CSC 2515: Introduction to Machine Learning

Tutorial - Algorithmic Fairness

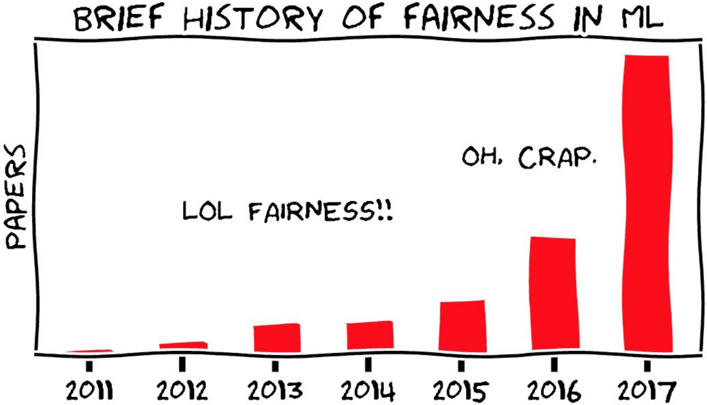
(Based on the slides of previous years)

University of Toronto

Overview

- 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!

Overview: Fairness

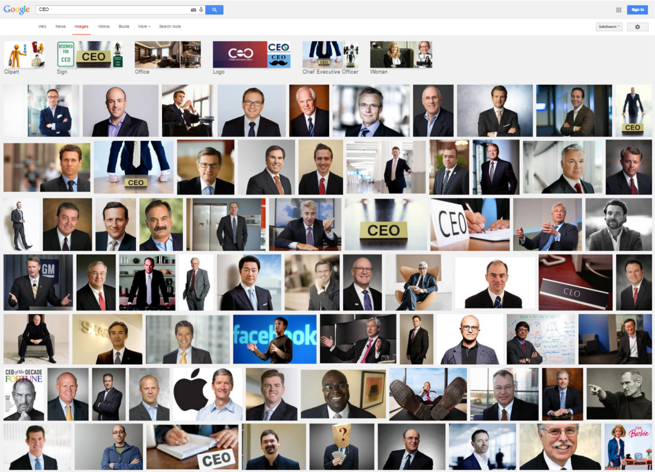


Credit: Moritz Hardt

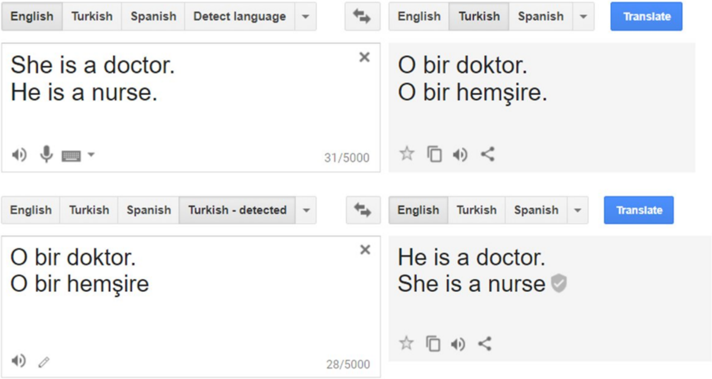
Overview: Fairness

FAIRNESS IN AUTOMATED DECISIONS

SUBTLER BIAS



Overview: Fairness



Turkish has gender neutral pronouns

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
    - Labeled data incorrect / noisy
  - Model
    - ML prediction error imbalanced
    - Compound injustices
- Important: Algorithmic fairness does not imply real fairness!

Learning Fair Representations

- A naïve attempt: simply don't use the sensitive feature.
  - Problem: the algorithm implicitly learns 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?

Overview: Fairness

- Notation
  - $X \in \mathbb{R}^D$: input to classifier
  - $S \in \{0, 1\}$: belongs to protected group (age, gender, race, etc.)
  - $Z \in \{1, 2, ..., K\}$: latent representation (code)
  - $T \in \{0, 1\}$: true label
  - $Y \in [0, 1]$: prediction ($p(T = 1 \mid X)$)

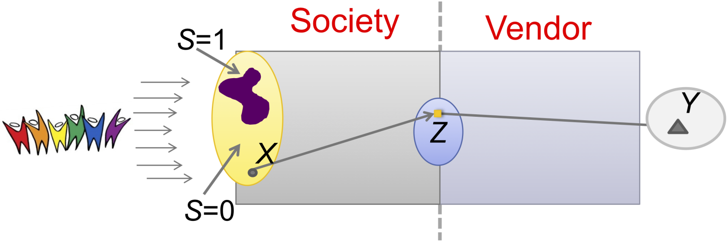
- We use capital letters to emphasize that these are random variables.

Fairness Criteria

- $X \perp\!\!\!\perp Y$ means X and Y are independent
- 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: $Y \perp\!\!\!\perp S \,|\, T$
  - Equal opportunity: $Y \perp\!\!\!\perp S \,|\, T = t$, for a fixed t
  - Equal (weak) calibration: $T \perp\!\!\!\perp S \,|\, Y$
  - Equal (strong) calibration: $T \perp\!\!\!\perp S \,|\, Y$ and $Y = \Pr(T = 1)$
  - Fair subgroup accuracy: $\mathbb{1}[T = Y] \perp\!\!\!\perp S$
- Many of these definitions are incompatible!

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!

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 code or representation (like K-means, PCA)
- Determine Z based on distance to (the cluster center in K-means)

$$r_k^{(i)} = p(Z = k \mid \mathbf{x}^{(i)}) \propto \exp(-\beta \|\mathbf{x}^{(i)} - \mathbf{v}_k\|^2),$$

  where $\beta > 0$ is a constant, and $\mathbf{v}_k$ is a prototype for the cluster.
- Need to fit the prototypes $\mathbf{v}_k$. They are unknown.
- Similar to EM update, we let the reconstruction be

$$\tilde{\mathbf{x}}^{(i)} = \sum_{k=1}^{K} r_k^{(i)} \mathbf{v}_k$$

  and enforce that $\mathbf{x}^{(i)} \approx \tilde{\mathbf{x}}^{(i)}$ by minimizing

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

Learning Fair Representations

- Remember, we want to train a fair **classifier**.
- We predict using a linear function of $\mathbf{r}^{(i)} = [r_1^{(i)}, r_2^{(i)}, ..., r_K^{(i)}]^\top$.

$$y^{(i)} = \sigma(\mathbf{w}^\top \mathbf{r}^{(i)}) = p(t^{(i)}|\mathbf{x}^{(i)})$$

- Need to fit weights $\mathbf{w}$. They are unknown.
- Loss: we can use cross-entropy

$$L_{\text{CE}}(y^{(i)}, t^{(i)}) = -t^{(i)} \log y^{(i)} - (1 - t^{(i)}) \log(1 - y^{(i)})$$

Learning Fair Representations

- Next, enforce a fairness constraint:

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

- $N_0 = \#\{i : s^{(i)} = 0\}$, $N_1 = \#\{i : s^{(i)} = 1\}$ and $N_0 + N_1 = N$.

- Next, we show this enforces **demographic parity**.

- Note that $(Z \mid X) \perp\!\!\!\perp S$.

Learning Fair Representations

- Enforce **demographic parity** by obfuscating S:

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

- $N_0 = \#\{i : s^{(i)} = 0\}$, $N_1 = \#\{i : s^{(i)} = 1\}$ and $N_0 + N_1 = N$.
- If the above discrimination loss is $\mathcal{L}_{\text{discrim}} = 0$, we have LHS=RHS for all $k = 1, 2, ..., K$. Therefore,

$$\begin{aligned}
p(Y=1\,|\,S=1) &= \sum_k p(Y=1\,|\,Z=k)p(Z=k\,|\,X, S=1) \\
&\approx \sum_k p(Y=1\,|\,Z=k)\frac{1}{N_1}\sum_{i:s^{(i)}=1} p(Z=k\,|\,\mathbf{x}^{(i)}, s^{(i)}=1) \\
&= \sum_k p(Y=1\,|\,Z=k)\frac{1}{N_0}\sum_{i:s^{(i)}=0} p(Z=k\,|\,\mathbf{x}^{(i)}, s^{(i)}=0) \\
&\approx \sum_k p(Y=1\,|\,Z=k)p(Z=k\,|\,X, S=0) \\
&= p(Y=1\,|\,S=0) \quad \textbf{demographic parity}
\end{aligned}$$

Learning Fair Representations

- We want to retain information about X: $\mathbf{x}^{(i)} \approx \tilde{\mathbf{x}}^{(i)}$ penalize reconstruction error

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

- Predict accurately: cross-entropy loss

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

- Obfuscate S:

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

Learning Fair Representations

- We can solve the following problem

$$\mathcal{L}_{\text{total}}(\{\mathbf{v}_k\}_{k=1}^K, \mathbf{w}) = \lambda_r \mathcal{L}_{\text{reconst}} + \lambda_p \mathcal{L}_{\text{pred}} + \lambda_d \mathcal{L}_{\text{discrim}}$$

  where λ_r, λ_p, and λ_d are hyperparameters governing the trade-off between losses.
- We can find the optimal parameter $\{\mathbf{v}_k\}_{k=1}^K, \mathbf{w}$ using an optimization method such as gradient descent.

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

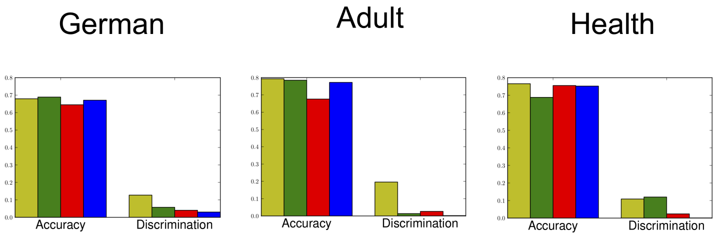
Task: predict whether patient spends any nights in hospital

Sensitive feature: Age

Metrics

- Classification accuracy
- Discrimination: measuring the difference in proportion of positive classification of individuals in the protected or unprotected groups.

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



Blue = theirs, others: logistic reg (LR), naive Bayes, regularized LR

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
  - Keep in mind that you'd need to carefully consider the assumptions when applying one of these methods!

Equalized Odds / Equal Opportunity

- There are several scores to measure the "fairness" of a model.
- Two notions of individual fairness (Hardt et al., 2016):
  - **Equalized odds**: equal true positive and false positive rates

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

  - **Equal opportunity**: equal true positive rates

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

Fairness 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
- Conference on Fairness, Accountability, and Transparency (FAT*)
- New textbook: `https://fairmlbook.org/`