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
What is a neural network?
Leveraging other models: fine-tuning
Bigger models are not more accurate

How to make AI accessible?

- Make it easy to reuse the model (ONNX format)
- Release the training code and the dataset under a Free licence
- Consider the number of FLOP when designing the model
Artificial Intelligence transparency
AI is used for critical matters

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

- Allows to interpret the result
- Builds trust in the model
- Makes debugging easier
Parameters are not meant to be transparent

xkcd.com
LIME: Debugging and selecting models
Local Interpretable Model-agnostic Explanations

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 does it work?

Original Image
P(tree frog) = 0.54

Perturbed Instances | P(tree frog)
--- | ---
| | 0.85
| | 0.00001
| | 0.52

Locally weighted regression

Explanation

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

Prediction probabilities

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;=50K</td>
<td>1.00</td>
</tr>
<tr>
<td>&gt;50K</td>
<td>0.00</td>
</tr>
</tbody>
</table>

<=50K

- Capital Gain <=...
  - 0.69
- Age <= 28.00
  - 0.11
- Hours per week <...
  - 0.10
- Marital Status...
  - 0.10
- Education-Num...
  - 0.09

>50K

<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 colors 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>female</td>
<td>women’s volleyball team</td>
<td>CS degree</td>
<td>35k</td>
</tr>
<tr>
<td>male</td>
<td>football team captain</td>
<td>self-taught</td>
<td>37k</td>
</tr>
<tr>
<td>male</td>
<td>chess</td>
<td>CS degree</td>
<td>37k</td>
</tr>
</tbody>
</table>

Think about correlation before removing an attribute
Vocabulary

- True Positive (TP)
- True Negative (TN)
- False Positive (FP)
- False Negative (FN)
## COMPAS recidivism scoring

<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>FP rate</td>
<td>32.35</td>
<td></td>
<td>44.85</td>
</tr>
<tr>
<td>FN rate</td>
<td>37.40</td>
<td></td>
<td>27.99</td>
</tr>
</tbody>
</table>

propublica.org (2016)
Racial bias in healthcare

Why an algorithm can be unfair?

- Bias in the input data itself
- Training with the wrong metric (bias by proxy)
- Bad prediction model
- Bias is hard to notice
- "With great power comes great responsibility" (Peter Parker)
A 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

$$loss = loss + \lambda \sum_{i=0}^{k} w_i f_i(y_{pred}, y_{true})$$

$$\min_{\forall i \in [0, k]} f_i(y_{pred}, y_{true})$$
Thank you! Questions?