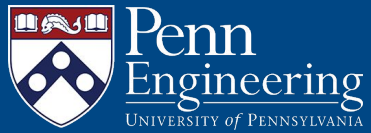


Fairness in Machine Learning

Harry Wang
harrywan@seas.upenn.edu
Computer and Information Science

STAT 991 – Topics in Modern Statistical Learning
2/15/2022

Overview

- Motivations
- ProPublica Case Study
- Fairness Definitions
- Methodologies
- ML Fairness in Natural Language Processing (NLP)

Motivations

- Machine Learning is everywhere!
- Used in high stake decisions sometimes
- Has potential to perpetuate societal bias

Fairness

- What is fairness?
- Somewhat of a philosophical question
- What do you think is a good definition of fairness?

Fairness

- Equality/Egalitarianism
- Meritocracy/Desert
- Protection from harm
- Traditional vs. Progressive Justice

Fairness

- Fairness and Justice are often related terms
- Legal dictionary definition of justice:

justice

n. 1) fairness. 2) moral rightness. 3) a scheme or system of law in which every person receives his/ her/its due from the system, including all rights, both natural and legal. One

Fairness

- We can think of fairness as a set of rules
- These rules have to be formally/quantitatively/algorithmically specified within the context of machine learning to construct algorithms that behave fairly
- Can be a sensitive topic given today's political climate, but is an important aspect of ML

Compas/ProPublica Case Study

- Northpointe's Algorithm Compas predicts whether criminal will recidivate
- ProPublica's study on Compas finds bias in the algorithm
- Criminals receive a risk scores from 1-10
  - Low Risk: 1-4
  - Medium Risk: 5-8
  - High Risk: 8-10

Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016

Compas Results

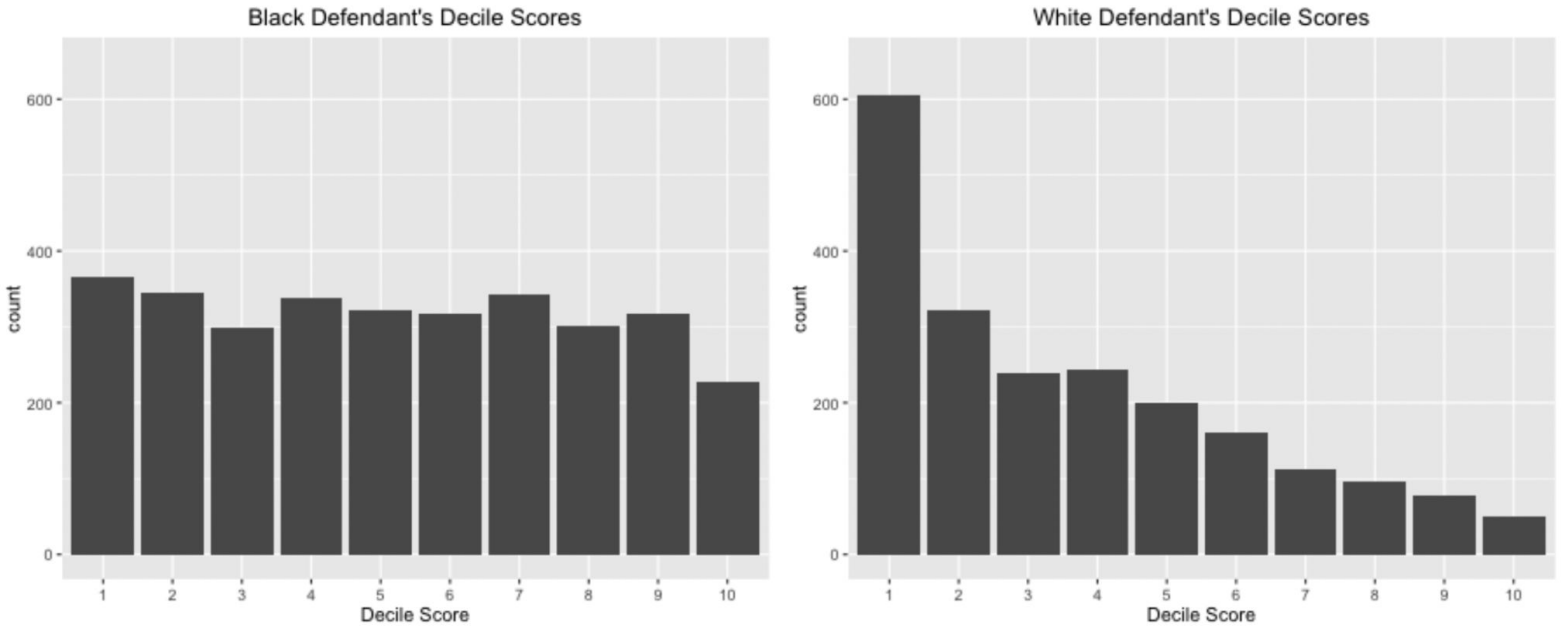
- ~62% Accuracy for both Black and White defendants
- Black defendants were 77.3% more likely than white defendants to receive a higher score, correcting for criminal history

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

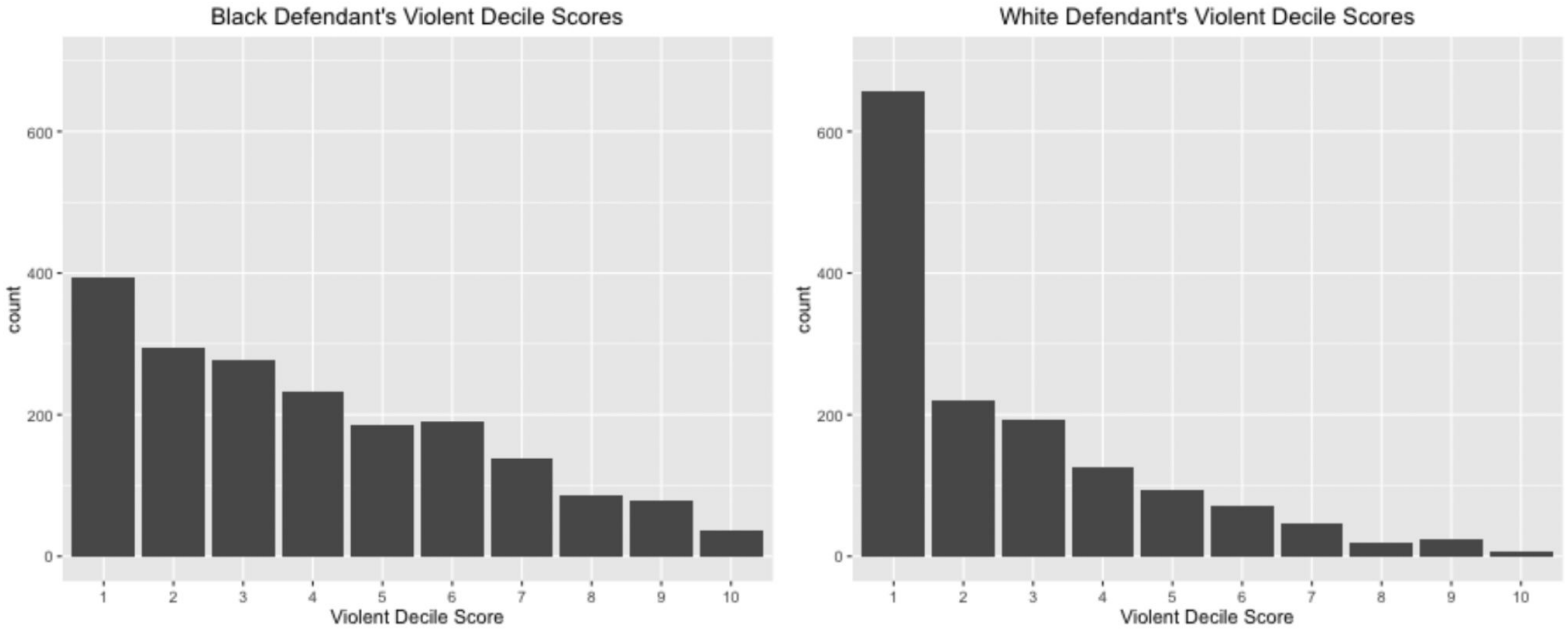
Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016

Compas Results



Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016

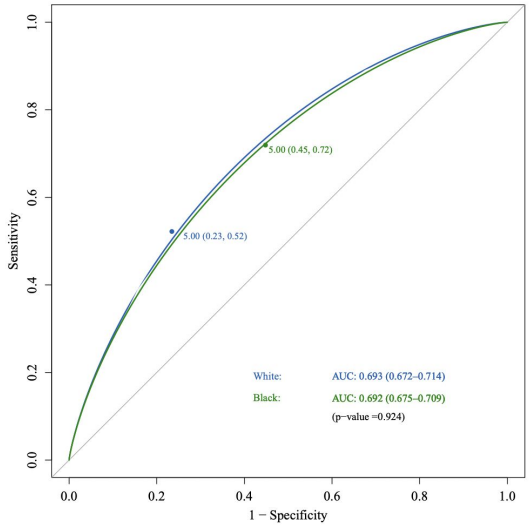
Compas Results



Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016

Compas Results

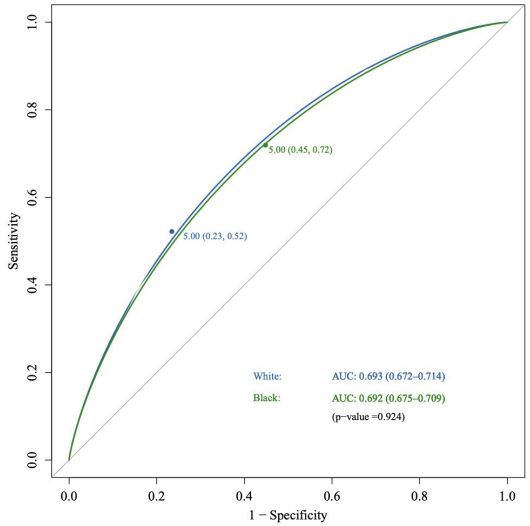
- ProPublica deemed Compas to be unfair due to different false positive and false negative rates amongst different groups
- Northpointe rebuttal: AUC is same amongst groups if two different thresholds are used



William Dieterich, Christina Mendoza, Tim Brennan
2016

Compas Results

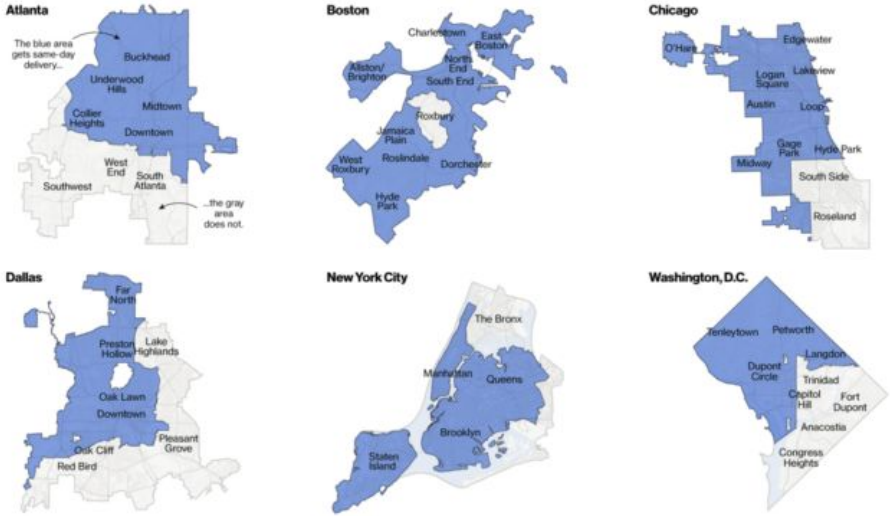
- Observation: different fairness criteria exist

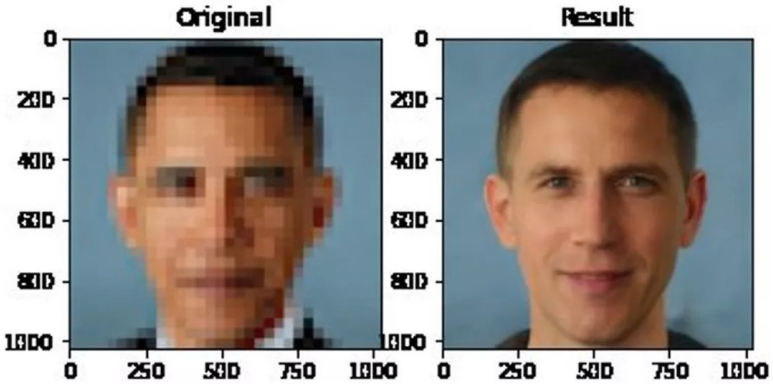


William Dieterich, Christina Mendoza, Tim Brennan
2016

Other example of ML Bias

- ML for allocating delivery services, face recognition, stationing police, etc.





Fairness Definitions

- There are various fairness definitions
- Many conflict with each other and cannot be simultaneously satisfied
- Individual vs. Group Fairness

$$\begin{aligned} \min_{\hat{h}} \quad & \epsilon(\hat{h}) \\ \text{s.t.} \quad & \text{some fairness constraint} \end{aligned}$$

Individual Fairness

- Individual Fairness: Similar individuals should have similar chance of receiving a certain label/output

$$(u_i, y_i) \sim \mathscr{D}$$

$$\cos(u_1, u_2) = \frac{u_1 \cdot u_2}{||u_1|| ||u_2||}$$

$$\cos(u_1, u_2) \leq \epsilon \implies |P(\hat{h}(u_1) = y) - P(\hat{h}(u_2) = y)| \leq \delta$$

$$\min \epsilon(\hat{h})$$

- Problem/Challenge: many different pairs of individuals to consider!
- Creates more constraints in optimization problem

Fairness Definitions

- Group Fairness: Ensure Fairness amongst groups (groups often defined by sensitive attribute)
- Group Fairness studied more often in research currently (less constraints in ML optimization problem)

Group Fairness Definitions

- Fairness Through Unawareness
- Demographic Parity
- Accuracy Parity
- Equal Opportunities
- Equalized Odds

Fairness Through Unawareness

- Given a set of sensitive attributes X (age, race, gender, zip code, etc), do not collect data with these attributes
- Seems like an intuitive first attempt
- US Laws/Policies enforce this in some industries
- What could go wrong?

Fairness Through Unawareness

- Data Leakage: there may be unforeseen proxies for sensitive variables in dataset (often the result of historical discrimination)

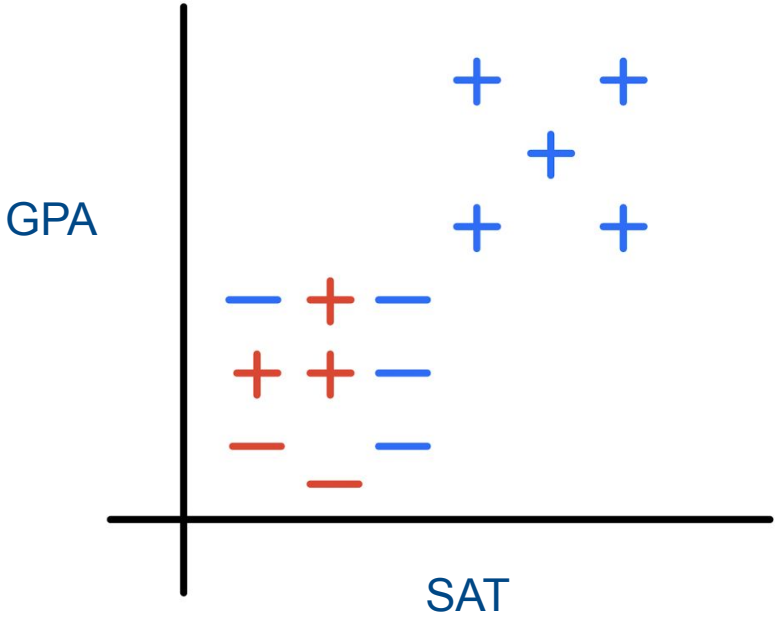
Sensitive Variable	Example Proxies
Gender	Education Level, Income, Occupation, Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), University Faculty, Working Hours
Marital Status	Education Level, Income
Race	Felony Data, Keywords in User Generated Content (e.g. CV, Social Media etc.), Zipcode

- Resulting Model may have similar levels of unfairness of “Sensitive Variable Blind” Model wrt sensitive feature of interest
- Makes it impossible to audit/evaluate fairness

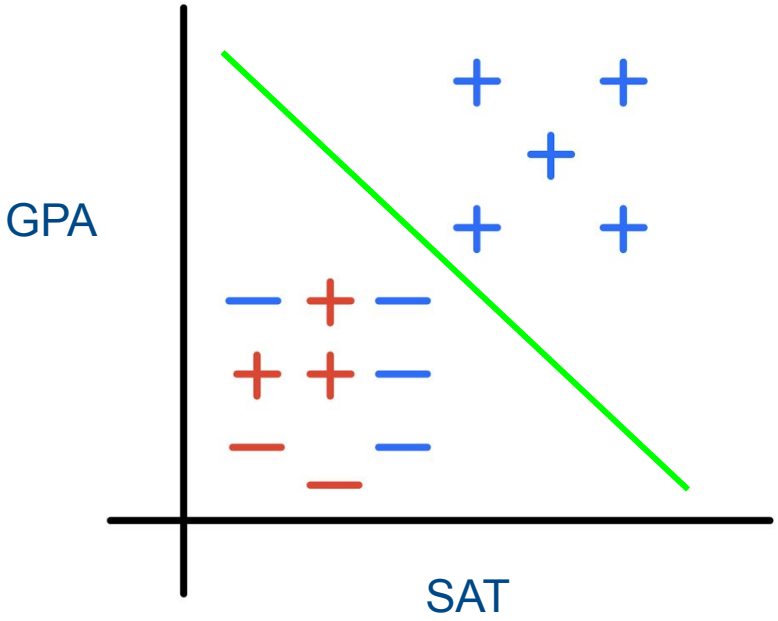
Example Problem: Predicting Undergraduate Admissions

- Given a set of features X (GPA, SAT score), predict whether a candidate will have a successful Undergraduate Career (graduating in a certain amount of time, with certain GPA)

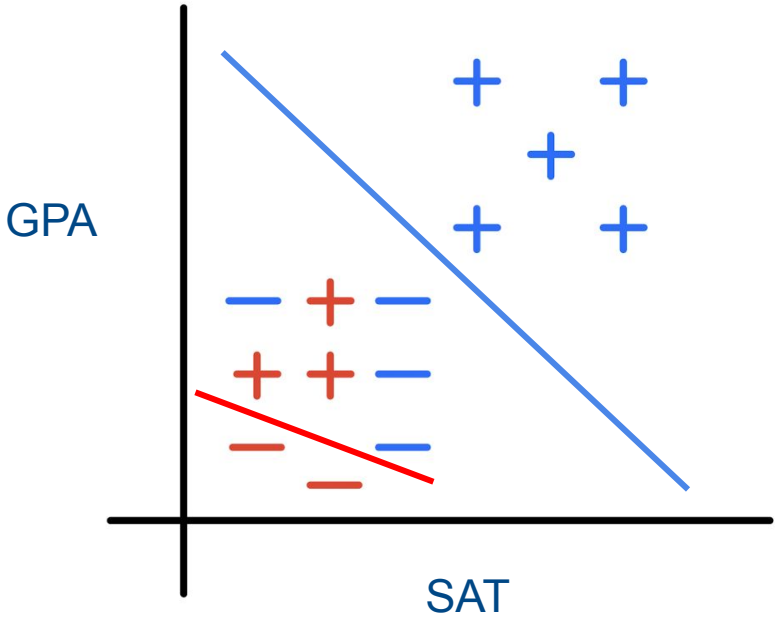
Example Problem: Predicting Undergraduate Admissions



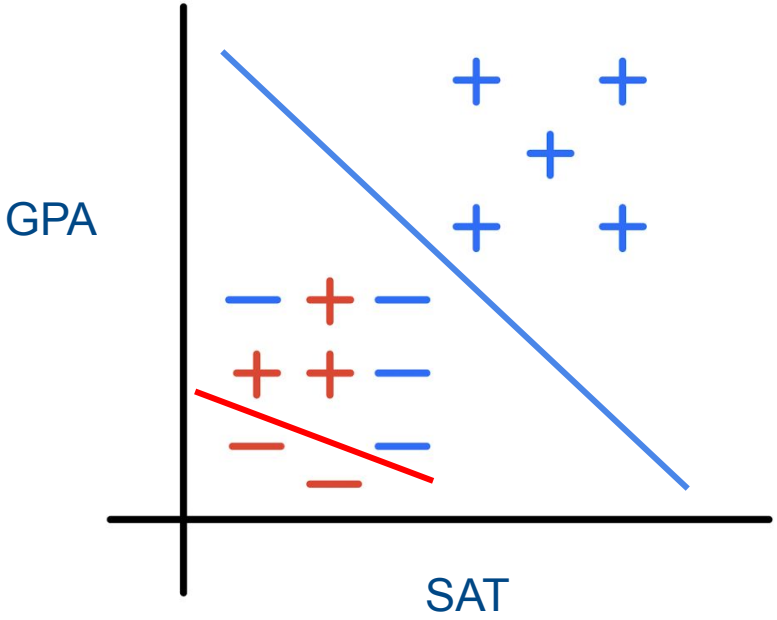
Example Problem: Predicting Undergraduate Admissions



Example Problem: Predicting Undergraduate Admissions



Example Problem: Predicting Undergraduate Admissions



- We actually need to explicitly specify the value of the sensitive variable to increase Accuracy, Recall, and Demographic Parity

Demographic Parity

- Explicitly takes sensitive variable into account
- Equalize “success (positive)” rates amongst groups
- Similar to “equality of outcomes”
- Given two groups A and B:

$$x[0] \in [A, B]$$

$$P(h(x) = 1 | x[0] = A) = P(h(x) = 1 | x[0] = B)$$

Demographic Parity

- Enforces distributed fairness amongst groups
- **Does not** enforce sensitive feature combinations
- Following constraints not enforced:

$$x[0] \in [A, B], x[1] \in [C, D]$$

$$\begin{aligned} & P(h(x) = 1 | x[0] = A, x[1] = C) \\ = & P(h(x) = 1 | x[0] = A, x[1] = D) \\ = & P(h(x) = 1 | x[0] = B, x[1] = C) \\ = & P(h(x) = 1 | x[0] = B, x[1] = D) \end{aligned}$$

Accuracy Parity

- Enforce accuracy of model across different groups

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Accuracy Parity Pitfalls in Compas

- ~62% Accuracy for both Black and White defendants
- Black defendants were 77.3 percent more likely than white defendants to receive a higher score, correcting for criminal history
- False positives in one group traded for false negatives in another group

Prediction Fails Differently for Black Defendants

	WHITE	AFRICAN AMERICAN
Labeled Higher Risk, But Didn't Re-Offend	23.5%	44.9%
Labeled Lower Risk, Yet Did Re-Offend	47.7%	28.0%

Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016

Equal Opportunity

- Equalize true positive rates amongst groups

$$x[0] \in [A, B]$$

$$P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 1, x[0] = A) = P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 1, x[0] = B)$$

Equalized Odds

- Equalize both true positives and false positive rates amongst groups
- More comprehensive than equal opportunity
- Harder constraints to satisfy

$$x[0] \in [A, B]$$

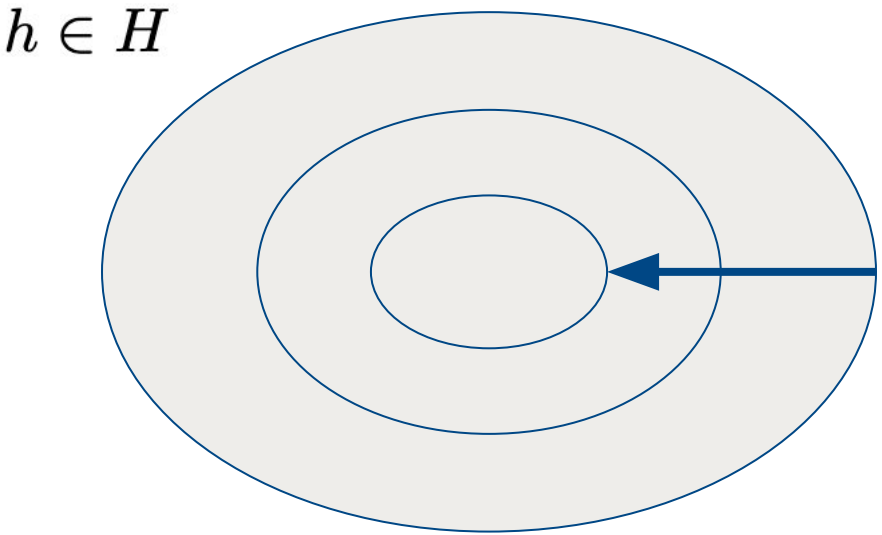
$$P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 1, x[0] = A) = P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 1, x[0] = B)$$

$$P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 0, x[0] = A) = P_{(x,y)\sim\mathscr{D}}(h(x) = 1 | y = 0, x[0] = B)$$

Trade-offs

- "Tolerate" a certain level of unfairness

$$|P(h(x) = 1|x[0] = A) - P(h(x) = 1|x[0] = B)| \leq \delta$$

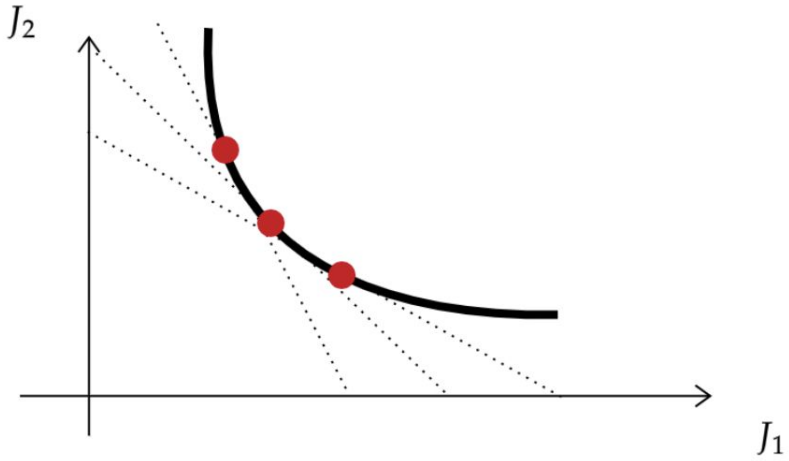


- Lower tolerance level leads to fewer candidate hypotheses that can optimize loss function

Trade-offs

- "Tolerate" a certain level of unfairness

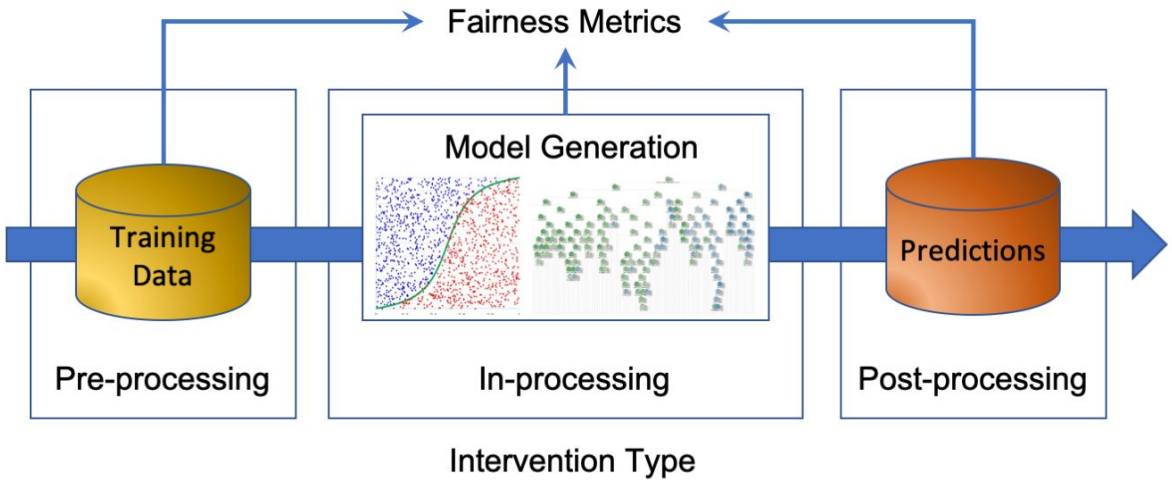
$$|P(h(x) = 1|x[0] = A) - P(h(x) = 1|x[0] = B)| \leq \delta$$



- Trade-off between error and fairness difficult to avoid in certain specific problems

Methodologies

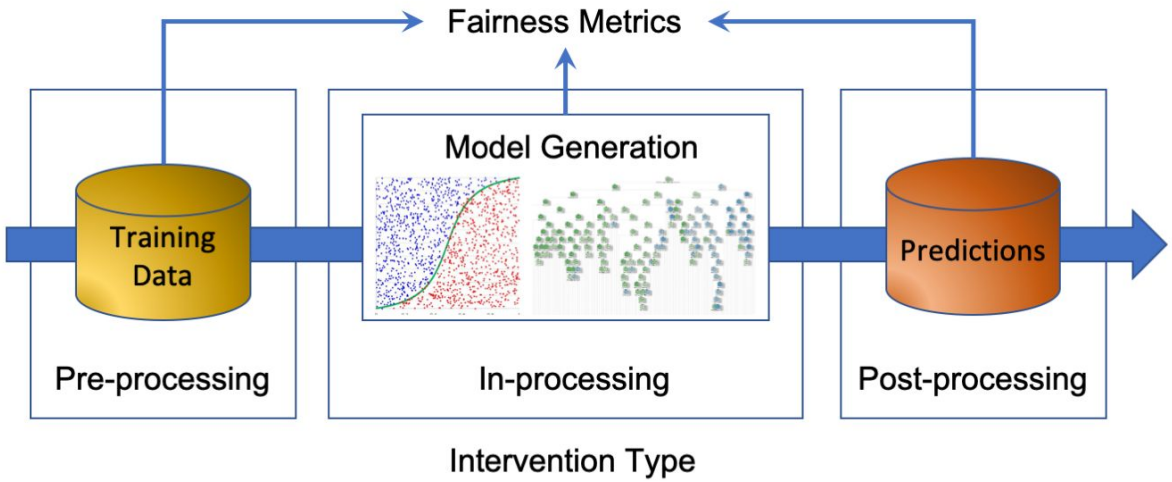
- Pre-processing: modify the training data itself
- In-processing: tweak model architecture to increase fairness
- Post-processing: add wrapper/extra component to model that ensures fairness



Simon Caton, Christian Haas 2020

Methodologies

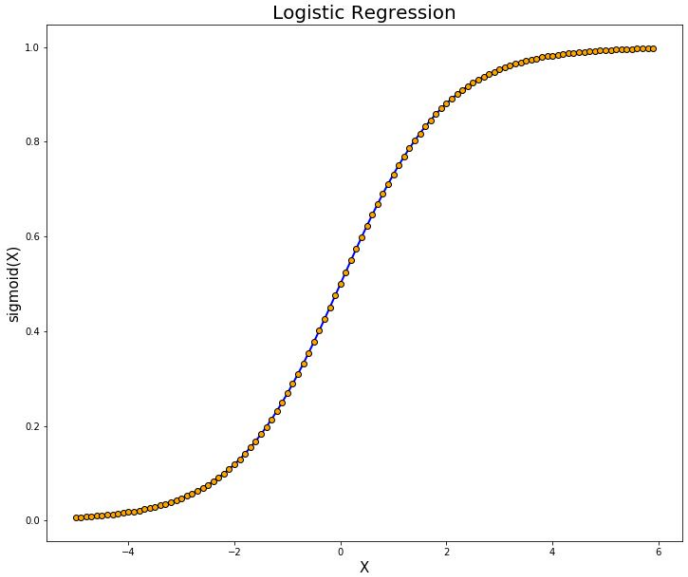
- How might we use uncertainty quantification in these methods?



Simon Caton, Christian Haas 2020

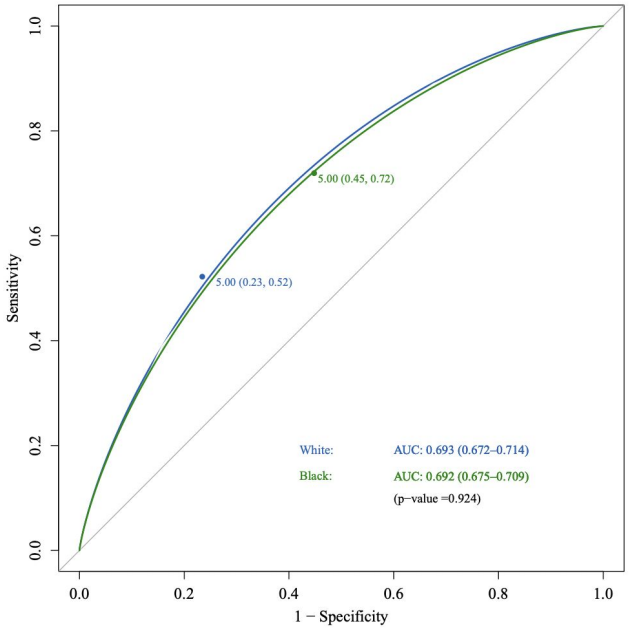
Post-Processing Logistic Regression Example

- Softmax output probability score
- Use different threshold for two different subgroups for labeling a point as positive (after the model has been trained) to ensure certain fairness constraint



Compas Results

- Used in Northpointe rebuttal against ProPublica



William Dieterich, Christina Mendoza, Tim Brennan
2016

Collecting better data

- Collect more representative data!
- Make sure that labels are balanced
- Make sure feature values are balanced
- Make sure labels and distribution of labels amongst groups are balanced

Re-weight Training Data

- Suppose a sensitive group has very few positive points
- Upweight minority data points

$$min \sum_{i=1}^{m} I(h(x) \neq y)$$

$$\beta < 1$$

Define $B^+ = \{x \in B \; s.t. \; y = 1\}$. Then let,

$$I'(h(x), y) = \begin{cases} \frac{1}{\beta} & h(x) \neq 1 \quad and \quad x \in B^+ \\ 0 & h(x) = 1 \quad and \quad x \in B^+ \\ I(h(x) \neq y) & otherwise \end{cases}$$

Avrim Blum,
Kevin Stangl
2019

Re-weight Training Data

- Penalizes False Negatives more for minority group

$$min \sum_{i=1}^{m} I(h(x) \neq y)$$

$$\beta < 1$$

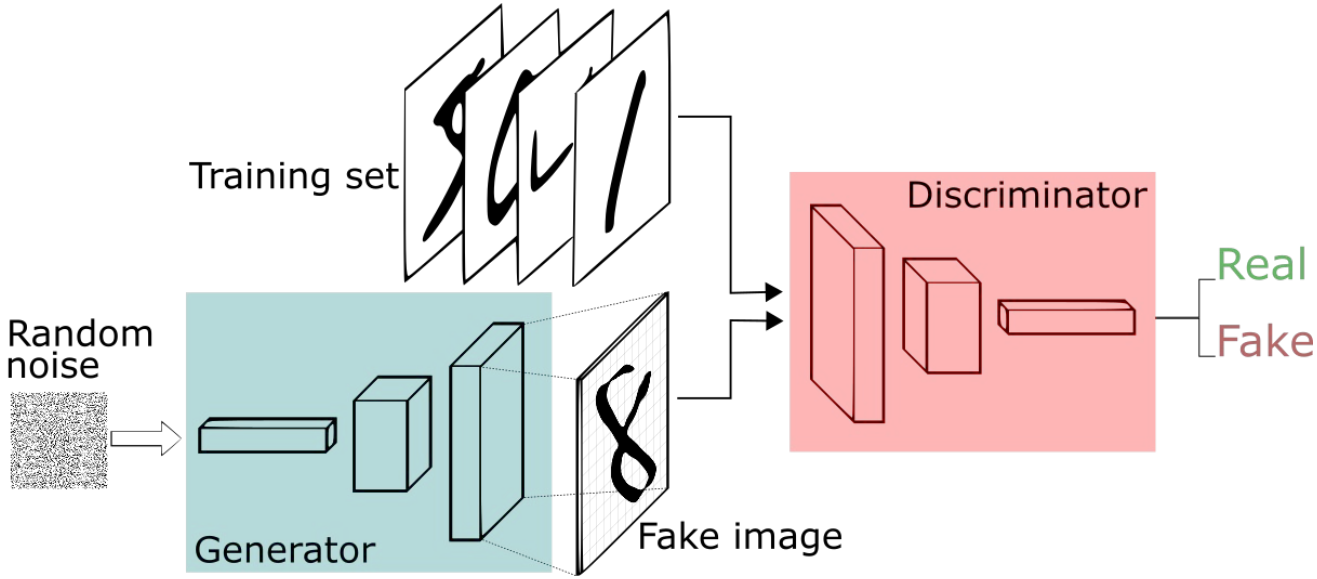
Define $B^+ = \{x \in B \; s.t. \; y = 1\}$. Then let,

$$I'(h(x), y) = \begin{cases} \frac{1}{\beta} & h(x) \neq 1 \quad and \quad x \in B^+ \\ 0 & h(x) = 1 \quad and \quad x \in B^+ \\ I(h(x) \neq y) & otherwise \end{cases}$$

Avrim Blum,
Kevin Stangl
2019

Generative Adversarial Network (GANs)

- Used to create fake images
- Two main components: Generator and Discriminator



Generative Adversarial Network (GANs)

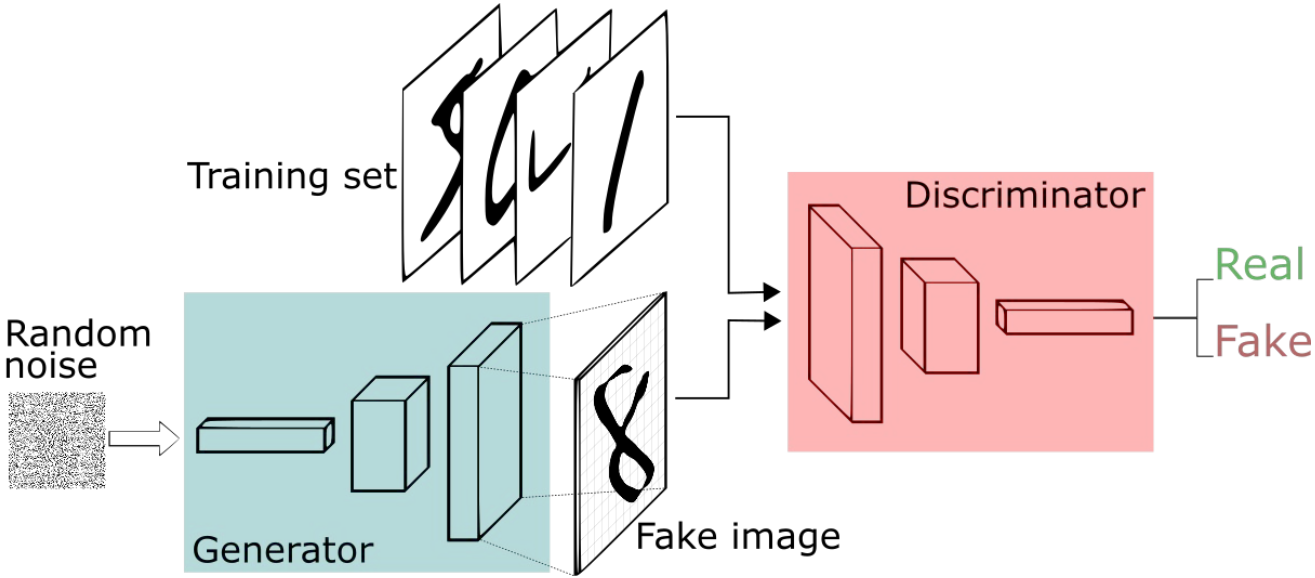
- Loss functions are based off of binary cross-entropy loss
- Generator creates new data points using noise while minimizing “reconstruction error”
- Discriminator tries to discriminate between generated and non-generated points

$$\text{Loss} = -\frac{1}{\substack{\text{output} \\ \text{size}}} \sum_{i=1}^{\substack{\text{output} \\ \text{size}}} y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log (1 - \hat{y}_i)$$

$$\min_G \max_D V(D, G) = \min_G \max_D \left(E_{x \sim P_{data}(x)}[logD(x)] + E_{z \sim P_z(z)}[log(1 - D(G(z)))] \right)$$

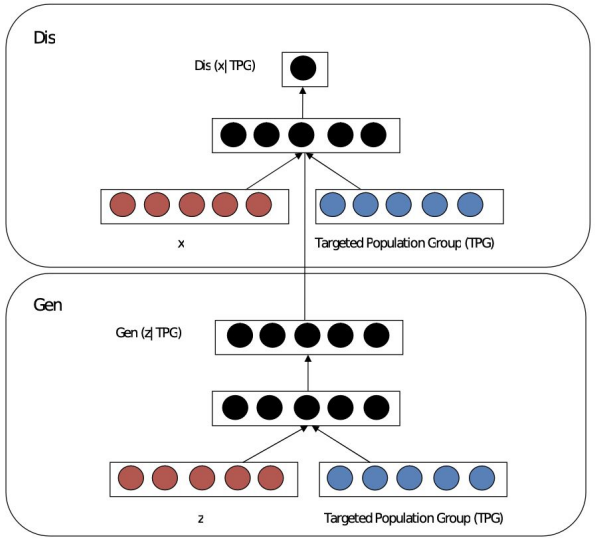
Generative Adversarial Network (GANs)

- Can also be used to create synthetic data in general



Conditional Generative Adversarial Network (cGANs)

- Generate Data with certain specified conditions
- Conditions specify properties of data-points involving Targeted Population Group (TPG)



$$min_{Gen} \ max_{Dis} V(Gen, Dis) =$$
$$\mathbf{E}_{x \sim p_{data}(x)} log[Dis(x|TPG)] + \mathbf{E}_{z \sim p_z(z)} log[1 - Dis(Gen(z|TPG))]$$

$$maximize \sum_{i=1}^{n} p_{data}(x_i|TPG) log(Dis(x_i|TPG))$$
$$Subject\ to: \ (1 - log(Dis(x_i|TPG)) \geq log(1/2), i = 1, ..., n$$
$$\boldsymbol{Dis} \in \mathscr{S},$$

Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019

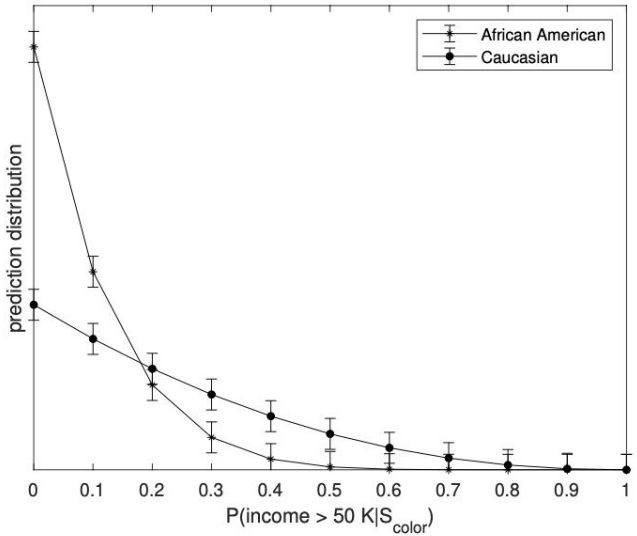
Conditional Generative Adversarial Network (cGANs)

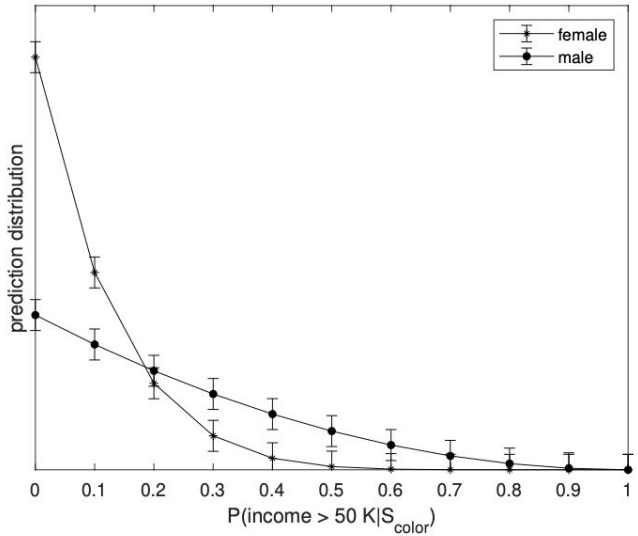
- Evaluation
  a. Use cGANS to generate synthetic data
  b. Train downstream ML classifier on original data
  c. Train downstream ML classifier on original data + generated data
  d. Compare Results

Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019

Conditional Generative Adversarial Network (cGANs)

- Neural Network trained (3 Hidden Layers + ReLU activation) on original data
- Predict income given race and gender

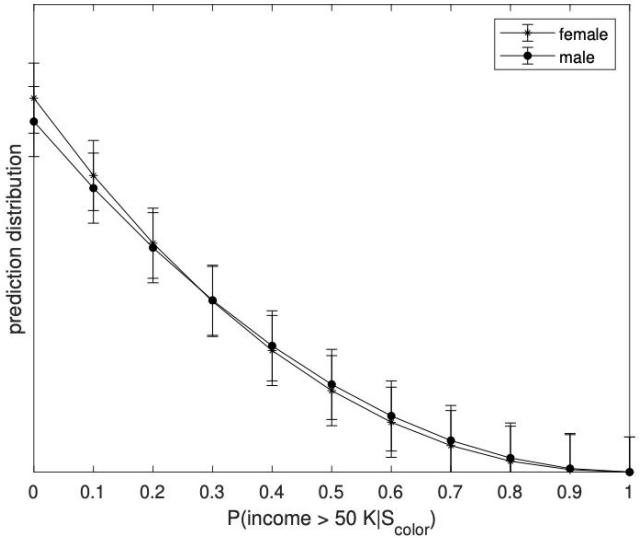


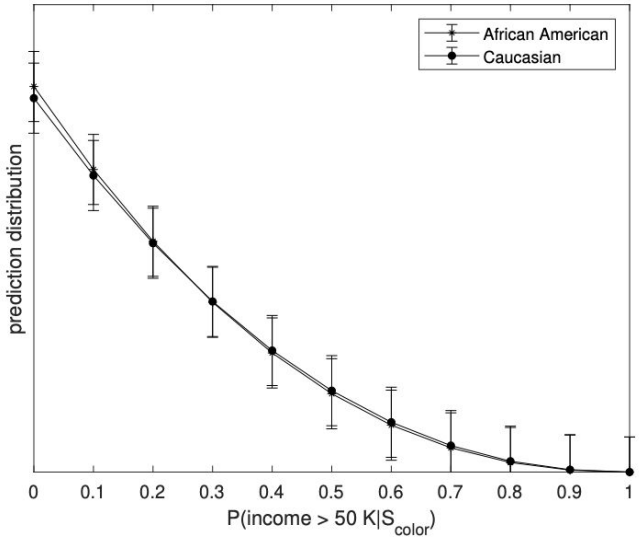


Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019

Conditional Generative Adversarial Network (cGANs)

- Neural Network trained (3 Hidden Layers + ReLU activation) on original data + 85% of generated data
- Distributions between demographics are much closer to each other!





Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019

Conditional Generative Adversarial Network (cGANs)

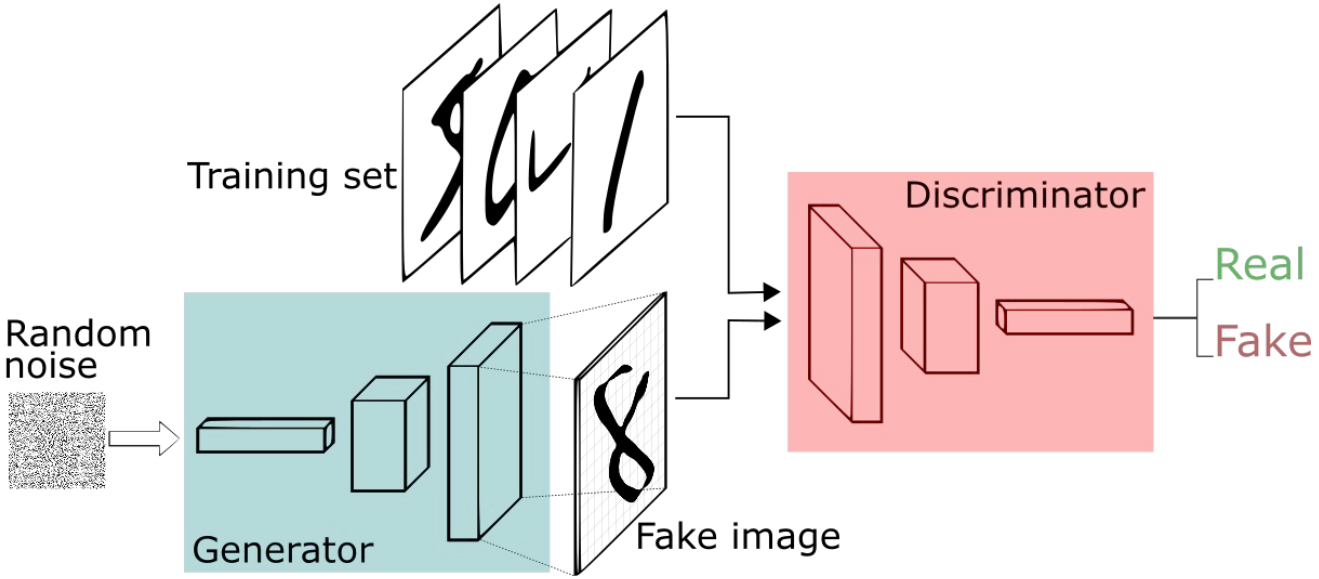
- Better accuracy as well!

	Acc. (300 HUs)	Acc. (500 HUs)	Acc. (700 HUs)	Acc. (900 HUs)
The Proposed Approach	**84.9± 1.14**	**85.1± 1.09**	**85.3± 1.92**	**85.5± 1.15**
Pivot-based Approach	76.1 ± 1.11	76.4 ± 1.84	77.1± 1.23	77.3± 1.78
Baseline	82.0± 1.16	82.3± 1.06	82.6 ± 1.90	82.9 ± 0.88

Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019

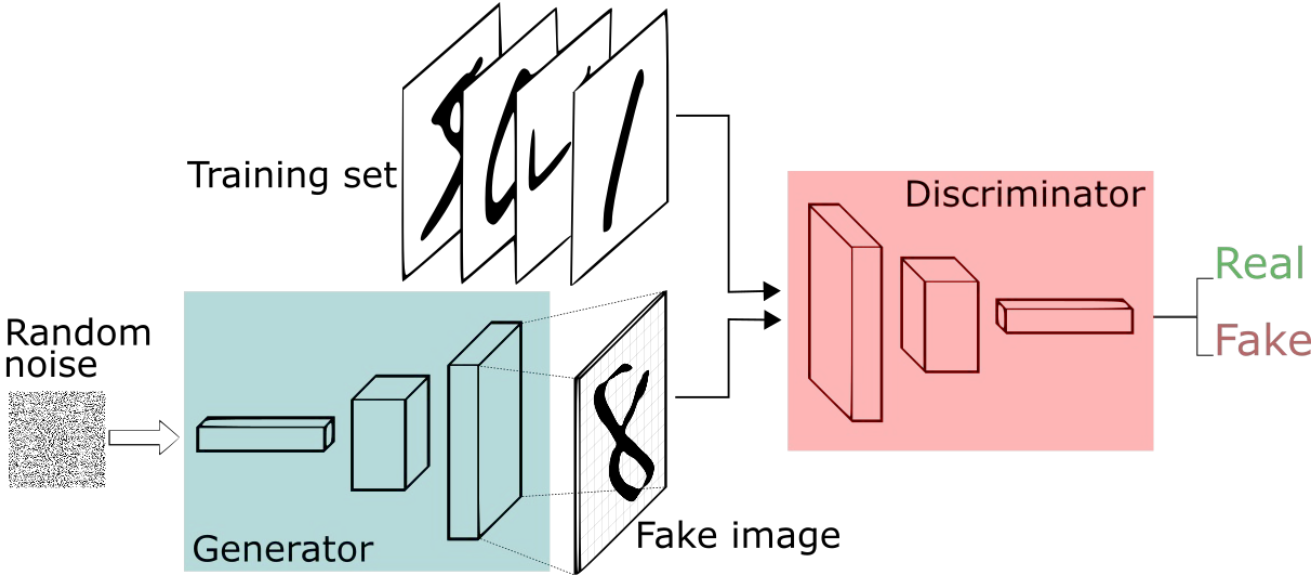
Pivot Method Uses GANs also

- Generator generates a model/classifier
- Discriminator enforces fairness constraints



Generative Adversarial Network (GANs)

- Drawback: hard to tune learning rates of Generator and Discriminator
- Discriminator often learns faster than Generator
- Can easily run into convergence issues



Equalized Coverage

With Malice Towards None: Assessing Uncertainty via Equalized Coverage

Yaniv Romano* Rina Foygel Barber† Chiara Sabatti*‡ Emmanuel J. Candès*§

August, 2019

Abstract

An important factor to guarantee a fair use of data-driven recommendation systems is that we should be able to communicate their uncertainty to decision makers. This can be accomplished by constructing prediction intervals, which provide an intuitive measure of the limits of predictive performance. To support equitable treatment, we force the construction of such intervals to be unbiased in the sense that their coverage must be equal across all protected groups of interest. We present an operational methodology that achieves this goal by offering rigorous distribution-free coverage guarantees holding in finite samples. Our methodology, *equalized coverage*, is flexible as it can be viewed as a wrapper around any predictive algorithm. We test the applicability of the proposed framework on real data, demonstrating that equalized coverage constructs unbiased prediction intervals, unlike competitive methods.

Equalized Coverage

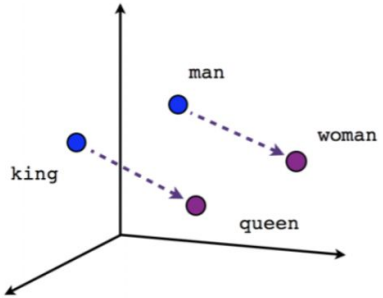
Method	Group	Avg. Coverage	Avg. Length
*Marginal CP	Non-white	0.920	2.907
	White	0.871	2.907
Conditional CP (groupwise)	Non-white	0.903	2.764
	White	0.901	3.182
Conditional CP (joint)	Non-white	0.904	2.738
	White	0.902	3.150
*Marginal CQR	Non-white	0.905	2.530
	White	0.894	3.081
Conditional CQR (groupwise)	Non-white	0.904	2.567
	White	0.900	3.203
Conditional CQR (joint)	Non-white	0.902	2.527
	White	0.901	3.102

Table 1: Length and coverage of both marginal and group-conditional prediction intervals ($\alpha = 0.1$) constructed by conformal prediction (CP) and CQR for MEPS dataset [37]. The results are averaged across 40 random train-test (80%/20%) splits. *Groupwise* – two independent predictive models are used, one for non-white and another for white individuals; *joint* – the same predictive model is used for all individuals. In all cases, the model is formulated as a neural network. The methods marked by an asterisk are not supported by a group-conditional coverage guarantee.

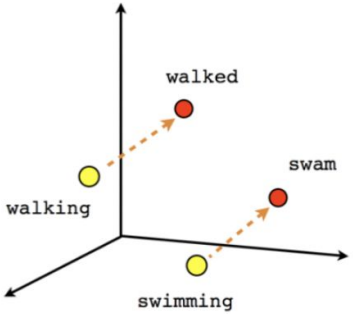
Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, Emmanuel J. Candes 2019

ML Fairness in NLP

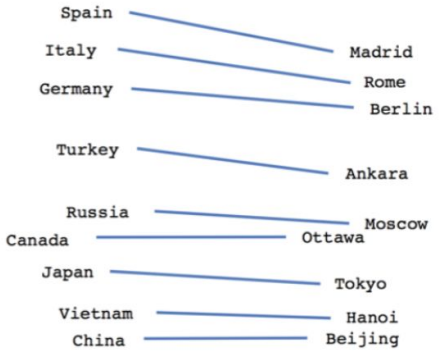
- Word Embeddings: vector/topological representation of a word
- Captures meaning of word - similar words have similar embeddings



Male-Female



Verb tense



Country-Capital

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean 2013

Word Embeddings

- Obtained/learned through models like word2vec, glove, BERT, etc
- Masked language model: given a stream of words, predict next word
- BERT: produces 784-dimensional word embeddings!

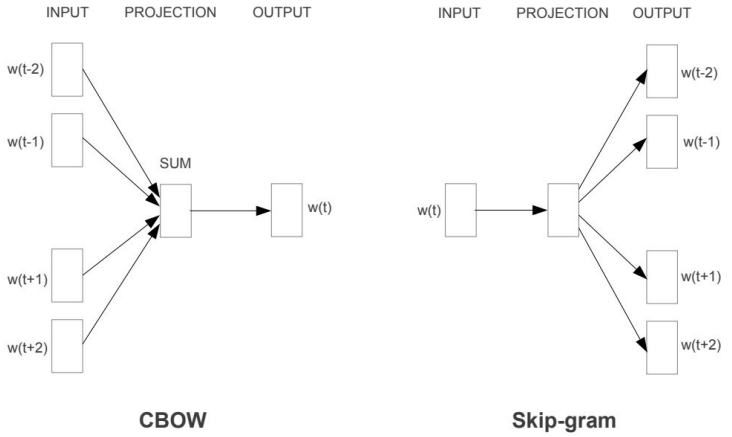


Figure 1: New model architectures. The CBOW architecture predicts the current word based on the context, and the Skip-gram predicts surrounding words given the current word.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean 2013

Word Embeddings

- Problem: language contains bias due to cultural/historical reasons
- Word Embeddings contain same biases

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016

Word Embeddings

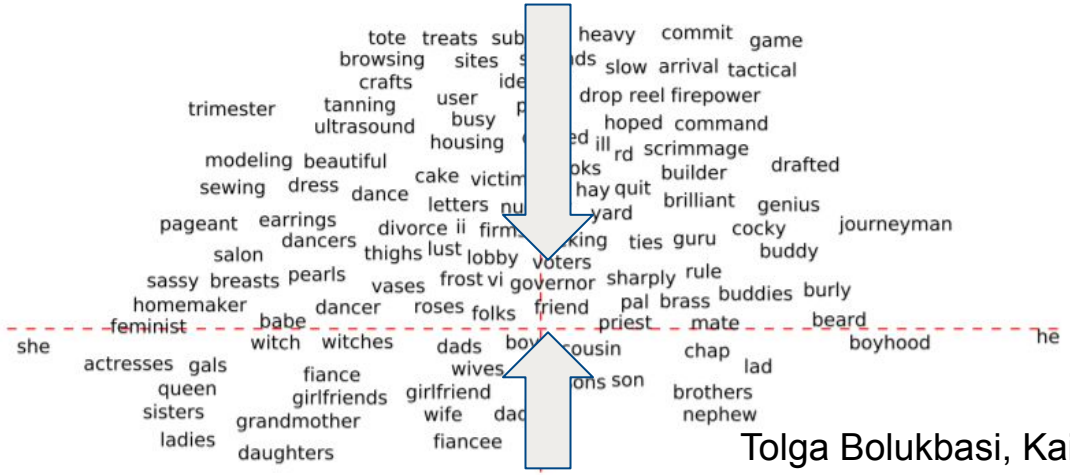
- Problem: language contains bias due to cultural/historical reasons

Association type	**Examples**
Extreme *she* occupations	1.) homemaker 2.) nurse 3.) receptionist 4.) librarian 5.) socialite 6.) hairdresser 7.) nanny 8.) bookkeeper 9.) stylist 10.) housekeeper 11.) interior designer 12.) guidance counselor
Extreme *he* occupations	1.) maestro 2.) skipper 3.) protege 4.) philosopher 5.) captain 6.) architect 7.) financier 8.) warrior 9.) broadcaster 10.) magician 11.) fighter pilot 12.) boss

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016

Debiasing Word Embeddings

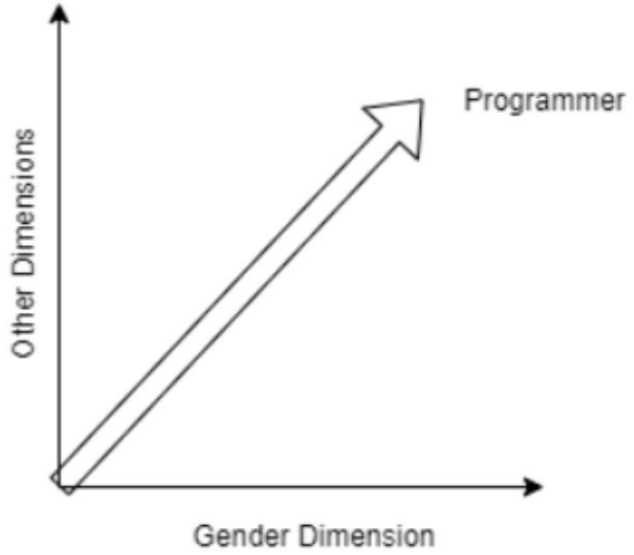
- Solution: debias word embeddings
- Take two opposite words that define a dimension we are interested in (eg. gender, ethnicity)
- “Flatten” embedding space to create bias subspace

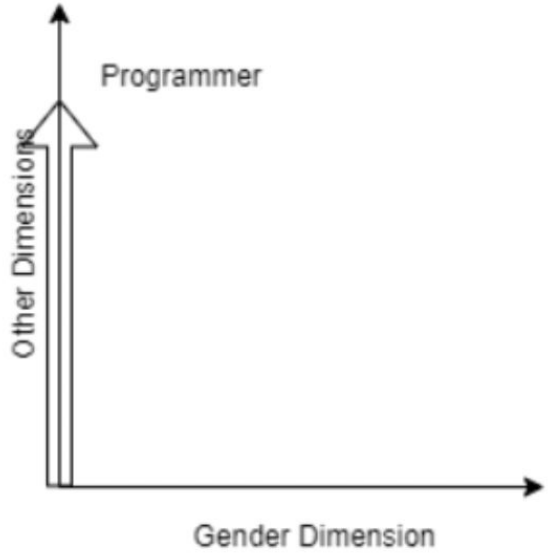


Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016

Debiasing Word Embeddings

- Examine projection of words that point in the bias subspace
- Subtract this component off
- Known as Hard Debiasing





Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016

Debiasing Word Embeddings

- Soft debiasing: debias words while retaining component of original embedding
- Lambda is a chosen hyperparameter

$$\min_{T} ||(TW)^T(TW) - W^TW||_F^2 + \lambda||(TN)^T(TB)||_F^2$$

Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016

Sample Colab Tutorial on Debiasing Word Embeddings

https://colab.research.google.com/github/ResponsiblyAI/word-embedding/blob/master/tutorial-bias-word-embedding.ipynb#scrollTo=KDLn6n64C3vC

Key Takeaways

- ML applications have potentially high-stake consequences
- There are different, sometimes conflicting notions of fairness
- Creating ML systems that have high performance and high fairness is a difficult, but important task given the prevalence of ML

Appendix

Algorithm 1: Algorithm for training a generator

Input: Targeted Population Group (TPG)

repeat

Sample n1 data samples x_i, i= 1, ..., n1 (minibatch sampling)

Sample n2 noise samples z_i, i= 1, ..., n2 (minibatch sampling)

for *K steps* **do**

Update the *Dis* through ascending the stochastic gradient:

$$\nabla_{\theta_{data}}[\frac{1}{n1}\sum_{1}^{n1} log(Dis(x_i|TPG))+ \frac{1}{n2}\sum_{1}^{n2} log(1-Dis(Gen(z_i|TPG)))] \quad (5)$$

end

Update the *Gen* distribution as follows:

$$\tilde{p}_{gen}(x_i|TPG) = p_{gen}(x_i|TPG) - \beta log(2(1-Dis(x_i|TPG))), i = 1, ...n1 \quad (6)$$

where β represents some step size and

$$p_{gen}(x_i|TPG) = \frac{1}{n2}\sum_{j=1}^{n2} k_\sigma(Gen(z_j|TPG) - x_i). \quad (7)$$

Update the Gen through descending the stochastic gradient:

$$\nabla_{\theta_{gen}}[\frac{1}{n2}\frac{1}{n2} log(1-Dis(Gen(z_j|TPG)))+ \frac{1}{n1}\sum_{1}^{n1}(\tilde{p}_{gen}(x_i|TPG) - p_{gen}(x_i|TPG))^2] \quad (8)$$

until ε *elapses*;

Appendix

Note to Screener: The following Criminal History Summary questions require you to add up the total number of specific types of offenses in the person's criminal history. Count an offense type if it was among the charges or counts within an arrest event. Exclude the current case for the following questions.

11. How many times has this person been arrested for a felony property offense that included an element of violence?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
12. How many prior murder/voluntary manslaughter offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
13. How many prior felony assault offense arrests (not murder, sex, or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
14. How many prior misdemeanor assault offense arrests (not sex or domestic violence) as an adult?
☑0 ☐1 ☐2 ☐3+
15. How many prior family violence offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
16. How many prior sex offense arrests (with force) as an adult?
☑0 ☐1 ☐2 ☐3+
17. How many prior weapons offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
18. How many prior drug trafficking/sales offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
19. How many prior drug possession/use offense arrests as an adult?
☑0 ☐1 ☐2 ☐3+
20. How many times has this person been sentenced to jail for 30 days or more?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
21. How many times has this person been sentenced (new commitment) to state or federal prison?
☐0 ☑1 ☐2 ☐3 ☐4 ☐5+
22. How many times has this person been sentenced to probation as an adult?
☑0 ☐1 ☐2 ☐3 ☐4 ☐5+

References

“How We Analyzed the COMPAS Recidivism Algorithm” Jeff Larson, Surya Mattu, Lauren Kirchner and Julia Angwin 2016 https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm

“COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity” William Dieterich, Christina Mendoza, Tim Brennan 2016 http://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

“Fairness in Machine Learning: A Survey” Simon Caton, Christian Haas 2020 https://arxiv.org/abs/2010.04053

“Recovering from Biased Data: Can Fairness Constraints Improve Accuracy?” Avrim Blum, Kevin Stangl 2019 https://arxiv.org/abs/1912.01094

“Generative Adversarial Networks for Mitigating Biases in Machine Learning Systems” Adel Abusitta, Esma Aïmeur, Omar Abdel Wahab 2019 https://arxiv.org/abs/1905.09972

“With Malice Towards None: Assessing Uncertainty via Equalized Coverage” Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, Emmanuel J. Candes 2019 https://arxiv.org/abs/1908.05428

“Distributed Representations of Words and Phrases and their Compositionality” Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, Jeffrey Dean 2013 https://arxiv.org/abs/1310.4546

References

“Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings” Tolga Bolukbasi, Kai-Wei Chang, James Zou, Venkatesh Saligrama, Adam Kalai 2016 https://arxiv.org/abs/1607.06520

Thank you!

Questions?