$ is dependent of $Y$.

• Equal Odds Vs Sufficiency: If $T$ is dependent of $Y$.
Achieving Fairness

• Preprocessing: While doing feature extraction, extract features that ensure independence of feature to T (Eg. Equal odds)

• While training: Find model that minimizes training error subject to fairness constraints

• Post-processing: Learn model as before on training data, as post processing use extra training data to learn a bias parameter to correct for fairness
On to the next social issue...
Another issue: ML methods are complex and we don’t understand semantic meaning
Another issue: ML methods are complex and we don’t understand semantic meaning

We need transparency of method for accountability
Another issue: ML methods are complex and we don’t understand semantic meaning

We need transparency of method for accountability

Transparency via interpretability.
- Provide explanation for each decision

What makes an instance a negative instance according to the algorithm
Data Transparency Lab

DTL Conferences

DTL 2017

Barcelona December 11-13th