Fairness in ML

Spring 2020
Today

• Fairness Overview
• Association and Fairness/Privacy
• Adversarial training for fair models
• Other approaches
Fairness in ML
Fairness in the real world (USA)

- Disparate Treatment
  - Given a job applicant pool, company uses a test only for black applicants, resulting in hiring more white applicants than black.

- Disparate Impact
  - Company uses assessment test on all applicants resulting in hiring more white applicants than black.

- Legal protections for certain classes in certain contexts.
  - Race, age, gender, ...
  - Hiring, lending, housing, ...

- See “Big Data’s Disparate Impact” by Barocas, Selbst.
Fairness in the real world (USA)

- Protected class may or may not be even present.
  - “redlining”
Examples in ML

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

• Performance of vision systems:
  • http://content.time.com/time/business/article/0,8599,1954643,00.html

• Credit worthiness:
  • https://www.wired.com/story/the-apple-card-didnt-see-genderand-thats-the-problem/

• Hiring worthiness:
  • https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
Examples in ML

• Vision
  • Trent’s datasets

• Crime predictions
  • Crime prediction based on vision systems

• Alexa
  • Different voice recognition performance for genders, accents, etc.

• Datasets
  • https://www.artsy.net/news/artsy-editorial-online-image-database-will-remove-600-000-pictures-art-project-revealed-systems-racist-bias
Why is ML biased?
Why is data biased?

• Historical injustices.
• Current injustices.
• Unbalanced data.
• Confidence imbalance.
• Feedback loops.
• Higher-level issues.
  • Costs of balancing data.
Association in Fairness and Privacy
Fairness and Privacy in ML (simplified)

• Fairness: Do not “use” protected class in some contexts.
• Privacy: Do not “use” sensitive/private attribute in some (other) contexts.

• “Use” ~ Association
  • Caveats
    • “association is not causation”
    • Causation is not association

• * More to fairness/privacy than this.
ML (simplified)

<table>
<thead>
<tr>
<th>X</th>
<th>Y</th>
<th>Z</th>
<th>...</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Machine Learning

Search terms,...

Clicked on Ad

Google broke Canada’s privacy laws with targeted health ads, watchdog says

Google broke Canada’s privacy laws with targeted health ads, watchdog says
Online Ads for High-Paying Jobs Are Targeting Men More Than Women

New study uncovers gender bias
Machine Learning

Past purchases,...

Purchase

Forbes / Tech

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
Big Data
Big Data
Big Data
Big Data
Big Data -- Non-adversarial
Big Data

Microsoft Finds Cancer Clues in Search Queries

By JOHN MARKOFF JUNE 7, 2016

Facebook, Google refer suicidal people to help lines

By Mark Milian, CNN
① Updated 5:46 PM ET, Tue December 13, 2011
Fairness and Privacy in ML (simplified)

• Fairness: Do not “use” protected class.
• Privacy: Do not “use” sensitive/private attribute.

• Attempt #1
  • Remove protected class / sensitive attributes from training data.
The Challenge: Proxies

Forbes / Tech
FEB 16, 2012 @ 11:02 AM 3,269,456

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did
Fairness and Privacy in ML (simplified)

• Fairness: Do not “use” protected class.
• Privacy: Do not “use” sensitive/private attribute.

• Attempt #1
  • Remove protected class / sensitive attributes from training data.
  • Associations persist.
  • Can cause more unfairness.
Adversarial Training for Fair Models
Adversarial training for fair models

- Definitions of fairness
- Fairness in model loss
- Experiments
Fairness Criteria

Definition 1. Demographic Parity. A predictor $\hat{Y}$ satisfies demographic parity if $\hat{Y}$ and $Z$ are independent.

This means that $P(\hat{Y} = \hat{y})$ is equal for all values of the protected variable $Z$: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$.

Definition 2. Equality of Odds. A predictor $\hat{Y}$ satisfies equality of odds if $\hat{Y}$ and $Z$ are conditionally independent given $Y$.

This means that, for all possible values of the true label $Y$, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

Definition 3. Equality of Opportunity. If the output variable $Y$ is discrete, a predictor $\hat{Y}$ satisfies equality of opportunity with respect to a class $y$ if $\hat{Y}$ and $Z$ are independent conditioned on $Y = y$.

This means that, for a particular value of the true label $Y$, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

Example: $Y$ is high skill, $\hat{Y}$ is predicted high skill (therefore job offer); $Z$ is gender.

The ratio of people who get the job is the same as the ratio of men who get the job, and the ratio of women who get the job.

The ratio of people with high skill who get the job is the same as the ratio of high-skilled men who get the job and the ratio of high-skilled women who get the job. Same for low-skilled.

As above, but only for high-skill.
Fairness Criteria

**Definition 1.** Demographic Parity. A predictor $\hat{Y}$ satisfies demographic parity if $\hat{Y}$ and $Z$ are independent.

This means that $P(\hat{Y} = \hat{y})$ is equal for all values of the protected variable $Z$: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$.

**Definition 2.** Equality of Odds. A predictor $\hat{Y}$ satisfies equality of odds if $\hat{Y}$ and $Z$ are conditionally independent given $Y$.

This means that, for all possible values of the true label $Y$, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

**Definition 3.** Equality of Opportunity. If the output variable $Y$ is discrete, a predictor $\hat{Y}$ satisfies equality of opportunity with respect to a class $y$ if $\hat{Y}$ and $Z$ are independent conditioned on $Y = y$.

This means that, for a particular value of the true label $Y$, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

<table>
<thead>
<tr>
<th>Without Debiassing</th>
<th>With Debiassing</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Female</strong></td>
<td><strong>Pred 0</strong></td>
</tr>
<tr>
<td>True 0</td>
<td>4711</td>
</tr>
<tr>
<td>True 1</td>
<td>265</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td><strong>Pred 0</strong></td>
</tr>
<tr>
<td>True 0</td>
<td>6907</td>
</tr>
<tr>
<td>True 1</td>
<td>1194</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrices on the UCI Adult dataset, with and without equality of odds enforcement.

In-class calculations:

- Def 1. $\Pr[\text{Pred0}] = (4711 + 265 + 6907 + 1194) / (\text{Same} + \text{other col}) = 0.803$
- $\Pr[\text{Pred0} | \text{Male}] = (6907 + 1194) / (\text{Same} + \text{other col}) = 0.74$
- $\Pr[\text{Pred0} | \text{Female}] = 0.91$
- Def 2.
  - $\Pr[\text{Pred0} | \text{True0}] =$
  - $\Pr[\text{Pred0} | \text{True0 Male}] =$
  - $\Pr[\text{Pred0} | \text{True0 Female}] =$
  - Def 3.
Adversaries

Figure 1: The architecture of the adversarial network.
Loss

• Adversary maximizes prediction of $z$. Follows gradient of standard prediction loss.

• Predictor optimizes for:

\[ \nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A \quad (1) \]
Experiments

- Note: we will cover word embeddings in more detail later.
- Toy scenario
- Adult dataset income prediction

<table>
<thead>
<tr>
<th></th>
<th>Without Debiasing</th>
<th>With Debiasing</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Pred 0</td>
<td>Pred 1</td>
</tr>
<tr>
<td><strong>Female</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True 0</td>
<td>4711</td>
<td>120</td>
</tr>
<tr>
<td>True 1</td>
<td>265</td>
<td>325</td>
</tr>
<tr>
<td><strong>Male</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>True 0</td>
<td>6907</td>
<td>697</td>
</tr>
<tr>
<td>True 1</td>
<td>1194</td>
<td>2062</td>
</tr>
</tbody>
</table>

Table 3: Confusion matrices on the UCI Adult dataset, with and without equality of odds enforcement.
Connection to Privacy

• Prevent $z$ from being predictable from model output.
• Prevent $z$ from being “used”.
• Because $z$ is race/gender/etc.

![Diagram of adversarial network](image)

Figure 1: The architecture of the adversarial network.

• Privacy: because $z$ is private.
Other approaches
Other approaches for association

- Debias the data (more common for privacy).
- Post-process model outputs.
- ”Repair” model after training.