Putting Artificial Intelligence Back into People’s Hands

Toward an Accessible, Transparent and Fair AI
Agenda

• How to create accessible Artificial Intelligence?
• Can AI be transparent and accurate?
• How to build fairness into AI?
Artificial Intelligence accessibility
Leveraging other models: fine-tuning
Bigger models are not more accurate

How to make AI accessible?

- Make it easy to reuse the model's parameters
- Release the training code and datasets under a Free licence
- Consider computational complexity when designing the model
Artificial Intelligence transparency
AI is used for critical matters

- Loan approval
- Justice
- Healthcare
- Self-driving cars
Why do we want AI transparency?

- Allows to interpret the result
- Builds trust in the model
- Helpful for debugging
- We require people to justify themselves
Parameters are not meant to be transparent

---

THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.
LIME: Debugging and selecting models
Local Interpretable Model-Agnostic Explanations

Algorithm 1
Words that A1 considers important:
- GOD
- mean
- anyone
- this
- Koresh
- through

Predicted: Atheism
Prediction correct: √

Algorithm 2
Words that A2 considers important:
- Posting
- Host
- Re
- by
- in
- Nntp

Predicted: Atheism
Prediction correct: √

Document
From: pauld@verdix.com (Paul Durbin)
Subject: Re: DAVID CORESH IS! GOD!
Nntp-Posting-Host: sarge.hq.verdix.com
Organization: Verdix Corp
Lines: 8

Making sense of images classification
How it works?

Original Image
P(tree frog) = 0.54

Perturbed Instances

<table>
<thead>
<tr>
<th>Instance</th>
<th>P(tree frog)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.85</td>
</tr>
<tr>
<td></td>
<td>0.00001</td>
</tr>
<tr>
<td></td>
<td>0.52</td>
</tr>
</tbody>
</table>

Locally weighted regression

Explanation

oreilly.com, Local Interpretable Model-Agnostic Explanations (LIME): An Introduction
Also for tabular data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capital Gain</td>
<td>0.00</td>
</tr>
<tr>
<td>Age</td>
<td>19.00</td>
</tr>
<tr>
<td>Hours per week</td>
<td>30.00</td>
</tr>
<tr>
<td>Marital Status=Never-married</td>
<td>True</td>
</tr>
<tr>
<td>Education-Num</td>
<td>9.00</td>
</tr>
</tbody>
</table>
Artificial Intelligence fairness
Protecting "car color" is easy

<table>
<thead>
<tr>
<th>Brand</th>
<th>Seats</th>
<th>Year</th>
<th>Color</th>
<th>Speed (km/h)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>5</td>
<td>2011</td>
<td>blue</td>
<td>150</td>
</tr>
<tr>
<td>B</td>
<td>2</td>
<td>2012</td>
<td>black</td>
<td>200</td>
</tr>
<tr>
<td>C</td>
<td>5</td>
<td>2010</td>
<td>red</td>
<td>250</td>
</tr>
</tbody>
</table>
Protecting gender is not easy

<table>
<thead>
<tr>
<th>gender</th>
<th>hobby</th>
<th>education</th>
<th>salary</th>
</tr>
</thead>
<tbody>
<tr>
<td>male</td>
<td>jogging</td>
<td>CS degree</td>
<td>35k</td>
</tr>
<tr>
<td>female</td>
<td>artistic swimming</td>
<td>self-taught</td>
<td>37k</td>
</tr>
<tr>
<td>female</td>
<td>women's volleyball team</td>
<td>PhD</td>
<td>35k</td>
</tr>
<tr>
<td>male</td>
<td>scuba-diving</td>
<td>CS degree</td>
<td>37k</td>
</tr>
</tbody>
</table>

⚠️ Think about correlation before removing an attribute
Why an algorithm can be unfair?

- Bias in the data itself
- Trained with the wrong metrics (bias by proxy)
- Bad prediction model
- Bias is hard to notice
- “With great power comes great responsibility” (Peter Parker)
Vocabulary

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
- Positive Predicted Values (PPV)
- Negative Predicted Values (NPV)
<table>
<thead>
<tr>
<th></th>
<th>All Defendants</th>
<th>Black Defendants</th>
<th>White Defendants</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Survived</td>
<td>2681</td>
<td>1282</td>
<td>990</td>
</tr>
<tr>
<td>Recidivated</td>
<td>1216</td>
<td>2035</td>
<td>532</td>
</tr>
<tr>
<td><strong>FP rate:</strong></td>
<td>32.35</td>
<td></td>
<td>44.85</td>
</tr>
<tr>
<td><strong>FN rate:</strong></td>
<td>37.40</td>
<td></td>
<td>27.99</td>
</tr>
<tr>
<td>PPV:</td>
<td>0.61</td>
<td></td>
<td>0.63</td>
</tr>
<tr>
<td>NPV:</td>
<td>0.69</td>
<td></td>
<td>0.65</td>
</tr>
<tr>
<td>LR+:</td>
<td>1.94</td>
<td></td>
<td>1.61</td>
</tr>
<tr>
<td>LR-:</td>
<td>0.55</td>
<td></td>
<td>0.51</td>
</tr>
</tbody>
</table>

[propublica.org, How We Analyzed the COMPAS Recidivism Algorithm (23 May 2016)](https://propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm)
Racial bias in US healthcare

A plausible fair loss function

Let $k$ be the number of values of a protected attribute
Let $f: y_{pred}, y_{true} \rightarrow s \in [0, 1]$ be a fairness function

$$loss = loss + \lambda \sum_{i=0}^{k} w_i f_i(y_{pred}, y_{true}) \left( \frac{\text{min}_{\forall i \in [0,k]} f_i(y_{pred}, y_{true})}{f_i(y_{pred}, y_{true})} \right)$$
Thank you! Questions?