INTERPRETING AND AUDITING MACHINE-LEARNING ALGORITHMS

Sorelle Friedler
Haverford College
How is a model making its decision?
How is a model making its decision?

- …for one person?
- …for all people?
How is a model made?

1. Training Data → Algorithm / Model creation → model
2. Personal Info → model → Decision
How does bias happen in model creation?

BBC Trending

Facebook challenges legitimacy of some Native names

By BBC Trending
What's popular and why

3 March 2015

Dane Lone Hill had trouble using her real name on Facebook

When Lance Browneyes of the Oglala Lakota community in South Dakota was blocked from Facebook for using a “fake” name, he submitted proof of his identification. Facebook then changed his name to Lance Brown.

# A hypothetical case study

## Training Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Top 1000 baby name</th>
<th>Dictionary word?</th>
<th>Real Name?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorelle Friedler</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Lady Gaga</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Big Bird</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Barack Obama</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Dana Lone Hill</td>
<td>yes</td>
<td>yes</td>
<td>???</td>
</tr>
</tbody>
</table>
Interpretable models

Personal Info

<table>
<thead>
<tr>
<th>Name</th>
<th>Phone</th>
<th>Email</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dana</td>
<td>Lone</td>
<td>Hill</td>
</tr>
<tr>
<td>yes</td>
<td>yes</td>
<td>???</td>
</tr>
</tbody>
</table>

model

- dictionary word?
  - yes
  - not real
    - yes
    - real name
  - no

Decision
Dear various parents, grandparents, co-workers, and other "not computer people."

We don't magically know how to do everything in every program. When we help you, we're usually just doing this:

START

FIND A MENU ITEM OR BUTTON WHICH LOOKS RELATED TO WHAT YOU WANT TO DO.

I CAN'T FIND ONE

PACK ONE AT RANDOM

I'VE TRIED THEM ALL

HAVE YOU BEEN TRYING THIS FOR OVER HALF AN HOUR?

NO

CLICK IT

YES

ASK SOMEONE FOR HELP OR GIVE UP

DID IT WORK?

NO

YOU'RE DONE

YES

GOOGLE THE NAME OF THE PROGRAM PLUS A FEW WORDS RELATED TO WHAT YOU WANT TO DO. FOLLOW ANY INSTRUCTIONS.

Please print this flowchart out and tape it near your screen. Congratulations; you're now the local computer expert!

https://xkcd.com/627/
Interpretable models?

http://dx.doi.org/10.1038/nature17439
Interpretable models

Cosmo Quiz

Is Your BFF Really on Your Side?

1. If you and your friend meet a cute guy you’re both into, what would she do after you declare that you like him?
   - A. Suggest you go for her and offer to be her wingwoman
   - B. Tell him she’s been checking him out

2. When she balls on plans to hang out with you to get together with her boyfriend instead, you’re:
   - A. Heartbroken. She’s happened many times before.
   - B. Touched. She always made such a big deal about putting friends first.
   - C. Disappointed yet aware that you’ve both failed occasionally.

3. You just came up with the coolest idea for a blog. How would she react to it when you tell her?
   - A. You should usually share stuff like that with her because she can be critical of your ideas.
   - B. That’s so enthusiastic, and it’s sooo smart that you haven’t thought through yet.
   - C. She’s 100 percent behind it and offers to help, no questions asked.

4. When you two have a disagreement, how does it usually get resolved?
   - A. You both speak your piece then back off and forget about it.
   - B. You end up going — it’s so much easier than arguing with her.

5. What role does your BFF usually take when you two are in public together?
   - A. Number two fan (the one to stay by your side and let you be the topic of conversation)
   - B. Complimentary to your thought view. She takes pride in you to get a laugh, but it’s all in good fun.
   - C. Partner in crime. You are not equal at the bar but do check in with each other often.

7 TO 10 POINTS
Frenemy in Disguise

She seems so interested in you but she’s really just out of reach in your book. But guess what? She does it so well, she’s been right under your nose this whole time! You didn’t think you could bring the bigger personality on anything worth the drive, but she more than matches you. She’s not just a friend, but a real girl who loves you the way you are. She’s supportive, warm, and always there for you. Now you can treat her like the BFF she is.

4 TO 6 POINTS
Crush-Case

You and her are on the same wavelength. You’re both so good at keeping your feelings to yourself and not taking risks. You’re both just too hard to get past. You should try your best to make her feel special and show her how much she means to you.

0 TO 3 POINTS
Too Faithful-Friend

You two are like hungry and faithful dogs. You’ve been through thick and thin together, but you didn’t expect that things are not as a little too easy and safe. Sometimes you just need to be there for a friend who you can’t help but trust. Remember, your friendship is important to keep it strong.

Log in to test yourself with more fun and insightful quizzes!
Interpretable models - SLIM

<table>
<thead>
<tr>
<th>Condition/Feature</th>
<th>Points</th>
<th>Condition/Feature</th>
<th>Points</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>age at release 18 to 24</td>
<td>2 points</td>
<td>prior arrests $\geq 5$</td>
<td>2 points</td>
<td>$+$</td>
</tr>
<tr>
<td>prior arrest for misdemeanor</td>
<td>1 point</td>
<td>no prior arrests</td>
<td>-1 point</td>
<td>$+$</td>
</tr>
<tr>
<td>age at release $\geq 40$</td>
<td>-1 point</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**ADD POINTS FROM ROWS 1–5**

<table>
<thead>
<tr>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\ldots \ldots$</td>
</tr>
</tbody>
</table>

What if using an interpretable model doesn’t make sense?

Interpretable models don’t always achieve the same level of accuracy as other models – there may be a tradeoff.

Revealing the model isn’t always possible.

If we have access to run the model, we can still find out some information about how it’s making decisions!
Audit options

Assumes access to appropriate input data and the ability to run the model and examine the outputs.

• Create an interpretable model of the model – use the predicted outputs as labels. Note: this is not the same model!
• Audit for direct influence - replace the feature with random noise and test the deterioration of the model.
• Audit for indirect influence – remove the feature and information about that feature contained in other features (e.g., proxy variables) and test the deterioration of the model.

Direct and Indirect Influence Audits
Synthetic Data (decision tree)

Synthetic Data:
A: item i number
B: 2i, C: -i
Constant, Random

Outcome:
  first half of items 1
  second half 2

Direct Influence:
A: 0.5
B: 0
Constant: 0

Indirect Influence:
A: 0.5
B: 0.5
Constant: 0

Direct vs. Indirect Influence Audits

Amazon Doesn’t Consider the Race of Its Customers. Should It?

By David Ingold and Spencer Soper
April 21, 2016

Zip code is a proxy for race.

http://www.bloomberg.com/graphics/2016-amazon.same-day/
Audit specifically for non-discrimination

Theorem:

the information content of a feature can be estimated by trying to predict it from the remaining features

If a protected feature can’t be predicted from the remaining features, then the information from that feature can’t influence the outcome of the model.

Audit: Build a classifier to try to predict the protected feature from the remaining training data. If the error is high, any trained model is non-discriminatory.

Policy points

• It’s possible to create interpretable models!
• Choosing interpretable models restricts model design choice, which may lower accuracy.
• We can audit a model even if it’s not interpretable:
  • model the model
  • direct influence
  • indirect influence
  • goal-specific audit (e.g., non-discrimination)
THANKS!

find me at:

sorelle@cs.haverford.edu
sorelle.friedler.net