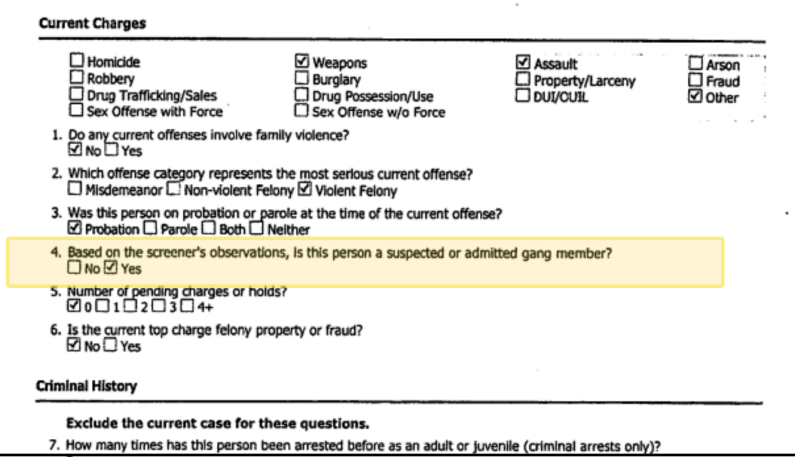
Fairness in Machine Learning

Mark Craven and David Page
Computer Sciences 760
Spring 2018

www.biostat.wisc.edu/~craven/cs760/

The COMPAS system

- used by many governments (including state of Wisconsin) to predict risk that those convicted of crimes will commit future crimes
- scores derived from 137 questions that are either answered by defendants or pulled from criminal records.



Current Charges

☐ Homicide ☐ Robbery ☐ Drug Trafficking/Sales ☐ Sex Offense with Force
☑ Weapons ☐ Burglary ☐ Drug Possession/Use ☐ Sex Offense w/o Force
☑ Assault ☐ Property/Larceny ☐ DUI/CUIL
☐ Arson ☐ Fraud ☑ Other

1. Do any current offenses involve family violence?
☑ No ☐ Yes
2. Which offense category represents the most serious current offense?
☐ Misdemeanor ☐ Non-violent Felony ☑ Violent Felony
3. Was this person on probation or parole at the time of the current offense?
☑ Probation ☐ Parole ☐ Both ☐ Neither
4. Based on the screener's observations, is this person a suspected or admitted gang member?
☐ No ☑ Yes
5. Number of pending charges or holds?
☑ 0 ☐ 1 ☐ 2 ☐ 3 ☐ 4+
6. Is the current top charge felony property or fraud?
☑ No ☐ Yes

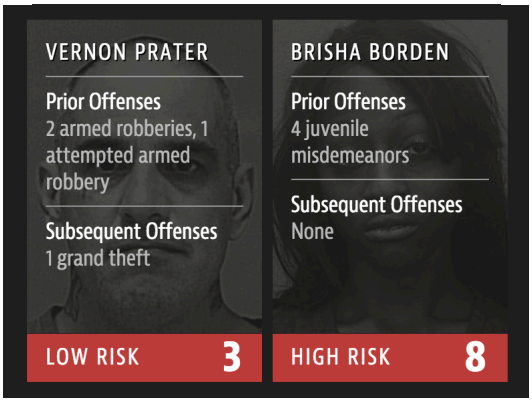
Criminal History

Exclude the current case for these questions.

7. How many times has this person been arrested before as an adult or juvenile (criminal arrests only)?

The COMPAS system

- ProPublica obtained the risk scores assigned to > 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over next 2 years



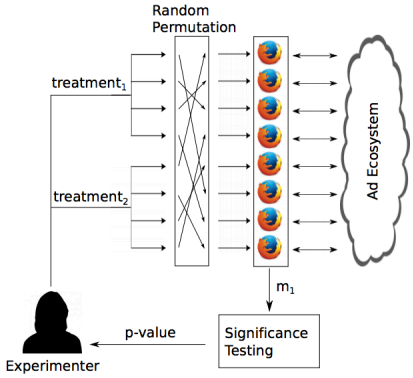
PROPUBLICA

The COMPAS system

- ProPublica obtained the risk scores assigned to > 7,000 people arrested in Broward County, Florida, in 2013 and 2014 and checked to see how many were charged with new crimes over next 2 years
- The system was particularly likely to falsely flag black defendants as future criminals
  - wrongly labeling them this way at almost twice the rate as white defendants
  - white defendants were mislabeled as low risk more often than black defendants

Google Ads Settings

- Datta et al. [PPET 2015] studied how user behaviors, Google's ads, and Ad Settings interact
- Setting gender to female in Google Ad Settings made it less likely that user would be shown ads for high paying jobs



Isn't discrimination the point of machine learning?

Yes, but we should be aware of

- unjustified bases for discrimination
- legal reasons to avoid unjust discrimination
- moral reasons to avoid unjust discrimination

Certain domains are legally regulated

- credit, education, employment, housing, public accommodation

Certain classes are legally protected in specific contexts

- race, color, sex, religion, national origin, citizenship, age, pregnancy, familial status, disability status, veteran status, genetic information

See http://mrtz.org/nips17/ for more detail

How does unfair bias arise in machine learning systems?

- selection, sampling, reporting bias in the data set
- bias in the objective function

Biases in data sets example

Garg et al. [*PNAS* 2017] "Word embeddings quantify 100 years of gender and ethnic stereotypes"

- tested relationships among concepts in Google word2vec vectors
- e.g. relatedness of occupations and words representing gender

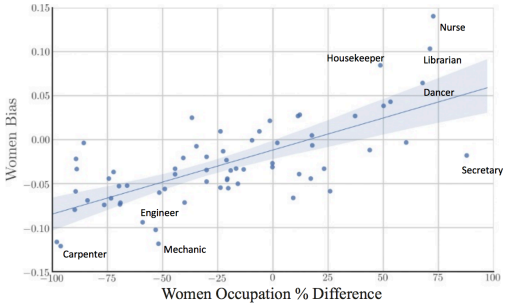


Fig. 1. Women's occupation relative percentage vs. embedding bias in Google News vectors. More positive indicates more associated with women on both axes. $P < 10^{-10}$, $r^2 = 0.499$. The shaded region is the 95% bootstrapped confidence interval of the regression line. In this single embedding, then, the association in the embedding effectively captures the percentage of women in an occupation.

Biases in data sets example

Garg et al. [*PNAS* 2017] "Word embeddings quantify 100 years of gender and ethnic stereotypes"

Table 1. The top 10 occupations most closely associated with each ethnic group in the Google News embedding

Hispanic	Asian	White
Housekeeper	Professor	Smith
Mason	Official	Blacksmith
Artist	Secretary	Surveyor
Janitor	Conductor	Sheriff
Dancer	Physicist	Weaver
Mechanic	Scientist	Administrator
Photographer	Chemist	Mason
Baker	Tailor	Statistician
Cashier	Accountant	Clergy
Driver	Engineer	Photographer

Table 2. Top adjectives associated with women in 1910, 1950, and 1990 by relative norm difference in the COHA embedding

1910	1950	1990
Charming	Delicate	Maternal
Placid	Sweet	Morbid
Delicate	Charming	Artificial
Passionate	Transparent	Physical
Sweet	Placid	Caring
Dreamy	Childish	Emotional
Indulgent	Soft	Protective
Playful	Colorless	Attractive
Mellow	Tasteless	Soft
Sentimental	Agreeable	Tidy

How to achieve fairness in ML

1. Blindness approach: don't use features that enable unfair classifications/predictions
   - this approach is generally not effective; the data usually contains many surrogates for such protected features
   - e.g. the COMPAS system does not explicitly use race
   - e.g. word embeddings case illustrates a lot of dependence between gender words and other words

How to achieve fairness in ML

2. Group fairness approach
   - given two groups, G_1 and G_2
   - enforce that $P(Outcome = o \mid G_1) \approx P(Outcome = o \mid G_2)$

How to achieve fairness in ML

3. Individual fairness approach
   - treat similar individuals similarly
   - $\mathrm{f}(\boldsymbol{x}^{(i)}) \approx \mathrm{f}(\boldsymbol{x}^{(j)}) \mid d(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}) \approx 0$
   - where $d: X \times X \to \mathbb{R}$ is a distance metric for individuals

An individual fairness approach

[Dwork et al. ITCS 2012]

- model outputs a probability distribution over set of outcomes $P(y \mid \boldsymbol{x})$
- the notion of individual fairness can be captured by a (D, d)-Lipschitz property

$$D\big(P(y \mid \boldsymbol{x}^{(i)}),\ P(y \mid \boldsymbol{x}^{j})\big) \leq d\big(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)}\big)$$

  where D is a distance measure for distributions

- learning is then a constrained optimization problem