Introduction to Machine Learning

Fairness in Machine Learning

Varun Chandola

Computer Science & Engineering
State University of New York at Buffalo
Buffalo, NY, USA
chandola@buffalo.edu

University at Buffalo
Department of Computer Science and Engineering
School of Engeering and Applied Sciences

Outline

Introduction

- Main text - `https://fairmlbook.org` [1]
  - Solon Barocas, Moritz Hardt, Arvind Narayanan
- Other recommended resources:
  - Fairness in machine learning (NeurIPS 2017)
  - 21 fairness definitions and their politics (FAT* 2018)
  - Machine Bias - COMPAS Study
- Must read - The **Machine Learning Fairness Primer** by Dakota Handzlik
- Programming Assignment 3 and Gradiance Quiz #10
- Also see - The Mozilla Responsible Computer Science Challenge

What will we learn in the module?

- What principles should guide the design of a machine learning solution?
  - Besides the usual performance metrics (accuracy, efficiency, etc.)

Ethical Considerations

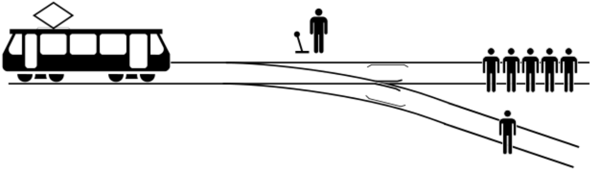
- What ethical principles to abide by?

Fairness and Bias

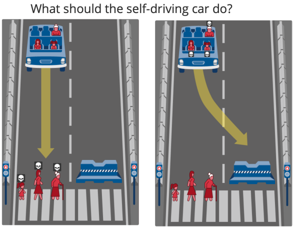
- Why is fairness important?
- How does bias get introduced?
- How do we measure fairness?
- How to make algorithms fair and remove bias?

Ethical Principles in ML

- What are the ethical implications of an ML Application?
- Ethics - The right thing to do
- The Trolley Problem



- Designing a self-driving car?
- *Moral machine*
  - `https://www.moralmachine.net`



Two Ethical Frameworks

Utilitarianism

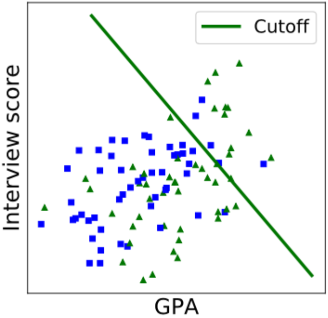
- Decisions made based on the amount of overall happiness or benefit they provide
  - Greater good in greater numbers
- Not the universal human approach to decision making
- Makes uncertain decisions

Deontological

- Decisions made based on a notion of moral duty or obligation
- What if the definition of moral duty is flawed?
- Makes certain decisions

Fariness - Toy Example

- *Task*: Learn a ML based job hiring algorithm
- *Inputs*: GPA, Interview Score
- *Target*: Average performance review
- *Sensitive attribute*: Binary (denoted by $\square$ and Δ), represents some demographic group
  - We note that GPA is correlated with the sensitive attribute



Process

1. Regression model to predict target
2. Apply a threshold (denoted by green line) to select candidates

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?
- It depends on the definition of fairness

Toy Example

- ML models does not use sensitive attribute
- Does it mean it is fair?
- It depends on the definition of fairness

Fairness-as-blindness notion

- Two individuals with similar features get similar treatment
- This model is fair

What about a different definition of fairness?

- Are candidates from the two groups equally likely to be hired?

What about a different definition of fairness?

- Are candidates from the two groups equally likely to be hired?
- No - triangles are more likely to be hired than squares
- Why did the model become unfair because of this definition?
  - In the training data, average performance review is lower for squares than triangles

Why this disparity in the data?

- Many factors could have led to this:
  - Managers who score employee's performance might have a bias
  - Workplace might be biased against one group
  - Socio-economic background of one group might have resulted in poor educational outcomes
  - Some intrinsic reason
  - Combination of these factors
- Let us assume that this disparity that was learnt by the ML model is unjustified
- How do we get rid of this?

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model
- Option 2: pick different thresholds for each sub-group
  - Model is no longer "blind"

Making ML model bias-free

- Option 1: ignore GPA as a feature
  - Might result in poor accuracy of the model
- Option 2: pick different thresholds for each sub-group
  - Model is no longer "blind"
- Option 3: add a diversity reward to the objective function
  - Could still result in poor accuracy

Why fairness?

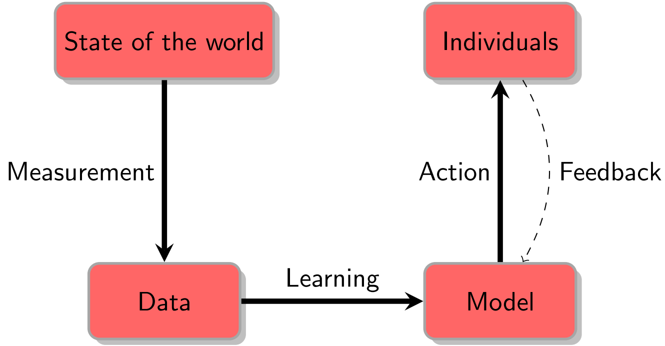
- We want/expect everything to be fair and bias-free
- Machine learning driven systems are everywhere
- Obviously we want them to be fair as well
  - Closely related are issues of ethics, trust, and accountability

What does fairness mean?

- **Consequential decision making**: ML system makes a decision that impacts individuals
  - admissions, job offers, bail granting, loan approvals
- Should use factors that are *relevant* to the outcome of interest

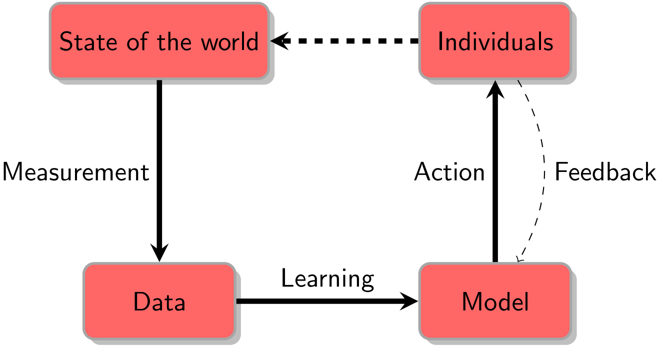
How does an ML algorithm becomes unfair?

- The "ML for People" Pipeline



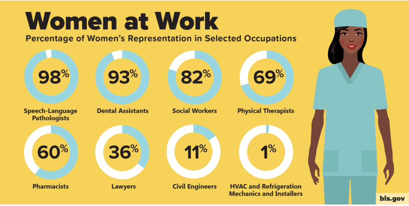
How does an ML algorithm becomes unfair?

- The "ML for People" Pipeline



Issues with the state of the society

- Most ML applications are about people
  - Even a pothole identification algorithm
- Demographic disparities exist in society
- These get embedded into the training data
- As ML practitioners we are not focused on removing these disparities
- We do not want ML to reinforce these disparities
- The dreaded **feedback loops** [3]



Understanding Bias in Data

- A data sample is considered **biased**, if it does not correctly represent the population parameter being estimated.
- There are several types of statistical and cognitive biases present in data acquisition and processing.

1. Selection bias
2. Base rate fallacy (or bias or neglect)
3. Conjunction fallacy
4. Response bias
5. Confirmation bias
6. Detection bias
7. Availability bias
8. Social biases
9. Measurement bias

- For exact definitions, refer to the *fairness primer*.

Selection Biases

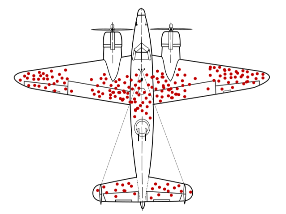
- Data instance are selected for analysis in a non-random way.

Sampling Bias

- Obtaining data in a non-random way
- Example - using opinions from Twitter to infer interest of population on a particular issue.

Survivorship Bias

- Bias due to applying critical thresholds to choose data for analysis



Base Rate Fallacy/Neglect/Bias

- Similar to the concept of ignoring the prior distribution in Bayesian analysis

How to make the ML model more fair

- Better objective functions that are fair to all sub-groups
  - More about this next

Fairness in Classification Problems

Notation

- Predict Y given $\mathbf{X}$
- Y is our target class $Y \in \{0, 1\}$
- $\mathbf{X}$ represents the input feature vector

Example

- Y - Will an applicant pay the loan back?
- $\mathbf{X}$ - Applicant characteristics - credit history, income, etc.

Supervised Learning

- Given training data: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$
- Either learn a function f, such that:

$$y^* = f(\mathbf{x}^*)$$

- Or, assume that the data was drawn from a probability distribution
- In either case, we can consider the classification output as a random variable $\hat{Y}$
- Now we have three random variables:

$$\mathbf{X}, Y, \hat{Y}$$

- We are going to ignore how we get $\hat{Y}$ from $\mathbf{X}$ for these discussions

How do we measure the quality of a classifier?

- So far we have been looking at accuracy

A different way to look at accuracy

$$\text{Accuracy} \equiv P(Y = \hat{Y})$$

- Probability of the predicted label to be equal to the true label
- How do we calculate this?

Accuracy is not everyting!

- Consider a test data set with 90 examples with true class 1 and 10 examples with true class 0
- A *degenerate* classifier that classifies everything as label 1, would still have a 90% accuracy on this data set

Other evaluation criteria

Event	Condition	Metric
$\hat{Y} = 1$	$Y = 1$	True positive rate (recall on positive class)
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate (recall on negative class)

- Here we are treating class label 1 as the positive class and class label 0 as the negative class.

We can swap the condition and the event

Event	**Condition**	**Metric**
$Y = 1$	$\hat{Y} = 1$	precision (on positive class)
$Y = 0$	$\hat{Y} = 0$	precision (on negative class)

Score Functions

- Often classification involves computing a **score** and then applying a threshold
- E.g., Logistic regression: first calculate $P(Y = 1|\mathbf{X} = \mathbf{x})$, then apply a threshold of 0.5
- Or, Support Vector Machine: first calculate $\mathbf{w}^\top \mathbf{x}$ and then apply a threshold of 0

Conditional Expectation

$$r(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

- We can treat it as a random variable too $R = \mathbb{E}[Y|\mathbf{X}]$
- This is what logistic regression uses.

From scores to classification

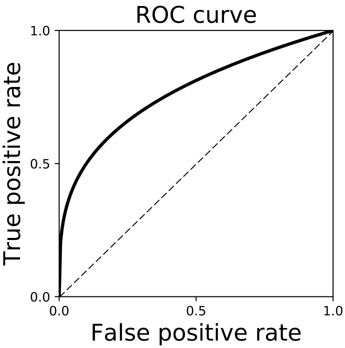
- Use a threshold t

$$y = \begin{cases} 1 & \text{if } r(\mathbf{x}) \geq t, \\ 0 & \text{otherwise} \end{cases}$$

- What threshold to choose?
  - If t is high, only few examples with very high score will be classified as 1 (accepted)
  - If t is low, only few examples with very low score will be classified as 0 (rejected)

The *Reciever Operating Characteristic* (ROC) Curve

- Exploring the entire range of t
- Each point on the plot is the FPR and TPR for a given value of t
- Area under the ROC curve or AUC is a quantitative metric derived from ROC curve

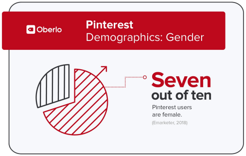


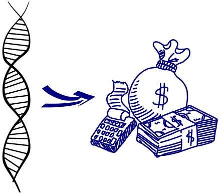
Sensitive Attributes

- Let A denote the attribute representing the sensitive characteristic of an individual
- There could be more than one sensitive attributes

Things to remember

- It is not always easy to identify A and differentiate it from $\mathbf{X}$
- Removing the sensitive attribute from $\mathbf{X}$ does not guarantee fairness
- Removing the sensitive attribute could make the classifier less accurate
- Not always a good idea to remove the impact of sensitive attributes





Quantifying Fairness

- Let us define some reasonable ways of measuring fairness
  - There are several ways to do this
  - All are debatable
- Three different categories

Independence	**Separation**	**Sufficiency**
$\hat{Y} \perp\!\!\!\perp A$	$\hat{Y} \perp\!\!\!\perp A \vert Y$	$Y \perp\!\!\!\perp A \vert \hat{Y}$

- Y - True label; $\hat{Y}$ - Predicted label; A - Sensitive attribute;

Conditional Independence

$$A \perp\!\!\!\perp B|C \Leftarrow P(A,B|C) = P(A|C)P(B|C)$$

- Amount of Speeding fine $\perp\!\!\!\perp$ Type of Car | Speed

Independence

$$P(\hat{Y} = 1 | A = a) = P(\hat{Y} = 1 | A = b), \forall a, b \in A$$

- Referred to as *demographic parity*, *statistical parity*, *group fairness*, *disparate impact*, etc.
- Probability of an individual to be assigned a class is equal for each group

Disparate Impact Law

$$\frac{P(\hat{Y} = 1 | A = a)}{P(\hat{Y} = 1 | A = b)} \geq 1 - \epsilon$$

For $\epsilon = 0.2$ - *80 percent rule*

Issues with independence measures

- *The self fulfilling prophecy* [2]
- Consider the hiring scenario where the model picks p excellent candidates from group a and p poor quality candidates from group b
  - Meets the independence criteria
  - However, it is still unfair

How to satisfy fairness criteria?

1. **Pre-processing phase**: Adjust the feature space to be uncorrelated with the sensitive attribute.
2. **Training phase**: Build the constraint into the optimization process for the classifier.
3. **Post-processing phase**: Adjust a learned classifier so that it is uncorrelated to the sensitive attribute

Separation

$$\hat{Y} \perp\!\!\!\perp A | Y$$

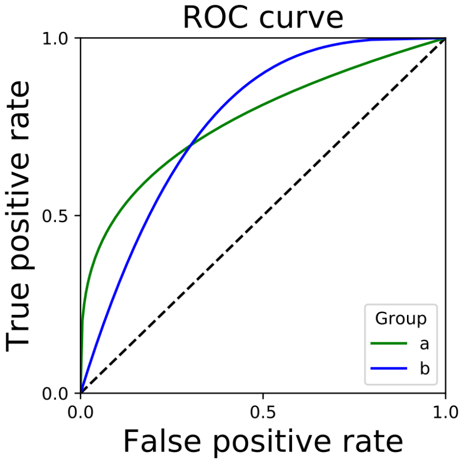
- Alternatively, the true positive rate and the false positive rate is equal for any pair of groups:

$$\begin{aligned} P(\hat{Y} = 1 | Y = 1, A = a) &= P(\hat{Y} = 1 | Y = 1, A = b) \\ P(\hat{Y} = 1 | Y = 0, A = a) &= P(\hat{Y} = 1 | Y = 0, A = b) \\ \forall a, b \in A \end{aligned}$$

- Can handle the discrepancy with the independence metric mentioned earlier

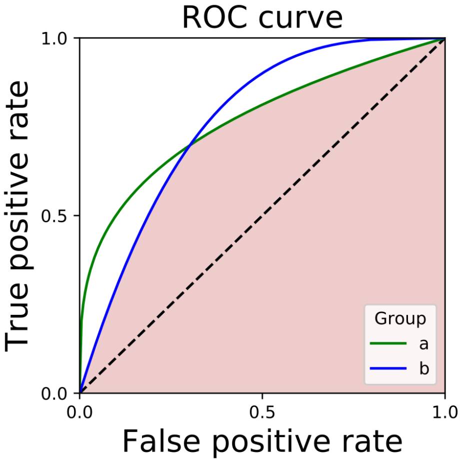
How to achieve separation

- Apply post-processing step using the ROC Curve
- Plot ROC curve for each group
- Within the constraint region (overlap), pick a classifier that minimizes the given cost



How to achieve separation

- Apply post-processing step using the ROC Curve
- Plot ROC curve for each group
- Within the constraint region (overlap), pick a classifier that minimizes the given cost



Sufficiency

$$Y \perp\!\!\!\perp A | R$$

- Alternatively, the true positive rate and the false positive rate is equal for any pair of groups:

$$P(Y = 1 | R = r, A = a) \quad = \quad P(Y = 1 | R = r, A = b)$$
$$\forall r \in dom(R) \text{ and } a, b \in A$$

Achieving sufficiency by calibration

What is calibration?

- Let us revert back to the score R
  - Recall that $\hat{Y}$ was obtained by applying a threshold on R
- R is *calibrated*, if for all r in the domain of R:

$$P(Y = 1 | R = r) = r$$

- Of course, this means that R should be between 0 and 1
- *Platt Scaling*: Converts an uncalibrated score to a calibrated score [4]

- Calibration by group implies sufficiency
  - Apply Platt scaling to each group defined by the sensitive attribute

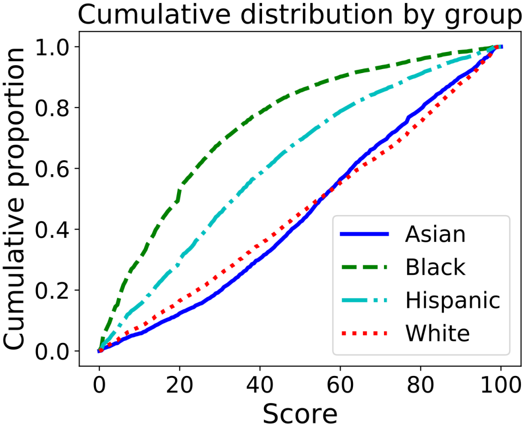
Case Study: Credit Scoring

- Extend loan or not - based on the risk that a loan applicant will default on a loan
- Data from the *Federal Reserve*
  - A - Demographic information (race)
  - R - Credit score
  - Y - Default or not (defined by credit bureau)

Table: Credit score distribution by race

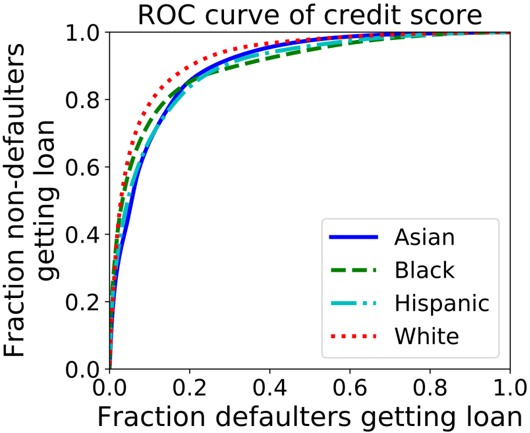
Race or ethnicity	Samples with both score and outcome
White	133,165
Black	18,274
Hispanic	14,702
Asian	7,906
Total	174,047

Group-wise distribution of credit score



- Strongly depends on the group

Using credit score for classification



- How make the classifier fair?

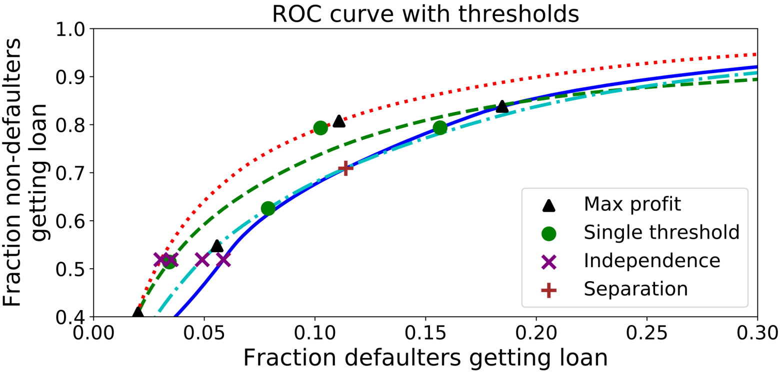
Four Strategies

1. *Maximum profit*: Pick group-dependent score thresholds in a way that maximizes profit
2. *Single threshold*: Pick a single uniform score threshold for all groups in a way that maximizes profit
3. *Separation*: Achieve an equal true/false positive rate in all groups. Subject to this constraint, maximize profit.
4. *Independence*: Achieve an equal acceptance rate in all groups. Subject to this constraint, maximize profit.

What is the profit?

- Need to assume a reward for a true positive classification and a cost/penalty for a false positive classification
- We will assume that cost of a false positive is 6 times greater than the reward for a true positive.

Comparing different criteria



References I

S. Barocas, M. Hardt, and A. Narayanan.
Fairness and Machine Learning.
fairmlbook.org, 2019.
http://www.fairmlbook.org.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel.
Fairness through awareness.
In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA, 2012. Association for Computing Machinery.

D. Ensign, S. A. Friedler, S. Neville, C. Scheidegger, and S. Venkatasubramanian.
Runaway feedback loops in predictive policing.
In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, volume 81 of *Proceedings of Machine Learning Research*, pages 160–171. PMLR, 2018.

References II

J. Platt.
Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods.
Adv. Large Margin Classif., 10, 06 2000.