
NYU

# DS-GA 3001.009 Responsible Data Science Lab 3

Center for Data Science
Udita Gupta | Tandon School of Engineering

# Disparate Impact Remover

# Define Bias

- D: Data set with attributes X, Y
  - X: protected attribute (eg: race, gender etc)
  - Y: unprotected attributes
- Goal: determine outcome C (hiring, admission etc)
- Direct discrimination: C = f(X)
  - Female not hired for programming jobs
  - People of certain ethnicity not allowed to eat at restaurant
- Indirect discrimination: C = f(Y), but Y strongly correlates with X
  - Undergraduates with more than 10 years of programming are hired for job (most women don't start programming till college)

# Detect Bias

- Players: Alice and Bob, and data D = (X, Y)
- Goal: C = f(Y), C $\in \{0, 1\}$
- Alice wishes to compute C = f(Y) using secret algorithm (A)
- Alice and Bob both have D
- Bob must trust that Alice is not using X in her Algorithm A
- Bob must certify that no algorithm f will discriminate against X
- Discrimination Test (disparate impact)
    - $Pr[C = 1 \,|X = 0] \geq 0.8\, Pr[C = 1 \,|X = 1]$
- So the final question: Given an algorithm and given this above model and set-up, can we determine if the algorithm is liable for a claim of disparate impact i.e., is it implicitly discriminating against the protected group X.

# Definitions – We can detect bias!

- You need to use Balanced Error Rate (BER) as the error measurement while doing machine learning. It is a class conditioned error rate:
  - $BER(f(X), Y) = \frac{Pr[f(Y)=0 \mid X=1] + Pr[f(Y)=1 \mid X=0]}{2}$
- D is $\varepsilon\, predictable$ if we can predict X from Y with BER $\leq \varepsilon$
- D is biased if we can determine against X without using X (i.e. there exists a classifier that admits disparate impact)

# Repair the data

- Once you do detect the bias, can you repair the data?
- Steps to do so:
  - Build BER-optimized classifier to predict X from Y
  - Evaluate BER
  - If BER is above threshold (given by a theorem in the paper), D does not admit bias
- While repairing the data you would want to preserve it's utility but making them more fair at the same time.
  - Eg: You have some ordered attribute (say SAT scores) and you want to preserve the relative ranking. So even if I modify the scores I would still want them to be relatively in the same order
- You can also generalize this repair to the case where you don't want to repair the data entirely.
  - Eg: you want to merge 2 distributions but not completely.
  - Like a trade-off between the amount of information you retain from the original data vs not

# Useful Links

- AIF360 Toolkit Algo https://aif360.readthedocs.io/en/latest/modules/preprocessing.html#disparate-impact-remover

- Disparate Impact Remover paper: https://arxiv.org/pdf/1412.3756.pdf