Fairness, Accountability, and Transparency

Machine Learning: Jordan Boyd-Graber
University of Maryland

NEED FOR INTERPRETABILITY
Trust Part of ML Pipeline

Learn model  Trust model  Deploy model

Trust AI system

Make better decisions

Improve model

Data
Evaluate
Features
Model

Improve
ML is Everywhere

- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment
ML is Everywhere

- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment
ML is Everywhere

- Authorizing credit
- Sentencing guidelines
- Prioritizing services
- College acceptance
- Suggesting medical treatment
- How do we know it isn’t being incompetent/evil?
To predict and serve?
Kristian Lum, William Isaac
First published: 7 October 2016  Full publication history

Discrimination in Online Ad Delivery
Latanya Sweeney
Harvard University
latanya@fas.harvard.edu
January 28, 2013

Abstract
Uber seems to offer better service in areas with more white people. That raises some tough questions.

Facebook Lets Advertisers Exclude Users by Race

Machine Bias
There's software used across the country to predict future criminals. And it's biased against blacks.
by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 25, 2016
Keep it Simple (Stupid)

- Clear preference for interpretability
- Even at the cost of performance: decision trees still popular
- But what about all of the great machine learning we’ve talked about?
Pneumonia Example (Caruana)

- **Prediction task:**
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)

- Most accurate ML method: multitask neural nets
Pneumonia Example (Caruana)

- Prediction task:
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)

- Most accurate ML method: multitask neural nets
- Used logistic regression
Pneumonia Example (Caruana)

- **Prediction task:**
  - LOW Risk: outpatient: antibiotics, call if not feeling better
  - HIGH Risk: admit to hospital (10% of pneumonia patients die)

- Most accurate ML method: multitask neural nets
- Used logistic regression
- Learned rule: HasAsthma($x$) $\rightarrow$ LessRisk($x$)
Why?

- asthmatics presenting with pneumonia considered very high risk
- receive aggressive treatment and often admitted to ICU
- history of asthma also means they often go to healthcare sooner
- treatment lowers risk of death compared to general population
Lessons Learned (Caruana)

- Always going to be risky to use data for purposes it was not designed for
  - Most data has unexpected landmines
  - Not ethical to collect correct data for asthma
- Much too difficult to fully understand the data
  - Our approach is to make the learned models as intelligible as possible for task at hand
- Experts must be able to understand models in critical apps like healthcare
  - Otherwise models can hurt patients because of true patterns in data
  - If you don’t understand and fix model it will make bad mistakes
- Same story for race, gender, socioeconomic bias
  - The problem is in data and training signals, not learning algorithm
- Only solution is to put humans in the machine learning loop
We’ve already seen problems

- Gender/racial bias
- Generalization failures
- Malicious Input
We’ve already seen problems

- Gender/racial bias
- Generalization failures
- Malicious Input
Can we just remove problematic variables?

- Not obvious *a priori*
- Can find correlated features
- More of a problem in deep learning
Subject for Today

- How to measure interpretability
- How to fix biased data
- How to unbias supervised algorithms