Fairness in ML

These slides were assembled by Eric Eaton, with grateful acknowledgement of the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution. Please send comments and corrections to Eric.

Robot Image Credit: Viktoria Khmelnitskaya © 123RF.com
Fairness

• Widespread algorithms with many small interactions
  • e.g. search, recommendations, social media
• Specialized algorithms with fewer but higher-stakes interactions
  • e.g. medicine, criminal justice, finance
• At this level of impact, algorithms can have unintended consequences
• Low classification error is not enough, need fairness

Based on materials by David Madras
Regulated domains

- **Credit** (Equal Credit Opportunity Act)
- **Education** (Civil Rights Act of 1964; Education Amendments of 1972)
- **Employment** (Civil Rights Act of 1964)
- **Housing** (Fair Housing Act)
- **Public Accommodation** (Civil Rights Act of 1964)

Extends to marketing and advertising; not limited to final decision
This list sets aside complex web of laws that regulates the government
Background on US Prison Population

Incarceration Rates per 100,000

Sentenced Federal and State Prisoners by Race and Hispanic Origin, 2009-2016

Data from 2014

Note: Whites and blacks include only those who are single-race, not Hispanic. Hispanics are of any race. Prison population is defined as inmates sentenced to more than a year in federal or state prison.

Source: Bureau of Justice Statistics.

PEW RESEARCH CENTER
Case Study: COMPAS

• Software by Northpointe that predicts recidivism
• Used by judges in determining sentencing and bail
• Scores derived from 137 questions answered by defendants or pulled from criminal records:
  • “Was one of your parents ever sent to jail or prison?”
  • “How many of your friends/acquaintances are taking drugs illegally?”
  • “How often did you get in fights while at school?”
  • Agree or disagree? “A hungry person has a right to steal”
  • Agree or disagree? “If people make me angry or lose my temper, I can be dangerous.”
  • Race is **not** one of the questions
• The exact method of determining the score is kept as a trade secret
Case Study: COMPAS

- African Americans are almost twice as likely as Caucasians to be incorrectly labeled as high risk.
- Software predictions can have real consequences.

Table 1: ProPublica Analysis of COMPAS Algorithm

<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>

Example: Bias in Word Embeddings (Bolukbasi et al. 2016)

- Studied word2vec word embeddings trained on Google News
- word2vec represents each word as a high-dimensional vector
- Vector arithmetic can be used to answer analogies like:
  - Paris : France ≅ London : England

- Other analogies with stereotyped answers:
  - man : woman ≅ programmer : homemaker
  - man : woman ≅ surgeon : nurse

Image from: https://www.analyticsvidhya.com/blog/2017/06/word-embeddings-count-word2vec/
Example: Bias in Image Classification

- Images from imSitu visual semantic role labeling (vSRL) dataset
  - Only 33% of cooking images are of men
  - Prediction with a (biased) conditional random field only predicts men in 16% of cooking images

Algorithmic Fairness

• How can we ensure that our algorithms act in ways that are fair?
  • This definition is vague and somewhat circular
  • Describes a broad set of problems, not a specific technical approach

• Related to ideas of:
  • **Accountability**: who is responsible for automated behavior? How do we supervise/audit machines that have large impact?
  • **Transparency/Explainability**: why does an algorithm behave in a certain way? Can we understand its decisions? Can it explain itself?
  • **AI safety**: how can we make AI without unintended negative consequences?
  • **Aligned AI**: How can AI make decisions that align with our values?
Why Fairness is Hard

• Suppose we are a bank trying to fairly decide who should get a loan i.e. Who is most likely to pay us back?

• Suppose we have two groups: A and B (the sensitive attribute)
  • This is where discrimination could occur

• The simplest approach is to remove the sensitive attribute from the data, so that our classifier doesn’t know the sensitive attribute

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Employed?</th>
<th>Zip Code</th>
<th>Requested Amount</th>
<th>A or B?</th>
<th>Grant Loan?</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>F</td>
<td>Yes</td>
<td>24729</td>
<td>$50,000</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>Yes</td>
<td>11038</td>
<td>$30,000</td>
<td>B</td>
<td>Yes</td>
</tr>
<tr>
<td>72</td>
<td>F</td>
<td>No</td>
<td>10038</td>
<td>$90,000</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>39</td>
<td>F</td>
<td>Yes</td>
<td>30499</td>
<td>$70,000</td>
<td>A</td>
<td>No</td>
</tr>
<tr>
<td>45</td>
<td>M</td>
<td>No</td>
<td>20199</td>
<td>$60,000</td>
<td>B</td>
<td>No</td>
</tr>
<tr>
<td>68</td>
<td>M</td>
<td>Yes</td>
<td>30029</td>
<td>$50,000</td>
<td>B</td>
<td>No</td>
</tr>
</tbody>
</table>
Legally Recognized “Protected classes”

- **Race** (Civil Rights Act of 1964)
- **Color** (Civil Rights Act of 1964)
- **Sex** (Equal Pay Act of 1963; Civil Rights Act of 1964)
- **Religion** (Civil Rights Act of 1964)
- **National origin** (Civil Rights Act of 1964)
- **Citizenship** (Immigration Reform and Control Act)
- **Age** (Age Discrimination in Employment Act of 1967)
- **Pregnancy** (Pregnancy Discrimination Act)
- **Familial status** (Civil Rights Act of 1968)
- **Disability status** (Rehabilitation Act of 1973; Americans with Disabilities Act of 1990)
- **Veteran status** (Vietnam Era Veterans' Readjustment Assistance Act of 1974; Uniformed Services Employment and Reemployment Rights Act)
- **Genetic information** (Genetic Information Nondiscrimination Act)
Why Fairness is Hard

- However, this won’t work if the sensitive attribute is correlated with others
  - E.g., it is easy to predict race given other info (home address, financials, etc.)
- We need more sophisticated approaches

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Employed?</th>
<th>Zip Code</th>
<th>Requested Amount</th>
<th>A or B?</th>
<th>Grant Loan?</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>F</td>
<td>Yes</td>
<td>24729</td>
<td>$50,000</td>
<td>?</td>
<td>Yes</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>Yes</td>
<td>11038</td>
<td>$30,000</td>
<td>?</td>
<td>Yes</td>
</tr>
<tr>
<td>72</td>
<td>F</td>
<td>No</td>
<td>10038</td>
<td>$90,000</td>
<td>?</td>
<td>Yes</td>
</tr>
<tr>
<td>39</td>
<td>F</td>
<td>Yes</td>
<td>30499</td>
<td>$70,000</td>
<td>?</td>
<td>No</td>
</tr>
<tr>
<td>45</td>
<td>M</td>
<td>No</td>
<td>20199</td>
<td>$60,000</td>
<td>?</td>
<td>No</td>
</tr>
<tr>
<td>68</td>
<td>M</td>
<td>Yes</td>
<td>30029</td>
<td>$50,000</td>
<td>?</td>
<td>No</td>
</tr>
</tbody>
</table>

Based on materials by David Madras
Group Fairness

• Key idea: “Treat different groups equally”

• Assess fairness based on demographic parity: require that the same percentage of A and B receive loans
  • What if 80% of A is likely to repay, but only 60% of B is?
  • Then demographic parity is too strong

• Could require equal false positive/negative rates
  • When we make an error, the direction of that error is equally likely for both groups
    • $P(\text{loan} | \text{no repay, A}) = P(\text{loan} | \text{no repay, B})$
    • $P(\text{no loan} | \text{would repay, A}) = P(\text{no loan} | \text{would repay, B})$

Based on materials by David Madras
Individual Fairness

• Key idea: “Treat similar examples similarly”

• Learn fair representations
  • Useful for classification, not for (unfair) discrimination
  • Related to domain adaptation
  • Generative modelling/adversarial approaches

Based on materials by David Madras
Looking Forward

• This is an open and active area of research
• Lots of progress, long way to go
• Law will catch up with ML technology eventually