Fairness in Machine Learning: An Overview

David Madras

Machine Learning Group, University of Toronto

November 27, 2017
AI affects our lives in many ways

- 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*
Example — COMPAS

- Fairness is morally and legally motivated
- Takes many forms
- Criminal justice: recidivism algorithms (COMPAS)
  - Predicting if a defendant should receive bail
  - Unbalanced false positive rates: more likely to wrongly deny a black person bail

Table 1: ProPublica Analysis of COMPAS Algorithm

<table>
<thead>
<tr>
<th></th>
<th>White</th>
<th>Black</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wrongly Labeled</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High-Risk</td>
<td>23.5%</td>
<td>44.9%</td>
</tr>
<tr>
<td>Low-Risk</td>
<td>47.7%</td>
<td>28.0%</td>
</tr>
</tbody>
</table>

Fairness is morally and legally motivated
Takes many forms
Bias found in word embeddings (Bolukbasi et al. 2016)
- Examined word embeddings (word2vec) trained on Google News
- Represent each word with high-dimensional vector
- Vector arithmetic: analogies like Paris - France = London - England
- Found also: man - woman = programmer - homemaker = surgeon - nurse

The good news: word embeddings learn so well!
The bad news: sometimes too well
Our chatbots should be less biased than we are
**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 **accountability**: who is responsible for automated behaviour? How do we supervise/audit machines which have large impact?
- Also **transparency**: why does an algorithm behave in a certain way? Can we understand its decisions? Can it explain itself?
- Connections to **AI safety** and **aligned AI**: how can we make AI without unintended negative consequences? Aligns 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 2: To Loan or Not to Loan?

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Postal Code</th>
<th>Req Amt</th>
<th>A or B?</th>
<th>Pay</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>F</td>
<td>M5E</td>
<td>$300</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>M4C</td>
<td>$1000</td>
<td>B</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>M</td>
<td>M3H</td>
<td>$250</td>
<td>A</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>F</td>
<td>M9C</td>
<td>$2000</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>F</td>
<td>M3B</td>
<td>$200</td>
<td>A</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>M</td>
<td>M5W</td>
<td>$1500</td>
<td>B</td>
<td>0</td>
</tr>
</tbody>
</table>
Why Fairness is Hard

- However, if the sensitive attribute is correlated with the other attributes, this isn’t good enough
- It is easy to predict race if you have lots of other information (e.g. home address, spending patterns)
- More advanced approaches are necessary

<table>
<thead>
<tr>
<th>Age</th>
<th>Gender</th>
<th>Postal Code</th>
<th>Req Amt</th>
<th>A or B?</th>
<th>Pay</th>
</tr>
</thead>
<tbody>
<tr>
<td>46</td>
<td>F</td>
<td>M5E</td>
<td>$300</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>M4C</td>
<td>$1000</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>33</td>
<td>M</td>
<td>M3H</td>
<td>$250</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>34</td>
<td>F</td>
<td>M9C</td>
<td>$2000</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>71</td>
<td>F</td>
<td>M3B</td>
<td>$200</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>28</td>
<td>M</td>
<td>M5W</td>
<td>$1500</td>
<td>?</td>
<td>0</td>
</tr>
</tbody>
</table>
So we’ve built our classifier . . . how do we know if we’re being fair?

One metric is *demographic parity* — requiring 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

\[
P(\text{loan}|\text{no repay}, A) = P(\text{loan}|\text{no repay}, B) \\
P(\text{no loan}|\text{would repay}, A) = P(\text{no loan}|\text{would repay}, B)
\]

These are definitions of *group fairness*

“Treat different groups equally”
Definitions of Fairness — Individual Fairness

- Also can talk about *individual fairness* — “Treat similar examples similarly”
- Learn fair representations
  - Useful for classification, not for (unfair) discrimination
  - Related to domain adaption
  - Generative modelling/adversarial approaches

Figure 1: “The Variational Fair Autoencoder” (Louizos et al., 2016)

(a) Unfair representations
(b) Fair(er) representations
Conclusion

- This is an exciting field, quickly developing
- Central definitions still up in the air
- AI moves fast — lots of (currently unchecked) power
- Law/policy will one day catch up with technology
- Those who work with AI should be ready
Thank you!