Graduate AI

Lecture 27: Ethics and AI II

Teachers:

Zico Kolter
Ariel Procaccia (this time)
FAIRNESS IN ML

• AI algorithms are supposedly unbiased
• But they may make use of features that interact with protected attributes
• For example, zip code is sometimes correlated with race
• There is a fast-growing body of evidence for discrimination by AI algorithms
EXAMPLE: AD DELIVERY

Racism is Poisoning Online Ad Delivery, Says Harvard Professor

Google searches involving black-sounding names are more likely to serve up ads suggestive of a criminal record than white-sounding names, says computer scientist

February 4, 2013

Have you ever been arrested? Imagine the question not appearing in the solitude of your thoughts as you read this paper, but appearing explicitly whenever someone queries your name in a search engine.”

So begins Latanya Sweeney at Harvard University in a compelling paper arguing that racial discrimination plagues online ad delivery.

Many people will have experience Googling friends, colleagues and relatives to find out about their online presence—the websites on which they appear, their pictures, hobbies and so on.
# Example: Ad Delivery

<table>
<thead>
<tr>
<th>Title</th>
<th>URL</th>
<th>Coefficient</th>
<th>Appears in Agents</th>
<th>Total Appearances</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Top ads for identifying the simulated female group</strong></td>
<td></td>
<td></td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>Jobs (Hiring Now)</td>
<td><a href="http://www.jobsinyourarea.co">www.jobsinyourarea.co</a></td>
<td>0.34</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>4Runner Parts Service</td>
<td><a href="http://www.westernpatoytosaservice.com">www.westernpatoytosaservice.com</a></td>
<td>0.281</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>Criminal Justice Program</td>
<td>www3.mc3.edu/Criminal+Justice</td>
<td>0.247</td>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>Goodwill - Hiring</td>
<td>goodwill.careerboutique.com</td>
<td>0.22</td>
<td>45</td>
<td>15</td>
</tr>
<tr>
<td>UMUC Cyber Training</td>
<td><a href="http://www.umuc.edu/cybersecuritytraining">www.umuc.edu/cybersecuritytraining</a></td>
<td>0.199</td>
<td>19</td>
<td>17</td>
</tr>
<tr>
<td><strong>Top ads for identifying agents in the simulated male group</strong></td>
<td></td>
<td></td>
<td>female</td>
<td>male</td>
</tr>
<tr>
<td>$200k+ Jobs - Execs Only</td>
<td>careerchange.com</td>
<td>-0.704</td>
<td>60</td>
<td>402</td>
</tr>
<tr>
<td>Find Next $200k+ Job</td>
<td>careerchange.com</td>
<td>-0.262</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Become a Youth Counselor</td>
<td><a href="http://www.youthcounseling.degreeleap.com">www.youthcounseling.degreeleap.com</a></td>
<td>-0.253</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>CDL-A OTR Trucking Jobs</td>
<td><a href="http://www.tadriver.com/OTRJobs">www.tadriver.com/OTRJobs</a></td>
<td>-0.149</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Free Resume Templates</td>
<td>resume-templates.resume-now.com</td>
<td>-0.149</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

[Datta et al. 2016]
INDIVIDUAL FAIRNESS

• Model introduced by Dwork et al. (2012)
• Set of individuals $V$ and outcomes $A$
• Metric on individuals $d: V \times V \to \mathbb{R}^+$
• Metric $D$ on distributions over outcomes
• Randomized classifier $M: V \to \Delta(A)$
• $M$ satisfies the **Lipschitz property** if for all $x, y \in V$,
  \[ D(M(x), M(y)) \leq d(x, y) \]
INDIVIDUAL FAIRNESS
**Individual Fairness**

- We can get a Lipschitz classifier by setting $M(x) = M(y)$ for all $x, y \in V$
- But we want to minimize a loss function $L: V \times A \to \mathbb{R}^+$
- This leads to the optimization problem

\[
\min \sum_{x \in V} \sum_{a \in A} \mu_x(a) \cdot L(x, a)
\]

s.t. $\forall x, y \in V, D(\mu_x, \mu_y) \leq d(x, y)$

$\forall x \in V, \mu_x \in \Delta(A)$
INDIVIDUAL FAIRNESS

• Various options for the metric $D$

• Example: \textbf{total variation norm}, defined for distributions $P$ and $Q$ as
  \[ D_{tv}(P, Q) = \max_{E \subseteq A} |P(E) - Q(E)| \]

• \textbf{Lemma}: (we skip the simple proof)
  \[ D_{tv}(P, Q) = \frac{1}{2} \sum_{a \in A} |P(a) - Q(a)| \]

• When $D = D_{tv}$, the optimization problem is a linear program!
Where would the similarity metric come from?
GROUP FAIRNESS

• Assume we are making a binary decision $\hat{Y} \in \{0,1\}$, and there is a protected attribute $A \in \{0,1\}$

• Demographic parity:

$$\Pr[\hat{Y} = 1 \mid A = 0] = \Pr[\hat{Y} = 1 \mid A = 1]$$

• May accept unqualified individuals when $A = 0$, and qualified individuals when $A = 1$!
GROUP FAIRNESS

\[ Y = 0 \]
\[ \hat{Y} = 0 \]
\[ Y = 1 \]
\[ \hat{Y} = 1 \]

\[ A = 0 \]

\[ Y = 0 \]
\[ \hat{Y} = 1 \]

\[ Y = 1 \]
\[ \hat{Y} = 0 \]

\[ A = 1 \]
GROUP FAIRNESS

• We will follow the exposition of Hardt et al. [2016]

• \( \hat{Y} \) satisfies equalized odds with respect to protected attribute \( A \) and outcome \( Y \) if \( \hat{Y} \) and \( A \) are independent conditional on \( Y \)

• That is, for all \( y \in \{0,1\} \),

\[
\Pr[\hat{Y} = 1 \mid A = 0, Y = y] = \Pr[\hat{Y} = 1 \mid A = 1, Y = y]
\]
Relations between Properties

• Demographic parity:
  \[ \Pr[\hat{Y} = 1 \mid A = 0] = \Pr[\hat{Y} = 1 \mid A = 1] \]

• Equalized odds: For all \( y \in \{0,1\}, \)
  \[ \Pr[\hat{Y} = 1 \mid A = 0, Y = y] = \Pr[\hat{Y} = 1 \mid A = 1, Y = y] \]

• Poll 1: Relation between demographic parity and equalized odds?
  1. Demographic parity \( \Rightarrow \) equalized odds
  2. Equalized odds \( \Rightarrow \) demographic parity
  3. Incomparable
GROUP FAIRNESS

\[ Y = 0 \]
\[ \hat{Y} = 0 \]
\[ Y = 1 \]
\[ \hat{Y} = 1 \]

\[ A = 0 \]
\[ Y = 0 \]
\[ \hat{Y} = 0 \]

\[ A = 1 \]
\[ Y = 1 \]
\[ \hat{Y} = 1 \]
\( \hat{Y} = Y \) may not satisfy demographic parity!
**Example: FICO scores**

- FICO scores are a proprietary classifier widely used in the United States to predict credit worthiness.
- Range from 300 to 850, where cutoff of 620 is commonly used for prime-rate loans.
- Based on features, such as number of bank accounts used, that may interact with race in unfair ways.
**Example: FICO Scores**

Non-default rate by FICO score

- **Asian**
- **White**
- **Hispanic**
- **Black**

CDF of FICO score by group

- **Asian**
- **White**
- **Hispanic**
- **Black**

[Hardt et al. 2016]

15780 Spring 2018: Lecture 27
EXAMPLE: FICO SCORES

[Hardt et al. 2016]

15780 Spring 2018: Lecture 27
ACHIEVING EQUALIZED ODDS

• We wish to derive a classifier \( \tilde{Y} \) from a possibly discriminatory classifier \( \hat{Y} \)

• \( \tilde{Y} \) is derived from \( \hat{Y} \) and \( A \) if it is a possibly randomized function of \( (\hat{Y}, A) \) alone

• \( \tilde{Y} \) is completely described by four parameters in \([0,1]\) corresponding to \( \Pr[\tilde{Y} = 1 \mid \hat{Y} = \hat{y}, A = a] \) for \( \hat{y}, a \in \{0,1\} \)
Achiving equalized odds

• Define $\gamma_a(\hat{Y})$ as 
  $\Pr[\hat{Y} = 1|A = a, Y = 0], \Pr[\hat{Y} = 1|A = a, Y = 1]$ 

• Poll 2: $\hat{Y}$ satisfies equalized odds iff 
  1. $\gamma_0(\hat{Y}) = \gamma_1(\hat{Y})$
  2. $\gamma_0(\hat{Y}) \leq \gamma_1(\hat{Y})$
  3. $\gamma_0(\hat{Y}) = 1 - \gamma_1(\hat{Y})$
  4. $\gamma_0(\hat{Y}) = 0$ and $\gamma_1(\hat{Y}) = 1$
Achiving Equalized Odds*

- **Lemma:** \( \tilde{Y} \) is derived iff for all \( a \in \{0,1\} \), 
  \( \gamma_a(\tilde{Y}) \in P_a(\hat{Y}) \), where \( P_a(\hat{Y}) \) is
  \( \text{conv}\{(0,0), \gamma_a(\hat{Y}), \gamma_a(1 - \hat{Y}), (1,1)\} \)

---

*Just for fun*
Achieving Equalized Odds*

- Loss function $\ell: \{0,1\}^2 \rightarrow \mathbb{R}^+$ gives the loss $\ell(\hat{y}, y)$ of predicting $\hat{y}$ when the label is $y$.
- The optimization problem is:

$$\min \mathbb{E}[\ell(\tilde{Y}, Y)]$$
$$\text{s.t. } \forall a \in \{0,1\}, \gamma_a(\tilde{Y}) \in P_a(\tilde{Y})$$
$$\gamma_0(\tilde{Y}) = \gamma_1(\tilde{Y})$$

- Theorem: This is a linear program

*Just for fun
The construction of $\tilde{Y}$ depends on the joint distribution of $(\hat{Y}, A, Y)$ at training time, but at prediction time we only have access to $(\hat{Y}, A)$.
SUMMARY

• Definitions
  o Lipschitz property
  o Statistical parity
  o Equalized odds

• Algorithms:
  o LP for Lipschitz classifiers
  o LP for deriving a classifier satisfying equalized odds