Fairness, Transparency and other Moral Issues in Machine Learning

Course Webpage:
http://www.cs.cornell.edu/Courses/cs4786/2016fa/
Announcements

• Survey 2, just over 80%

• Make sure you fill out the course eval

  • (If the participation on this is above 90% I will still drop worst assignment)
What product you would like…
What news would you prefer to read…
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.)*
Why Now?
Why Now?

Loads of data collected everywhere!
Why Now?

galvanize

Machine Learning

The Data Incubator

METIS

Startup.ML

datascience@berkeley

Coursera
Why Now?

galvanize

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
Ignore 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
FAIRNESS THROUGH AWARENESS
FAIRNESS THROUGH AWARENESS

Population

$T^C$
FAIRNESS THROUGH AWARENESS

Population

$T$ : Protected subset

$T^c$ : Rest of the population
$T$: Protected subset

$T^c$: Rest of the population

$$P(\text{Outcome}|T) \approx P(\text{Outcome}|T^c)$$
Eg. Fraction of people shown high paying jobs in $T$ and in $T^c$ is equal
Eg. Fraction of people shown high paying jobs in $T$ and in $T^c$ is equal
Eg. Fraction of people shown high paying jobs in T and in $T^c$ is equal
Eg. Fair K-means Clustering (very naive)
Eg. Fair K-means Clustering (very naive)

Objective = \[ \sum_{j=1}^{K} \sum_{t \in C_j} \left\| x_t - r_j \right\|^2 \]

where \( r_j = \frac{1}{|C_j|} \sum_{x_t \in C_j} x_t \)
**Eg. Fair K-means Clustering (very naive)**

\[
\text{Objective} = \sum_{j=1}^{K} \sum_{t \in C_j} \|x_t - r_j\|^2
\]

where \( r_j = \frac{1}{|C_j|} \sum_{x_t \in C_j} x_t \)

Faithfulness constraints: \( \forall j \in [K], \sum_{t : c_t = j} 1_{x_t \in T} = \sum_{t : c_t = j} 1_{x_t \notin T} \)
Eg. Fair $K$-means Clustering (very naive)

Objective \[= \sum_{j=1}^{K} \sum_{t \in C_j} \|x_t - r_j\|^2 \]

where \[r_j = \frac{1}{|C_j|} \sum_{x_t \in C_j} x_t \]

Fairness constraints: \[\forall j \in [K], \sum_{t: c_t = j} 1_{x_t \in T} = \sum_{t: c_t = j} 1_{x_t \notin T} \]

Number of protected in cluster $j$ = Number of unprotected in cluster $j$
A view from a mile above:
A view from a mile above:

Minimize Classification objective
(or whatever other surrogate loss you use usually)
A view from a mile above:

Minimize Classification objective
(or whatever other surrogate loss you use usually)

Added Constraint: subject to proportion of labels in each class being same for protected and unprotected population
• Is this good enough?
Fair Classification

• 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 simple way to satisfy the fairness constraint:
• 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 simple way to satisfy the fairness constraint:
  • 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 simple way to satisfy the fairness constraint:
  
  • Make offer to higher income people in the unprotected class
  
  • Make offer to lower income people in protected class (in same proportion)
• 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 simple way to satisfy the fairness constraint:
  • 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!
A view from a mile above:

Minimize Classification objective
(or whatever other surrogate loss you use usually)
A view from a mile above:

Minimize Classification objective
(or whatever other surrogate loss you use usually)

Added Constraint: subject to proportion of labels for
similar instances
in each class being same for protected
and unprotected population
Its not always about fairness
There are other issues too!
ML Dream

What news would user prefer to read…
Predict for every user what they would like
Show Ads, products, news, ...
• Just because we can predict, should we?
• Just because we can predict, should we?

• Say we have a fair, unbiased algorithm for prediction
• Just because we can predict, should we?
• Say we have a fair, unbiased algorithm for prediction
• Can there be other issues?
Extremizing Effect of ML
Extremizing Effect of ML

User 1

User 2
**Extremizing Effect of ML**

**User 1**

*Apples* are extremely rich in important antioxidants, flavonoids, and dietary fiber. The phytonutrients and antioxidants in *apples* may help reduce the risk of developing cancer, hypertension, diabetes, and heart disease.

**User 2**

For fewer calories per fruit, *oranges* have higher levels of Vitamin C, folate, potassium, and protein.
EXTREMIZING EFFECT OF ML

User 1

Apples are extremely rich in important antioxidants, flavanoids, and dietary fiber. The phytonutrients and antioxidants in apples may help reduce the risk of developing cancer, hypertension, diabetes, and heart disease.

User 2

For fewer calories per fruit, oranges have higher levels of Vitamin C, folate, potassium, and protein.

The Health Benefits of Oranges
- Packed with fiber to promote healthy digestion
- Full of folate to help the body form red blood cells
- A good source of immune-boosting vitamin C
- Can help maintain a healthy heart
- Keeps blood and eyes healthy with its content of vitamin A
- A great source of calcium for healthy and strong bones
EXTREMIZING EFFECT OF ML

User 1

Apples are extremely rich in important antioxidants, flavonoids, and dietary fiber. The phytonutrients and antioxidants in apples may help reduce the risk of developing cancer, hypertension, diabetes, and heart disease.

User 2

For fewer calories per fruit, oranges have higher levels of Vitamin C, folate, potassium, and protein.

The Health Benefits of Oranges
- Packed with fiber to promote healthy digestion
- Full of folate to help the body form red blood cells
- A good source of immune-boosting vitamin C
- Contains potassium to ensure a healthy heart
- Keeps vision clear and eye healthy with its content of vitamin A
- A great source of calcium for healthy and strong bones
**Extremizing Effect of ML**

**User 1**

*Apples* are extremely rich in important antioxidants, flavonoids, and dietary fiber. The phytonutrients and antioxidants in *apples* may help reduce the risk of developing cancer, hypertension, diabetes, and heart disease.

**User 2**

For fewer calories per fruit, *oranges* have higher levels of Vitamin C, folate, potassium, and protein.

**The Health Benefits of Oranges**

- Packed with fiber to promote healthy digestion
- Full of folate to help the body form red blood cells
- A good source of immune-boosting vitamin C
- Contains potassium to ensure a healthy heart
- Keeps vision clear and eyes healthy with its content of vitamin A
- A great source of calcium for healthy and strong bones
EXTREMIZING EFFECT OF ML
Possible fixes:
• Possible fixes:
  • Good contents is the best fix
Possible fixes:

• Good contents is the best fix

• Mix up articles: provide both sides
Possible fixes:

- Good contents is the best fix
- Mix up articles: provide both sides
  - what if there are multiple sides and worse yet you can’t identify the sides
Possible fixes:

- Good contents is the best fix

- Mix up articles: provide both sides
  - what if there are multiple sides and worse yet you can’t identify the sides

- Mix up user profiles from time to time (or have a canonical user and every user is mix of individual and the canonical user)
Possible fixes:

- Good contents is the best fix

- Mix up articles: provide both sides
  - what if there are multiple sides and worse yet you can’t identify the sides

- Mix up user profiles from time to time (or have a canonical user and every user is mix of individual and the canonical user)

- This is a completely open topic …
Extremizing Effect of ML

- Possible fixes:
  - Good contents is the best fix
  - Mix up articles: provide both sides
    - what if there are multiple sides and worse yet you can’t identify the sides
  - Mix up user profiles from time to time (or have a canonical user and every user is mix of individual and the canonical user)
  - This is a completely open topic …
    But a very important one …
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
November 18th 2016 / New York University, NYC

Fairness, Accountability, and Transparency in Machine Learning

Co-located with the Data Transparency Lab Conference and the Workshop on Data and Algorithmic Transparency

The workshop is now over but a recording of the event will be available later.

Sponsors

Google
MIT Connection Science
mozilla

Microsoft
DATA
TRANSPARENCY

NYU Technology Law and Policy Clinic

Telefónica
Time for General Questions