Bias in Machine Learning

CS 584: Spring 2022

Sanmay Das
George Mason University

Bias Case Study 1: Language Models

- Word embeddings: Represent words as points in a high-dimensional vector space
  - Words that co-occur a lot in the same neighborhood will be more similar in the vector space
- Authors used a large-scale corpus from an Internet crawl and then measured similarities between target group words and standard lists of, e.g., pleasant and unpleasant words
  - Replicated known human biases: e.g.
    - European-American names are considered more “pleasant” than African-American names
    - Women’s names are more associated with “family” and men with “career”

Caliskan, Bryson, Narayanan; Science 2017

Related Cases in the News: Hiring and Selection

- Amazon case: Scrapped plans to use AI to screen resumes since it was showing bias towards men over women
- UT Austin CS PhD admissions used an assistant trained on a database of past PhD admissions, but scrapped it recently
  - Had cut total time reviewing by 74% (at least one faculty member still reviewed each applicant)
  - Letters of reference using words like “best” “award” “research” were highly rated, while “good” “class” “programming” “technology” were rated lower
  - Privileged degrees from elite institutions
  - Built in the biases of the last fully-human admissions committee!

Bias Case Study 2: Facial Analysis (Gender Shades)

- Use of automated face recognition is high (e.g. in law enforcement)
- Benchmark task: Gender recognition
- Facial analysis datasets are highly skewed, with 79.6% and 86.2% of two important benchmarks being lighter-skinned
- Authors created a new, more balanced benchmark dataset and also tested existing commercial classifiers on that

GENDER SHADES

Classifier	Metric	All	F	M	Darker	Lighter	DF	DM	LF	LM
MSFT	PPV(%)	93.7	89.3	97.4	87.1	99.3	79.2	94.0	98.3	**100**
	Error Rate(%)	6.3	10.7	2.6	12.9	0.7	**20.8**	6.0	1.7	0.0
	TPR (%)	93.7	96.5	91.7	87.1	99.3	92.1	83.7	**100**	98.7
	FPR (%)	6.3	8.3	3.5	12.9	0.7	**16.3**	7.9	1.3	0.0
Face++	PPV(%)	90.0	78.7	99.3	83.5	95.3	65.5	**99.3**	94.0	99.2
	Error Rate(%)	10.0	21.3	0.7	16.5	4.7	**34.5**	0.7	6.0	0.8
	TPR (%)	90.0	98.9	85.1	83.5	95.3	98.8	76.6	**98.9**	92.9
	FPR (%)	10.0	14.9	1.1	16.5	4.7	**23.4**	1.2	7.1	1.1
IBM	PPV(%)	87.9	79.7	94.4	77.6	96.8	65.3	88.0	92.9	**99.7**
	Error Rate(%)	12.1	20.3	5.6	22.4	3.2	**34.7**	12.0	7.1	0.3
	TPR (%)	87.9	92.1	85.2	77.6	96.8	82.3	74.8	**99.6**	94.8
	FPR (%)	12.1	14.8	7.9	22.4	3.2	**25.2**	17.7	5.20	0.4

Table 4: Gender classification performance as measured by the positive predictive value (PPV), error rate (1-PPV), true positive rate (TPR), and false positive rate (FPR) of the 3 evaluated commercial classifiers on the PPB dataset. All classifiers have the highest error rates for darker-skinned females (ranging from 20.8% for Microsoft to 34.7% for IBM).

Buolamwini and Gebru, FAccT 2018

Bias Case Study 3: Recidivism Prediction

- ProPublica story on COMPAS algorithm
  - Scores from 1-10, indicating risk of recidivism
  - Trained on features not including race
  - Used by judges to decide whether to release defendants awaiting trial
  - ProPublica's "smoking gun"

	White	African-American
Pr(Labeled high risk | Didn't reoffend, Race)	23.5%	44.9%
Pr (Labeled low risk | Reoffended, Race)	47.7%	28.0%

  - But…within each risk score category, proportion who re-offend is approximately equal, regardless of race

How is this possible?

Population: 70% of sample who are Black reoffend, 40% of sample who are White reoffend

	Black	White
Reoffenders	60 High Risk 10 Low Risk	12 High Risk 28 Low Risk
Non Reoffenders	10 High Risk 20 Low Risk	2 High Risk 58 Low Risk

Pr (High Risk | Not Reoffend, Black) = 1/3
Pr (High Risk | Not Reoffend, White) = .033
Pr (Low Risk | Reoffend, Black) = 1/7
Pr (Low Risk | Reoffend, White) = 0.7

	Black	White
High Risk	60 Reoffenders 10 Not	12 Reoffenders 2 Not
Low Risk	10 Reoffenders 20 Not	28 Reoffenders 58 Not

Pr (Reoffend | High Risk, Black) = 6/7
Pr (Reoffend | High Risk, White) = 6/7
Pr (Not reoffend | Low Risk, Black) = 0.667
Pr (Not reoffend | Low Risk, White) = 0.674

Possible Definitions

- Anti-classification: Don’t use protected attributes (race, gender, proxies) to make decisions.
- Classification parity: Measures like False Positive Rate and False Negative Rate are equal across subgroups defined by protected attributes
- Calibration: Outcomes are independent of protected attributes conditional on risk estimates.

Mathematical Formulation

- Decision rule $d : X \to \{0,1\}$
- Actions are lend/don't lend, release/don't release, etc.
- $X = (X_p, X_u)$: Protected, Unprotected
- $Y \in \{0,1\}$: Target of prediction: crime, default, etc.)
- $r(x) = \Pr(y = 1 \mid x)$ (True risk function)
- $s(x)$ is a risk score. Typically $d(x) = 1 \iff s(x) \geq t$ where *t* is some threshold

- Anti-classification:

$$d(x) = d(x') \quad \forall x, x' \text{ s.t. } x_u = x_{u'}$$

- Classification parity (e.g. FPR parity):

$$\Pr(d(x) = 1 \mid y = 0, x_p) = \Pr(d(x) = 1 \mid y = 0)$$

- Calibration:

$$\Pr(y = 1 \mid s(x), x_p) = \Pr(y = 1 \mid s(x))$$

An Impossibility Result

- Groups A and B (protected) of equal size and they have different re-offense rates
- High risk group
  - Re-offends $x_h\%$ of the time and is $y_h\%$ of the population
- Low risk group
  - Re-offends $x_l\%$ of the time and is $y_l\%$ of the population
- Theorem: If $x_h \neq x_l$ then **cannot get** both calibration and parity of FPRs at the same time from any classifier.

Problems With Fairness Definitions

- Anti-classification: Existence of proxies, times when maybe we *should* consider protected attributes (e.g. gender-specific recidivism)
- Classification parity: If different subpopulations have different risk distributions, thresholds will have to vary by group
- Calibration: Suppose you *want* to discriminate. Redlining (base score only on zip code, for example)
- Don’t forget about bias in the estimation of probabilities (e.g. from learning a classifier on a training set!)

Shaping Algorithmic Decision-Making

- There is no panacea. We have to carefully formulate goals
- Given a particular group fairness objective M, a common practice is to design machine learning algorithms that take this objective into account
  - So, for example, learn a model $\min_f E(f; D)$ subject to $|M(G_1) - M(G_2)| \leq \gamma$
  - Many algorithmic approaches. For example
    - GerryFair (Kearns et al, ICML 2018): e.g. set FPR to be within γ for subgroups
    - DI-Remove (Feldman et al, KDD 2015): Specifically tries to remove ability to predict subgroup, which has a connection to removing disparate impact