Fairness in ML 2: Equal opportunity and odds

Privacy & Fairness in Data Science
CompSci 590.01 Fall 2018

Slides adapted from https://fairmlclass.github.io/4.html
Outline

• Observational measure of fairness
  – Issues with Disparate Impact
  – Equal opportunity and Equalized odds
  – Positive Rate Parity
  – Tradeoff

• Achieving Equalized Odds
  – Binary Classifier
Supervised Learning

<table>
<thead>
<tr>
<th>X1</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>...</th>
<th>Race</th>
<th>Bail</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>...</td>
<td>0</td>
<td>1</td>
<td>...</td>
<td>1</td>
<td>Y</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>1</td>
<td>N</td>
</tr>
<tr>
<td>1</td>
<td>...</td>
<td>1</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>N</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

\[ P_a\{E\} = P\{E \mid A = a\}. \]
Demographic parity
(or the reverse of disparate impact)

**Definition.** Classifier $C$ satisfies *demographic parity* if $C$ is independent of $A$.

When $C$ is binary 0/1-variables, this means

$$\mathbb{P}_a\{C = 1\} = \mathbb{P}_b\{C = 1\}$$

for all groups $a, b$.

Approximate versions:

$$\frac{\mathbb{P}_a\{C = 1\}}{\mathbb{P}_b\{C = 1\}} \geq 1 - \epsilon$$

$$|\mathbb{P}_a\{C = 1\} - \mathbb{P}_b\{C = 1\}| \leq \epsilon$$
Demographic parity Issues

\[ Y = 1 \]

[Diagram showing parity with \( A = 1 \) and \( A = 0 \)]
Demographic parity Issues

• Does not seem “fair” to allow random performance on \( A = 0 \)
• Perfect classification is impossible
Perfect Classifier and Fairness

• The perfect classifier may not ensure demographic parity
  – $Y$ is correlated with $A$

• What if we did not know how the classifier $C$ was created?
  – No access to the classifier (to retrain)
  – No access to the training data (human created classifier)
True Positive Parity (TPP)  
(or equal opportunity)

Assume $C$ and $Y$ are binary 0/1-variables.

**Definition.** Classifier $C$ satisfies *true positive parity* if 

$$P_a\{C = 1 \mid Y = 1\} = P_b\{C = 1 \mid Y = 1\} \text{ for all groups } a, b.$$

- When positive outcome (1) is desirable
- Equivalently, primary harm is due to false negatives
  - Deny bail when person will not recidivate
TPP

• Forces similar performance on $Y = 1$
False Positive Parity (FPP)

Assume $C$ and $Y$ are binary 0/1-variables.

**Definition.** Classifier $C$ satisfies *false positive parity* if

\[ P_a\{C = 1 \mid Y = 0\} = P_b\{C = 1 \mid Y = 0\} \] for all groups $a, b$.

• TPP + FPP: Equalized Odds, or Positive Rate Parity

\[ R \text{ satisfies equalized odds if} \]

\[ R \text{ is conditionally independent of } A \text{ given } Y. \]
Positive Rate Parity

\[ A = 1 \]

\[ A = 0 \]
Predictive Value Parity

Assume \( C \) and \( Y \) are binary 0/1-variables.

**Definition.** Classifier \( C \) satisfies
- **positive predictive value parity** if for all groups \( a, b \):
  \[
P_a\{Y = 1 \mid C = 1\} = P_b\{Y = 1 \mid C = 1\}
\]
- **negative predictive value parity** if for all groups \( a, b \):
  \[
P_a\{Y = 1 \mid C = 0\} = P_b\{Y = 1 \mid C = 0\}
\]
- **predictive value parity** if it satisfies both of the above.

Equalized chance of success given acceptance
Predictive Value Parity

\[
P_1[Y = 1 \mid C = 1] = \frac{7}{10}
\]

\[
P_0[Y = 1 \mid C = 1] = \frac{3}{5}
\]

\[
P_1[Y = 1 \mid C = 0] = \frac{4}{10}
\]

\[
P_0[Y = 1 \mid C = 0] = \frac{1}{5}
\]
Predictive Value Parity

For $A = 1$:

- $P_1[Y = 1 | C = 1] = \frac{8}{9}$
- $P_1[Y = 1 | C = 0] = 0$

For $A = 0$:

- $P_0[Y = 1 | C = 1] = \frac{1}{3}$
- $P_0[Y = 1 | C = 0] = 0$
Trade-off

Proposition. Assume differing base rates and an imperfect classifier $C \neq Y$. Then, either

- positive rate parity fails, or
- predictive value parity fails.

• We will look at a similar result later in the course due to Kleinberg, Mullainathan and Raghavan (2016)
Outline

• Observational measure of fairness
  – Issues with Disparate Impact
  – Equal opportunity and Equalized odds
  – Positive Rate Parity
  – Tradeoff

• Achieving Equalized Odds
  – Binary Classifier
Equalized Odds

\( R \) satisfies equalized odds if \( R \) is conditionally independent of \( A \) given \( Y \).

• Derived Classifier: A new classifier \( \tilde{C} \) that only depends on \( C, A \) (and \( Y \))
Derived Classifier

$P_1[C = 1 \mid Y = 0] \neq P_0[C = 1 \mid Y = 0]$
Derived Classifier

• Options for $\tilde{C}$:
  
  – $\tilde{C} = C$
  
  – $\tilde{C} = 1 - C$
  
  – $\tilde{C} = 1$
  
  – $\tilde{C} = 0$
  
  – Or some randomized combination of these

$\tilde{C}$ is in the enclosed region
Derived Classifier

For equal odds, result lies below all ROC curves.

$P_A[C = 1 \mid Y = 1]$  

$P_A[C = 1 \mid Y = 0]$  

$\tilde{C}$ is in this region for $A = 0$  

$\tilde{C}$ is in this region for $A = 1$
Summary: Multiple fairness measures

- Demographic parity or disparate impact
  - Pro: Used in the law
  - Con: Perfect classification is impossible
  - Achieved by modifying training data

- Equal Odds/ Opportunity
  - Pro: Perfect classification is possible
  - Con: Different groups can get rates of positive prediction
  - Achieved by post processing the classifier
Summary: Multiple fairness measures

• Equal odds/opportunity
  – Different groups may be treated unequally
  – Maybe due to the problem
  – Maybe due to bias in the dataset

• While demographic parity seems like a good fairness goal for the society, …
  Equal odds/opportunity seems to be measuring whether an algorithm is fair (independent of other factors like input data).
Summary: Multiple fairness measures

• Fairness through Awareness:
  – Need to define a distance function $d(x,x')$
  – A guarantee at the individual level (rather than on groups)
  – How does this connect to other notions of fairness?