CSC 411 Lecture 20: Algorithmic Fairness

Roger Grosse, Amir-massoud Farahmand, and Juan Carrasquilla

University of Toronto

Final Exam

- Tuesday, Dec. 11, from 7–10pm. See course web page for room assignments.
- Covers all lectures except the final week (Lectures 23 and 24)
- Similar in format and difficulty to midterm (except that Questions 8 and 9 on the midterm were too hard)
- You are only responsible for material covered in lecture, but topics additionally covered in tutorials and homeworks will receive more emphasis.
- See e-mail announcement for what you need to know about Gaussians.
- Practice exams will be posted.

Overview

- Most of this course has been concerned with getting ML algorithms to do something useful (e.g. make good predictions, find patterns, learn policies).
- As ML starts to be applied to critical applications involving humans, the field is wrestling with the societal impacts
  - **Security:** what if an attacker tries to poison the training data, fool the system with malicious inputs, "steal" the model, etc.?
  - **Privacy:** avoid leaking (much) information about the data the system was trained on (e.g. medical diagnosis)
  - **Fairness:** ensure that the system doesn't somehow disadvantage particular individuals or groups
  - **Transparency:** be able to understand why one decision was made rather than another
  - **Accountability:** an outside auditor should be able to verify that the system is functioning as intended
- If some of these definitions sound vague, that's because formalizing them is half the challenge!

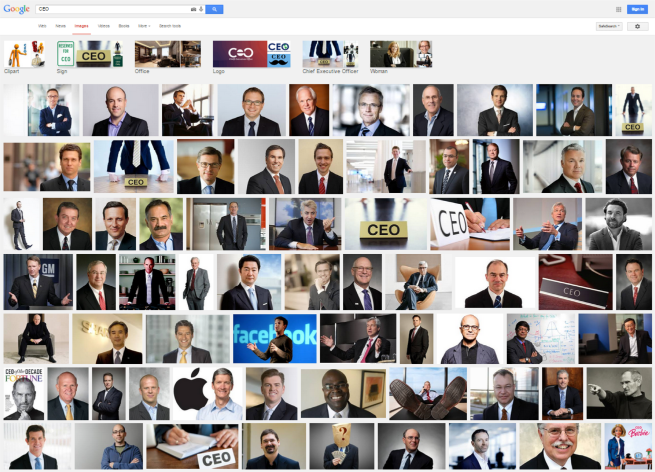
WHY WAS I NOT SHOWN THIS AD?

Credit: Richard Zemel

FAIRNESS IN AUTOMATED DECISIONS

Credit: Richard Zemel

SUBTLER BIAS



Credit: Richard Zemel

Overview: Fairness

- This lecture: algorithmic fairness
- Goal: identify and mitigate bias in ML-based decision making, in all aspects of the pipeline
- Sources of bias/discrimination
  - Data
    - Imbalanced/impoverished data
    - Labeled data imbalance (more data on white recidivism outcomes)
    - Labeled data incorrect / noisy (historical bias)
  - Model
    - ML prediction error imbalanced
    - Compound injustices (Hellman)

Credit: Richard Zemel

Overview: Fairness

- Notation
  - X: input to classifier
  - S: sensitive feature (age, gender, race, etc.)
  - Z: latent representation
  - Y: prediction
  - T: true label
- We use capital letters to emphasize that these are random variables.

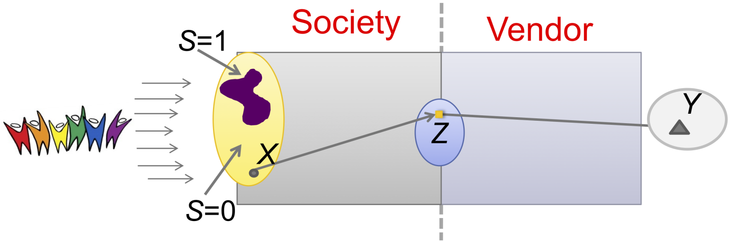
Fairness Criteria

- Most common way to define fair classification is to require some invariance with respect to the sensitive attribute
  - Demographic parity: $Y \perp\!\!\!\perp S$
  - Equalized odds: $Y \perp\!\!\!\perp S \mid T$
  - Equal opportunity: $Y \perp\!\!\!\perp S \mid T = t$, for some t
  - Equal (weak) calibration: $T \perp\!\!\!\perp S \mid Y$
  - Equal (strong) calibration: $T \perp\!\!\!\perp S \mid Y$ and $Y = \Pr(T = 1)$
  - Fair subgroup accuracy: $\mathbb{1}[T = Y] \perp\!\!\!\perp S$
- $\perp\!\!\!\perp$ denotes stochastic independence
- Many of these definitions are incompatible!

Credit: Richard Zemel

Learning Fair Representations

- Idea: separate the responsibilities of the (trusted) society and (untrusted) vendor



- Goal: find a representation Z that removes any information about the sensitive attribute
- Then the vendor can do whatever they want!

Image Credit: Richard Zemel

Learning Fair Representations

- A naïve attempt: simply don't use the sensitive feature.
  - Problem: the algorithm implicitly learn to predict the sensitive feature from other features (e.g. race from zip code)
- Another idea: limit the algorithm to a small set of features you're pretty sure are safe and task-relevant
  - This is the conservative approach, and commonly used for both human and machine decision making
  - But removing features hurts the classification accuracy. Maybe we can make more accurate decisions if we include more features and somehow enforce fairness algorithmically?
- Can we learn fair representations, which can make accurate classifications without implicitly using the sensitive attribute?

Learning Fair Representations

Desiderata for the representation:

Retain information about X	$\Rightarrow$	high mutual information between X and Z
Obfuscate S	$\Rightarrow$	low mutual information between S and Z
Allow high classification accuracy	$\Rightarrow$	high mutual information between T and Z

Learning Fair Representations

First approach: Zemel et al., 2013, "Learning fair representations"

- Let Z be a discrete representation (like K-means)
- Determine Z stochastically based on distance to a prototype for the cluster (like the cluster center in K-means)

$$\Pr(Z = k \mid \mathbf{x}) \propto \exp(-d(\mathbf{x}, \mathbf{v}_k)),$$

  where d is some distance function (e.g. Euclidean distance)
- Use the Bayes classifier $y = \Pr(T = 1 \mid Z)$
- Need to fit the prototypes $\mathbf{v}_k$

Learning Fair Representations

- Retain information about X: penalize reconstruction error

$$\mathcal{L}_{\mathrm{reconst}} = \frac{1}{N} \sum_{i=1}^{N} \|\mathbf{x}^{(i)} - \tilde{\mathbf{x}}^{(i)}\|^2$$

- Predict accurately: cross-entropy loss

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{N} \sum_{i=1}^{N} -t^{(i)} \log y^{(i)} - (1 - t^{(i)}) \log(1 - y^{(i)})$$

- Obfuscate S:

$$\mathcal{L}_{\mathrm{discrim}} = \frac{1}{K} \sum_{k=1}^{K} \left| \frac{1}{N_0} \sum_{i:s^{(i)}=0} \Pr(Z = k \mid \mathbf{x}^{(i)}) - \frac{1}{N_1} \sum_{i:s^{(i)}=1} \Pr(Z = k \mid \mathbf{x}^{(i)}) \right|,$$

where we assume for simplicity $S \in \{0, 1\}$ and N_0 is the count for $s = 0$.

Learning Fair Representations

- Obfuscate S:

$$\mathcal{L}_{\text{discrim}} = \frac{1}{K}\sum_{k=1}^{K}\left|\frac{1}{N_0}\sum_{i:s^{(i)}=0}\Pr(Z=k\,|\,\mathbf{x}^{(i)}) - \frac{1}{N_1}\sum_{i:s^{(i)}=1}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right|,$$

- Is this about individual-level or group-level fairness?
- If discrimination loss is 0, we satisfy demographic parity

$$\begin{aligned}
\Pr(Y=1\,|\,s^{(i)}=1) &= \frac{1}{N_1}\sum_{i:s^{(i)}=1}\sum_{k=1}^{K}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\Pr(Y=1\,|\,Z=k) \\
&= \sum_{k=1}^{K}\left[\frac{1}{N_1}\sum_{i:s^{(i)}=1}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right]\Pr(Y=1\,|\,Z=k) \\
&= \sum_{k=1}^{K}\left[\frac{1}{N_0}\sum_{i:s^{(i)}=0}\Pr(Z=k\,|\,\mathbf{x}^{(i)})\right]\Pr(Y=1\,|\,Z=k) \\
&= \Pr(Y=1\,|\,s^{(i)}=0)
\end{aligned}$$

Learning Fair Representations

Datasets

1. German Credit

Task: **classify individual as good or bad credit risk**

Sensitive feature: **Age**

2. Adult Income

Size: **45,222 instances, 14 attributes**

Task: **predict whether or not annual income > 50K**

Sensitive feature: **Gender**

3. Heritage Health

Size: **147,473 instances, 139 attributes**

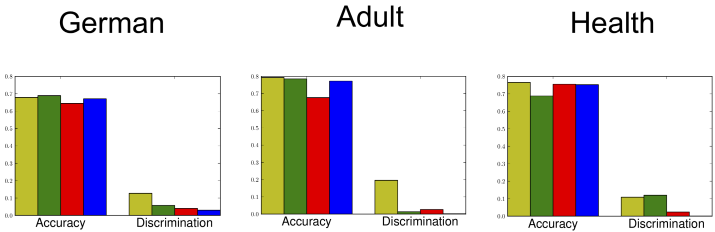
Task: **predict whether patient spends any nights in hospital**

Sensitive feature: **Age**

Metrics

- Classification accuracy
- Discrimination

$$\left| \frac{\sum_{i:s^{(i)}=1}^{N} y^{(i)}}{N_1} - \frac{\sum_{i:s^{(i)}=0}^{N} y^{(i)}}{N_0} \right|$$



Yellow = unrestricted; Blue = theirs

Fair VAE

- Discrete Z based on prototypes is very limiting. Can we learn a more flexible representation?
- Louizos et al., 2015, "The variational fair autoencoder"
- The variational autoencoder (VAE) is a kind of autoencoder that represents a probabilistic model, and can be trained with a variational objective similar to the one we used for E-M.
  - For this lecture, just think of it as an autoencoder.
  - How can we learn an autoencoder such that the code vector $\mathbf{z}$ loses information about s?

Fair VAE: Maximum Mean Discrepancy

- Our previous non-discrimination criterion only makes sense for discrete Z.
- New criterion: ensure that $p(Z \mid s)$ is indistinguishable for different values of s.
- Maximum mean discrepancy (MMD) is a quantitative measure of distance between two distributions. Pick a feature map $\boldsymbol{\psi}$.

$$\mathrm{MMD}(p; q) = \left\| \mathbb{E}_{\mathbf{z} \sim p}[\boldsymbol{\psi}(\mathbf{z})] - \mathbb{E}_{\mathbf{z} \sim q}[\boldsymbol{\psi}(\mathbf{z})] \right\|^2$$

Fair VAE: Maximum Mean Discrepancy

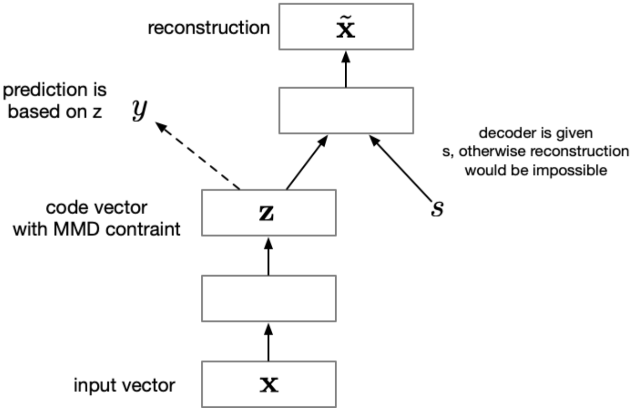
- MMD can be kernelized by expressing it in terms of $k(\mathbf{z}, \mathbf{z}') = \psi(\mathbf{z})^\top \psi(\mathbf{z}')$.
- Let $\{\mathbf{z}_i\}_{i=1}^{N_0}$ and $\{\mathbf{z}'_i\}_{i=1}^{N_1}$ be sets of samples from p and q. The empirical MMD is given by:

$$\left\| \frac{1}{N_0} \sum_{i=1}^{N_0} \psi(\mathbf{z}_i) - \frac{1}{N_1} \sum_{i=1}^{N_1} \psi(\mathbf{z}'_i) \right\|^2$$
$$= \frac{1}{N_0^2} \sum_{i=1}^{N_0} \sum_{j=1}^{N_0} k(\mathbf{z}_i, \mathbf{z}_j) + \frac{1}{N_1^2} \sum_{i=1}^{N_1} \sum_{j=1}^{N_1} k(\mathbf{z}'_i, \mathbf{z}'_j) - 2 \frac{1}{N_0 N_1} \sum_{i=1}^{N_0} \sum_{j=1}^{N_1} k(\mathbf{z}_i, \mathbf{z}'_j)$$

- You can show that for certain kernels (e.g. RBF), the MMD is 0 iff $p = q$. So MMD is a very powerful distance metric.

Fair VAE

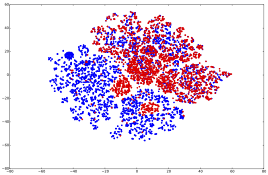
Train a VAE, with the constraint that the MMD between $p(\mathbf{z} \mid s = 0)$ and $p(\mathbf{z} \mid s = 1)$ is small.



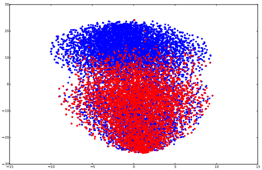
Fair VAE: tSNE embeddings

- tSNE is an unsupervised learning algorithm for visualizing high-dimensional datasets. It tries to embed points in low dimensions in a way that preserves distances as accurately as possible.
- Here are tSNE embeddings of different distributions, color-coded by the sensitive feature:

Original inputs



VAE latent space



Fair VAE latent space

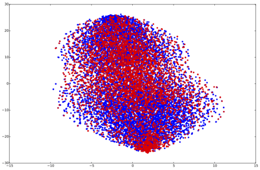


Figure Credit: Louizos et al., 2015

Individual Fairness

- The work on fair representations was geared towards group fairness
- Another notion of fairness is individual level: ensuring that similar individuals are treated similarly by the algorithm
  - This depends heavily on the notion of "similar".
- One way to define similarity is in terms of the "true label" T (e.g. whether this individual is in fact likely to repay their loan)
  - Can you think of a problem with this definition?
  - The label may itself be biased
    - if based on human judgments
    - if, e.g., societal biases make it harder for one group to pay off their loans
  - We'll ignore this issue in our analysis. But keep in mind that you'd need to carefully consider the assumptions when applying one of these methods!

Equal Opportunity

- Now we'll turn to Hardt et al., 2016, "Equality of opportunity in supervised learning".
- Assume we make a binary prediction by computing a real-valued score $R = f(X, S)$, and then thresholding this score to obtain the prediction Y.
- As before, assume $S \in \{0, 1\}$.
- Motivating example: predict whether an individual is likely to repay their loan
- Two notions of individual fairness:
  - Equalized odds: equal false positive and false negative rates

$$\Pr(Y = 1 \mid S = 0, T = t) = \Pr(Y = 1 \mid S = 1, T = t) \quad \text{for } t \in \{0, 1\}$$

  - Equal opportunity: equal false negative rates

$$\Pr(Y = 1 \mid S = 0, T = 1) = \Pr(Y = 1 \mid S = 1, T = 1)$$

Equal Opportunity

- Consider derived predictors, which are a function of the real-valued score R and the sensitive feature S.
  - I.e., we don't need to check the original input X. This simplifies the analysis.
- Define a loss function $\mathcal{L}(Y, T)$. Since Y and T are binary, there are 4 values to specify.
- They show that:
  - Without a constraint, the optimal predictor is obtained from thresholding R.
  - With an equal opportunity constraints, the optimal predictor is obtained by thresholding R, but with a different treshold for different values of S.
  - Satisfying equalized odds is overconstrained, and may require randomizing Y.

- Case study: FICO scores
- Aim to predict whether an individual has less than an 18% rate of default (which is the treshold for profitability)

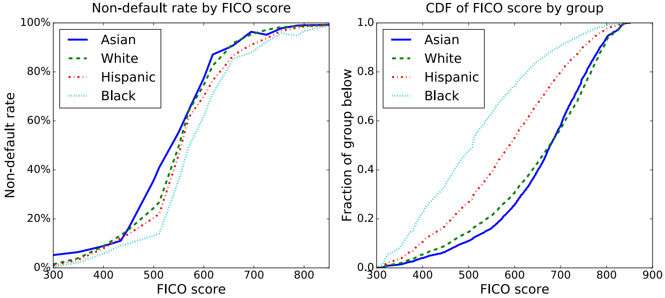


Figure: Hardt et al., 2016

Equal Opportunity

- The "race-blind" solution applies the same threshold for all the groups.
- Problem: non-defaulting black applicants are much less likely to be approved than non-defaulting white applicants.
  - Fraction of non-defaulting applicants in each group = fraction of area under curve which is shaded

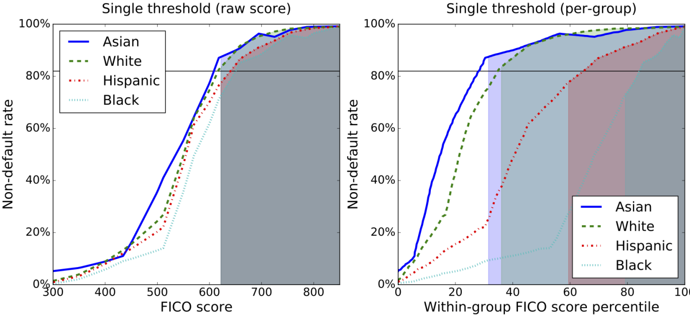


Figure: Hardt et al., 2016

Equal Opportunity

- Can obtain equal opportunity, equalized odds, demographic parity by setting group-specific thresholds (except equalized odds requires randomizing).

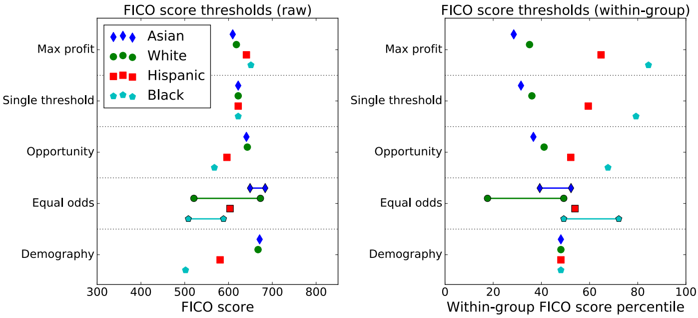


Figure: Hardt et al., 2016

Equal Opportunity

- Different notions of fairness often come into conflict. E.g., demographic parity conflicts with equal opportunity (left).
- Some notions of fairness are harder to achieve than others, in terms of lost profit (right).
- Choosing the right criterion requires careful consideration of the causal relationships between the variables.

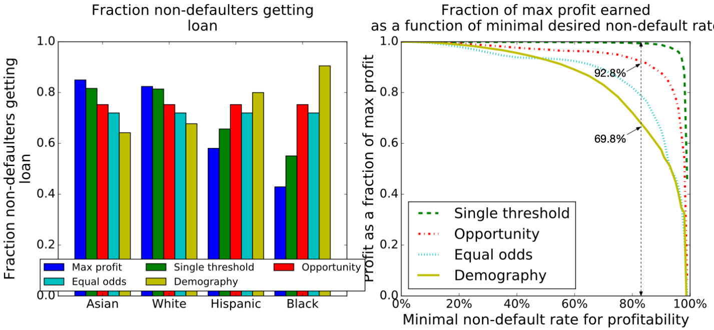


Figure: Hardt et al., 2016

Summary

- Fairness is a challenging issue to address
  - Not something you can just measure on a validation set
  - Philosophers and lawyers have been trying to define it for thousands of years
  - Different notions are incompatible. Need to carefully consider the particular problem.
    - individual vs. group
- Explosion of interest in ML over the last few years
- New conference on Fairness, Accountability, and Transparency (FAT*)
- New textbook: `https://fairmlbook.org/`