Intro to Fairness + Bias in Classification

CREDIT TO

CS 294: Fairness in Machine Learning at Berkeley
Instructor: Moritz Hardt

https://mrtz.org/nips17/#/
https://vimeo.com/248490141

Background

- Pro-publica article about automated sentencing in 2016: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
  - More false positives related to black defendants.
- Since then, many conflicting analyses of bias in COMPAS
  - Northpointe: Classifications are calibrated and reflect training data: https://www.documentcloud.org/documents/2998391-ProPublica-Commentary-Final-070616.html
  - Neill et. al: Bias relates more strongly to female defendants without priors than black defendants: https://arxiv.org/abs/1611.08292

So … huh?

Bias in Classification

Bias in classifiers impacts:

- resource allocation (COMPAS is just one example)
- identity construction and associated opportunities (Latanya Sweeney, Joy Buolamwini) https://www.radcliffe.harvard.edu/video/race-technology-and-algorithmic-bias-vision-justice

NIPS 2017 Keynote on the topic:
https://www.youtube.com/watch?v=fMym_BKWQzk

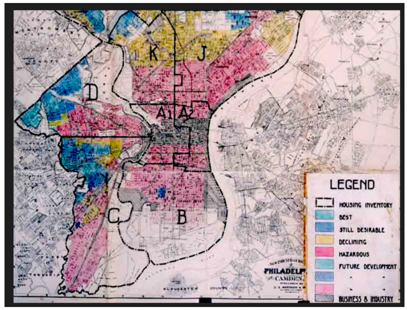
Formal classification: pros and cons

Formalizing decision making can limit opportunities to exercise prejudicial discretion or fall victim to implicit bias

"Automated underwriting increased approval rates for minority and low-income applicants by 30% while improving the overall accuracy of default predictions"

Gates, Perry, Zorn (2002)

Formal classification: pros and cons



But, of course, formal procedures can just as easily encode or reinforce bias. Example: Redlining
https://en.wikipedia.org/wiki/Redlining

So what is a classifier?

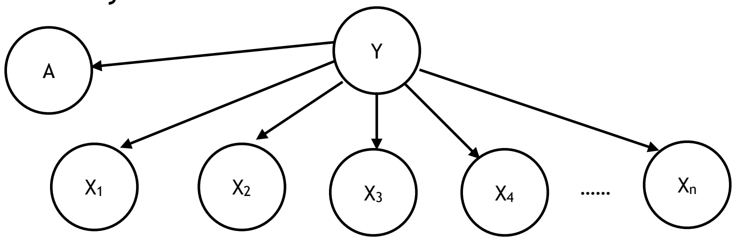
Assume a classifier relies on:

- X - features of an individual (browsing history etc.)
- A - features that include sensitive attributes (e.g. gender)
- Y - target variable or 'label' (what you want to predict)
- C - a function, C(X,A), which returns a binary classification (Y')
    - Note that Y' may be a threshold of a value (R(X,A)) between 0 and 1

A classifier will be trained on data where you know Y, i.e. data that is labelled. The classification function could be something based on regression, for example, or something else.

A familiar looking classifier

Naive Bayes classifier



How can we use this structure to compute $P(Y|X_1,X_1,X_1\dots X_n, A)$?
How might we use this to make a binary classification?

Bias may start with your training data

Skewed sample: Example is predictive policing, which relies on reported incidents of crime. But reported incidents are not necessarily accurate!

Tainted examples: Labels in data might be unreliable. Performance reviews, for example, are forms of labels that already may be subject to bias.

Limited features: Some features may work well to classify one group (e.g. men) but not others (e.g. women).

Sample size disparity: If we have few examples from one group, we can't model the group accurately.

Proxies: Many features are correlated with "sensitive" features (e.g. use of Pinterest as proxy for gender).

B, Selbst (2016)

Adjusting for (coping with) bias

At the point of sampling

At the point of training

After training

Example: Placing Ads for Software Engineers

- X - features of an individual (e.g. browsing history)
- A - sensitive attribute (e.g. gender)
- C - C(X,A) binary predictor (show ad or not)
- Y - target variable ("is a Software Engineer")

Also: We may also have a score function $R=r(X,A) \in [0,1]$
This can be turned into (binary) predictor C by thresholding
e.g.
Bayes optimal score given by r(x,a) = the expected value of Y given X=x,A=a.

How can we enforce a lack of "bias"?

We can require:

Independence: C independent of A
Separation: C independent of A, conditional on Y
Sufficiency: Y independent of A, conditional on C

Independence

Means $P(C|A) = P(C)$ is the same for all values that A can take on, so
C doesn't depend on A.

This is sometimes called *demographic parity* or *statistical parity,*

e.g. "70% of all applicants received a mortgage regardless of gender or race."

Is this good?

Ignores possible correlation between Y and A.

Also, permits laziness:

We can accept “qualified” in one group, “random people” in other

And, allows us to trade false negatives for false positives.

Sufficiency

Y independent of A, conditional on R *(which we can threshold to create C)*

Sufficiency implied by *calibration by group*:

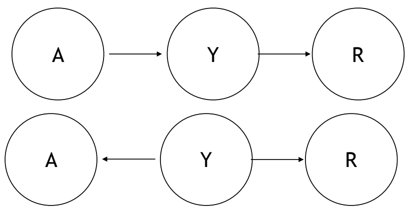
$P(Y=1 \mid R=r, A=a)=r$

Means if we have a risk score of 40%, there is a 40% chance that Y will be 1, on average.

Separation

Means C is independent of A, conditional on Y

So $P(C \mid Y=y, A=a) = P(C \mid Y=y)$



Separation

More specifically, call

False positives: $P(C = 1 \mid Y = 0, A)$, True positives: $P(C=1 \mid Y=1, A)$

1. We get *equalized odds* if both false and true positives are equal across groups
2. We get *equalized opportunity* if just true positives are equal across groups

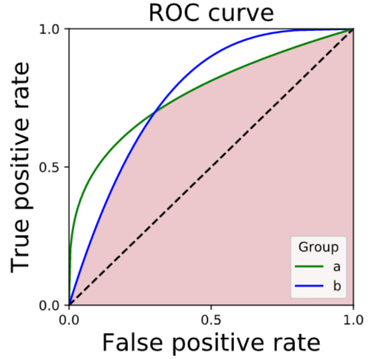
Is this good?

Possibly, as it forces us to distribute errors across groups (we can't be lazy)

We can strive to achieve this by post-processing
(i.e. by thresholding R in some way that may depend on A)

Or, we could try enforcing equal error distribution during data collection or when training (which is hard)

Separation



https://research.google.com/bigpicture/attacking-discrimination-in-ml/

Example: COMPAS data

Do we have Demographic Parity?

P(C=High Risk|African-American) = 0.28
P(C=High Risk|White) = 0.11
P(C=High Risk) = 0.21

…. no.

Example: COMPAS data

Do we have Sufficiency?

P(*Re-offender*|*C=High,A=White*)=P(*Re-offender*|*C=High,A=African-American*)=0.7
P(*Re-offender*|*C=Medium,A=White*)=P(*Re-offender*|*C=Medium,A=African-American*)=0.5
P(*Re-offender*|*C=Low,A=White*)=P(*Re-offender*|*C=Low,A=African-American*)=~0.3

…. more or less.

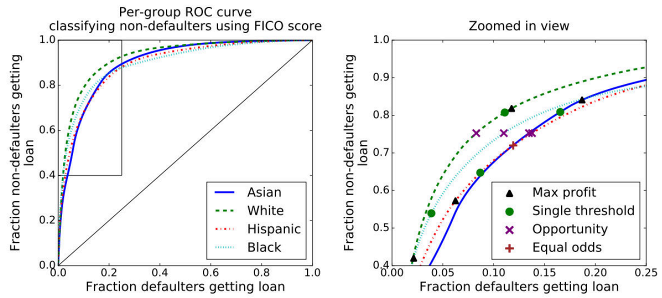
Example: COMPAS data

Do we have Separation?

P(*C=High|No Re-offence,A=White*) = 0.05
P(*C=High|No Re-offence,A=African-American*) = 0.16

…. no, not equalized odds.

Example: FICO scores



Max profit picks a threshold for each group the threshold that maximizes profit.
Race blind (single threshold) requires the threshold to be the same for each group.
Equal opportunity picks a threshold such that the fraction of non-defaulting group members that qualify for loans is the same.
Equalized odds requires the fraction of non-defaulters that qualify and the fraction of defaulters that qualify to be constant across groups

Of interest

Sufficiency, Independence and Separation are all mutually exclusive

You can't have them all. You have to choose one or the other!

Tradeoffs

Which tradeoff is "fair"?

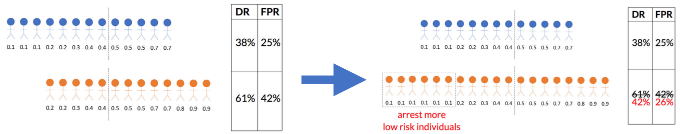
Pro-publica says:
COMPAS does not enforce **equality of odds**

Northpointe says:
But, we calibrated by group! We went for **sufficiency**, not **separation**.

All situations admit “unfair” practices

Calibration by group:

Based on averages in training data that may not reflect individuals. Those with “risk” of 0.4 will be re-offenders 40% of the time, on average.



Equality of odds: False positive rates can be adjusted by arresting more “low risk” people.