Fairness, bias, and transparency in Machine Learning

Module 3: Fairness and bias interventions

This module

- 3.1: Introduction and categorization of fairness interventions
  - 3.2: Debiasing by pre-processing
  - 3.3: Debiasing by in-processing
  - 3.4: Debiasing by post-processing
- 3.5: Limits of bias and fairness interventions (mostly just final thoughts)
- Case study: Does mitigating ML's impact disparity require treatment disparity?
- Case study: AI-moderated decision-making: capturing and balancing anchoring bias in sequential decision tasks

(approx. 1.5 weeks)

Fairness and bias interventions

3.1: Introduction and categorization of fairness interventions

This section

- Introduce the topic of model interventions non-technically
- Explore general principles in terms of what is desired from a model “intervention”

What do we want from a fairness intervention?

Food for thought: (former) CSE enrollment lottery:

Instead of enrolling students holistically or based on GPA, the department selects at random — assuming they exceed the 3.3 CSE GPA threshold. With the lottery system, all students are equally considered despite differences in their experience, drive, and ability.

When asked about the implications of the new system — and possible disadvantage to high-performing students — CSE Chair [redacted] explained, ***"a lottery, by definition, is fair."***

"I think there's this false assumption that the students who work harder are the ones who are getting the 4.0s, that hard work directly translates to a higher grade. [The lottery system will] admit a lot of hard-working students who weren't getting in before," [CSE Vice-Chair for Undergraduate Education] added.

https://computinged.wordpress.com/2020/06/22/managing-cs-major-enrollment-boom-with-a-lottery-a-lottery-by-definition-is-fair/

What do we want from a fairness intervention?

Food for thought: "race blind" admissions

- Conducting “race blind” (for example) college admissions might have potential benefits, e.g. by helping to eliminate preconceived biases
- E.g. resumes with white-sounding names receive 50% more callbacks (see study below which explores randomly swapping names on resumes); symphony orchestras went from < 5% female to over 30% female after blinding auditions (among other things)
- Arguably, this type of blinding helps to reduce *implicit discrimination* simply by removing access to biased variables

see: Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination (2004)
https://www.theguardian.com/women-in-leadership/2013/oct/14/blind-auditions-orchestras-gender-bias

What do we want from a fairness intervention?

Both of the above interventions make "fairer" decisions by *removing* information from the decision-making process

Even if it increases fairness, won't that reduce accuracy, or have other unintended consequences?

What general principles should we use to design fairness interventions, and how can we reconcile improvements in fairness with other potential harms?

What do we want from a fairness intervention?

General (or “ethical”) principles:

- “Do no harm” (sometimes called “*non-maleficence*”): methods for fair machine learning should not harm *any* group

- Weaker statement: methods for fair machine learning should not harm members of the *protected* group.
- E.g. no member of the protected group should have their job application rejected if it would have been accepted *without* the intervention

see e.g. https://proceedings.mlr.press/v97/ustun19a/ustun19a.pdf

What do we want from a fairness intervention?

General (or “ethical”) principles:

- “Do one’s best” (sometimes called “*beneficence*”): methods for fair machine learning should attempt to be as accurate as possible for each group
- E.g. one could trivially make a classifier “fairer” by randomizing its outputs, or by always returning true (or false)

see e.g. https://proceedings.mlr.press/v97/ustun19a/ustun19a.pdf

What do we want from a fairness intervention?

Other than what is desirable, what is even allowable?

- Can we use the sensitive attribute to make decisions (illegal for college applications!)
- If we *can't* use it for decisions, can we use it for tuning a model, or selecting/weighting samples from the dataset (assuming the final model doesn't use the sensitive attribute)?
- What is the price (e.g. in terms of accuracy) for making a model more "fair"?
- Can any model be truly fair, other than a trivial model?

(we'll revisit these questions at the end of the module)

see e.g. https://proceedings.mlr.press/v97/ustun19a/ustun19a.pdf

Intervention strategies

In the following sections, we'll divide intervention strategies into a few broad categories:

- **Pre-processing :** modify the *dataset* to improve the outcomes of methods trained on that dataset
- **In-processing :** modify the *training objective* e.g. to include a fairness penalty
- **Post-processing :** modify the model's *outputs* (e.g. predictions) to correct outcomes after-the-fact

Fairness and bias interventions

3.2: Debiasing by pre-processing

This section

- Explore three specific schemes to pre-process datasets in order to get more fair outcomes (massaging, reweighting, sampling)

How can we process data to make outcomes more fair?

We'll look at three types of approaches; very roughly:

- **Massaging:** reduce discrimination in the training set by switching some of the labels (advantaged/disadvantaged group to negative/positive)
- **Reweighting:** instead of changing the labels in the training set, attach different weights to each dataset instance so that some training instances are more "important" than others
- **Sampling:** similar to reweighting, but some training samples are (randomly) used more often than others

from "Data Pre-Processing Techniques for Classification without Discrimination", Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

The paper (in footnote) isn't something I'd treat as "the" reference on the topic, but is a very readable introductory paper that gives a sense of how an intervention should be designed (even if the specific interventions described here are just one possible solution)

from "Data Pre-Processing Techniques for Classification without Discrimination", Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Some notation from this paper:

- **D** : dataset
- **X** : instance in D
- **S** : sensitive attribute
- **b / w** : disadvantaged or advantaged group (X(S) = b or X(S) = w)
- **+ / -** : positive and negative class
- **X(Class) = +** : label from the dataset
- **C(X) = +** : prediction from the classifier
- **D_w** : number of instances X(S) = w (etc.)
- **p_w** : number of *positively labeled* instances with X(S) = w (etc.)

from “Data Pre-Processing Techniques for Classification without Discrimination”, Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Mapping it to our own notation...

- **D** : dataset
- **x_i** : instance in D
- **z** : sensitive attribute
- **z_i = 1 / z_i = 0** : disadvantaged or advantaged group
- **y_i = 1 / y_i = 0** : positive and negative class
- **y_i** : label from the dataset
- **\hat{y}** : prediction from the classifier
- **D_0** : number of instances with z_i = 0 (etc.)
- **p_0** : number of *positively labeled* instances with z_i = 0 (etc.)

from “Data Pre-Processing Techniques for Classification without Discrimination”, Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Some definitions: discrimination in a labeled **dataset**

from “Data Pre-Processing Techniques for Classification without Discrimination”, Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Some definitions: discrimination in a labeled **dataset**

$$disc_{S=b}(D) := \frac{|\{X \in D \mid X(S) = w, X(Class) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, X(Class) = +\}|}{|\{X \in D \mid X(S) = b\}|}$$

from “Data Pre-Processing Techniques for Classification without Discrimination”, Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Some definitions: discrimination in a **classifier's predictions**

from "Data Pre-Processing Techniques for Classification without Discrimination", Kamiran and Calders, 2011

How can we process data to make outcomes more fair?

Some definitions: discrimination in a **classifier's predictions**

$$disc_{S=b}(C, D) := \frac{|\{X \in D \mid X(S) = w, C(X) = +\}|}{|\{X \in D \mid X(S) = w\}|} - \frac{|\{X \in D \mid X(S) = b, C(X) = +\}|}{|\{X \in D \mid X(S) = b\}|}$$

from "Data Pre-Processing Techniques for Classification without Discrimination", Kamiran and Calders, 2011

Massaging

- Change the label of some objects with X(S) = b from - to +
- Change the label of some objects with X(S) = w from + to -

(S = sensitive attribute; b = “deprived” class; w = “non-deprived” class; -/+ = negative/positive label)

i.e., change the labels for some members of the disadvantaged group to positive, and change *the same number of* labels of some members of the non-disadvantaged group to negative

Discrimination (in the training set) reduces, but the class distribution remains fixed

Massaging

What strategy should we use to select promotion candidates and demotion candidates?

(e.g. for disadvantage=female) “Promote” the highest scoring female *with a negative label*, and “demote” the lowest scoring male *with a positive label*

Paper argues that this strategy will have minimal effect on accuracy (we are basically relabeling the instances the classifier “wants” to relabel anyway). Why is this a good heuristic?

Massaging

Algorithm 2: *Rank*

Input: Labeled dataset D, Sensitive attribute and value S, b, desired class $+$
Output: Ordered promotion list pr and demotion list dem

1: Learn a ranker R for prediction $+$ using D as training data
2: $pr := \{X \in D \mid X(S) = b, X(Class) = -\}$
3: $dem := \{X \in D \mid X(S) = w, X(Class) = +\}$
4: Order pr descending w.r.t. the scores by R
5: Order dem ascending w.r.t. the scores by R
6: **return** (pr, dem)

Massaging

How many instances do we have to change to reach zero discrimination (pg. 15 from paper)?

Food for thought

Do you have any feelings about this? Flipping labels in a dataset (basically making the data “fake”!) seems like an intrusive operation

But is it any worse than (e.g.) removing some outliers or assigning more/less weight to some instances?

Reweighting

Instead of *changing* the labels in the dataset (which is a rather intrusive operation!), attach different *weights* to each dataset instance

recall: optimizing a balanced error rate, from Module 1

(**Q:** which fairness objective does this optimize?)

Reweighting

- Instead of *changing* the labels in the dataset (which is a rather intrusive operation!), attach different *weights* to each dataset instance
- This time, we want something like the following:

    - objects with *X(S)=b* and *X(class)=+* (i.e., disadvantaged class, positive label) should have higher weights than objects with *X(S)=b* and *X(Class)=-* (i.e., disadvantaged class, negative label)

    - objects with *X(S)=w* and *X(Class)=+* will get lower weights than objects with *X(S)=w* and *X(Class)=-*

Reweighting

Our fairness goal here is that the sensitive attribute should be independent of the class label, i.e.,

should match the observed probability (pg. 16 from paper):

Reweighting

Q: How can we weight instances (roughly speaking, attaching a “probability” to each instance) to achieve independence?

A: Weight of an object should be the *expected probability of seeing an instance with X(S) and X(Class)*, ***divided by the observed probability*** (pg. 17 from paper):

Reweighting

Algorithm 3: *Reweighing*

Input: $(D, S, Class)$
Output: Classifier learned on reweighed D

1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \frac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: $D_W := \{\}$
7: **for** X in D **do**
8: Add $(X, W(X(S), X(Class)))$ to D_W
9: **end for**
10: Train a classifier C on training set D_W, taking onto account the weights
11: **return** Classifier C

Sampling

- Basic idea is very similar to reweighting – just a way of adjusting the idea
- Instead of upweighting some data points, just *sample* the data such that certain data points show up more (or less) frequently than others

Sampling

Algorithm 4: *Uniform Sampling*

Input: $(D, S, Class)$
Output: Classifier C learned on resampled D

1: **for** $s \in \{b, w\}$ **do**
2: **for** $c \in \{-, +\}$ **do**
3: Let $W(s, c) := \dfrac{|\{X \in D \mid X(S) = s\}| \times |\{X \in D \mid X(Class) = c\}|}{|D| \times |\{X \in D \mid X(Class) = c \text{ and } X(S) = s\}|}$
4: **end for**
5: **end for**
6: Sample uniformly $W(b, +) \times |DP|$ objects from DP;
7: Sample uniformly $W(w, +) \times |FP|$ objects from FP;
8: Sample uniformly $W(b, -) \times |DN|$ objects from DN;
9: Sample uniformly $W(w, -) \times |FN|$ objects from FN;
10: Let D_{US} be the bag of all samples generated in steps 6 to 9
11: **return** Classifier C learned on D_{US}

Experiment

Code example: Data massaging (mostly, just clarifying how some of the quantities in the paper are computed)

workbook3.iypnb

Other strategies

Lots of other approaches (see e.g. https://cseweb.ucsd.edu/~jmcauley/pdfs/nips18.pdf for brief survey); above are chosen mostly because of their simplicity:

- Flipping labels of training samples
- Learn representations of data points such that cluster assignments can't be inferred from representations (e.g. clustering)

Study points & take-homes

- Can pretty easily get “more fair” outcomes by simple perturbation of datasets
- Try to implement interventions based on each of the three schemes
- The interventions we saw are (arguably) quite “invasive,” especially when replacing real data with “fake” data; on the one hand, training samples represent historical samples which can’t be “harmed;” on the other hand, such interventions would likely violate (e.g.) affirmative action rules (see later)

Bias and fairness interventions

3.3: Debiasing by in-processing

This section

- Explore schemes to modify model objectives in order to make models fairer (for classifiers and regressors), both based on convex optimization

Pre-processing vs in-processing

Our previous set of interventions (pre-processing) manipulated the *dataset*, which will hopefully encourage "standard" machine learning approaches to yield fair(er) results downstream

Instead, why not incorporate fairness objectives into the learning algorithm directly?

Doing so will (perhaps) give us tighter control when optimizing toward fairness goals, rather than "hoping" that pre-processing will work

Pre-processing vs in-processing

(Again) Haven’t we seen this sort of thing before (balanced classifier)?

Is this “pre-processing” or “in-processing”?

Disparate learning processes

Arguably, the approaches we've seen so far have a few issues:

- (Somewhat) heuristic: dataset interventions might not be guaranteed to yield fairer outcomes
- Many of the specific fairness goals we've seen so far (in the previous module) may be hard to address with a dataset intervention
- May have other side-effects, such as a loss in accuracy
- Not obvious how to generalize, e.g. if there are multiple sensitive attributes

more in "Mechanisms for Fair Classification", Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

Recall: our “reweighting” pre-processing intervention strategy tried to correct for dependence between the sensitive attribute and the class label, i.e., the *dataset* was adjusted (instances were reweighted) so that the sensitive attribute and class label would look independent.

Could we apply this type of intervention to a learning algorithm?

Idea:

Disparate learning processes

Some notation:

- **x:** feature vector
- **z:** sensitive attribute (not included in the feature vector)
- **d_\theta(x):** (signed) distance from decision boundary

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

Covariance between sensitive attribute and distance from decision boundary:

Recall (?): covariance:

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

Covariance between sensitive attribute and distance from decision boundary:

Recall (?): covariance:

$$\begin{aligned}\operatorname{cov}(X,Y) &= \operatorname{E}[(X - \operatorname{E}[X])\,(Y - \operatorname{E}[Y])] \\ &= \operatorname{E}[XY - X\operatorname{E}[Y] - \operatorname{E}[X]Y + \operatorname{E}[X]\operatorname{E}[Y]] \\ &= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y] - \operatorname{E}[X]\operatorname{E}[Y] + \operatorname{E}[X]\operatorname{E}[Y] \\ &= \operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y],\end{aligned}$$

(from wikipedia)

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

Covariance between sensitive attribute and distance from decision boundary:

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

The covariance now becomes a *constraint* on the model. That is, we want the best possible (most accurate) model *among models that have little correlation between the prediction and the sensitive attribute* (eq. 4 from paper):

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

E.g. for logistic regression:

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

I won't go through exactly how to optimize this, but this set of (convex) constraints doesn't make the problem much harder

from "Mechanisms for Fair Classification", Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

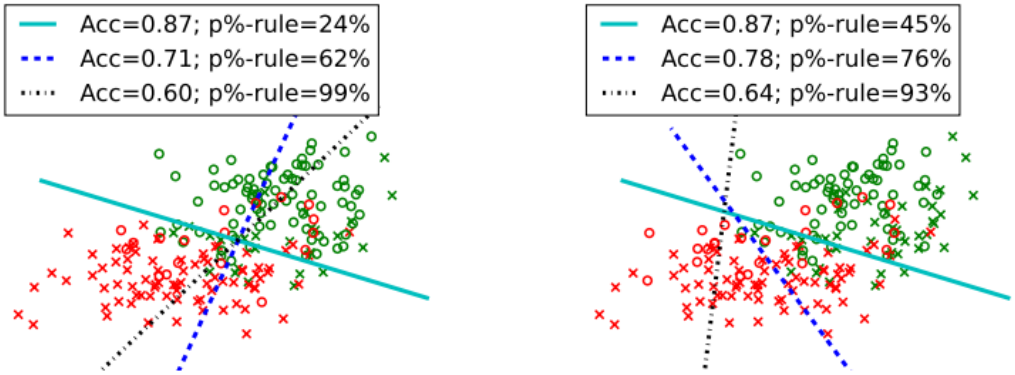
Disparate learning processes

Can easily (?) rewrite this as *maximizing fairness given an accuracy constraint* instead of *maximizing accuracy given a fairness constraint:*

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

Disparate learning processes

Tuning the threshold yields different classifiers that balance accuracy with fairness objectives (in this case, p%-rule):



max acc. s.t.
fairness constraint

max fairness s.t.
acc constraint

from “Mechanisms for Fair Classification”, Zafar et al. 2017 (https://arxiv.org/pdf/1507.05259)

What about regression?

Most of the methods in this module (and really, in this course) are focused on classification. But can the same ideas be applied to regression?

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

Yet more notation...

- **y** : regression target (in [-1, 1], why?)
- **x** : feature vector
- **S_1, S_2** : two “groups” (no rule about which contains the “sensitive” attribute); groups are made up of (x,y) pairs

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

Previously, we said that the *distance from the decision boundary* should not be correlated with the *sensitive attribute:*

$$\begin{aligned} \mathrm{Cov}(\mathbf{z}, d_{\boldsymbol{\theta}}(\mathbf{x})) &= \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})d_{\boldsymbol{\theta}}(\mathbf{x})] - \mathbb{E}[(\mathbf{z}-\bar{\mathbf{z}})]\bar{d_{\boldsymbol{\theta}}}(\mathbf{x}) \\ &\approx \frac{1}{N}\sum_{i=1}^{N}(\mathbf{z}_i-\bar{\mathbf{z}})\, d_{\boldsymbol{\theta}}(\mathbf{x}_i), \end{aligned}$$

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

For regression we'll explore a couple of related goals

Roughly speaking: if two individuals from different groups have *similar labels*, they should also have *similar predictions* (e.g. one group should not have their values overestimated more than the other group)

(**Q:** which of our fairness desiderata does this goal align with?)

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

We’ll extend our regression model to include an additional “fairness” term:

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

Our fairness term should then somehow express that *individuals from different groups should have similar predictions if they have similar labels*

Different groups:

Similar labels:

Similar predictions:

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

Individual fairness: a model is penalized for how differently it treats instances from two groups *x* and *x'*:

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

Group fairness: in the individual fairness definition, no “cancellation” occurs: it’s not okay to overestimate a bunch of predictions from one group if you overestimate a bunch of predictions from the *other* group. A weaker definition says that the errors in a group should just be *on average* the same as errors in the other group:

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

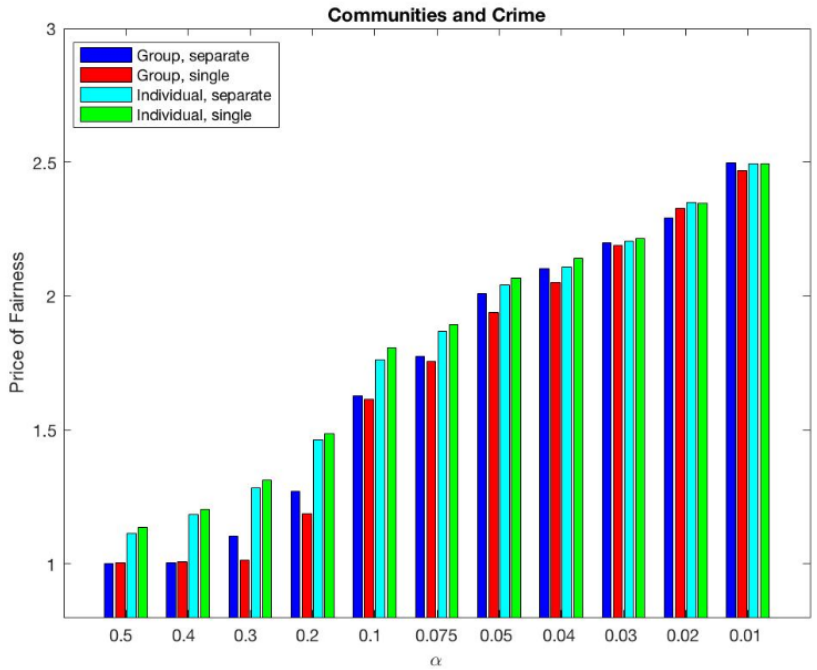
A convex framework for fair regression

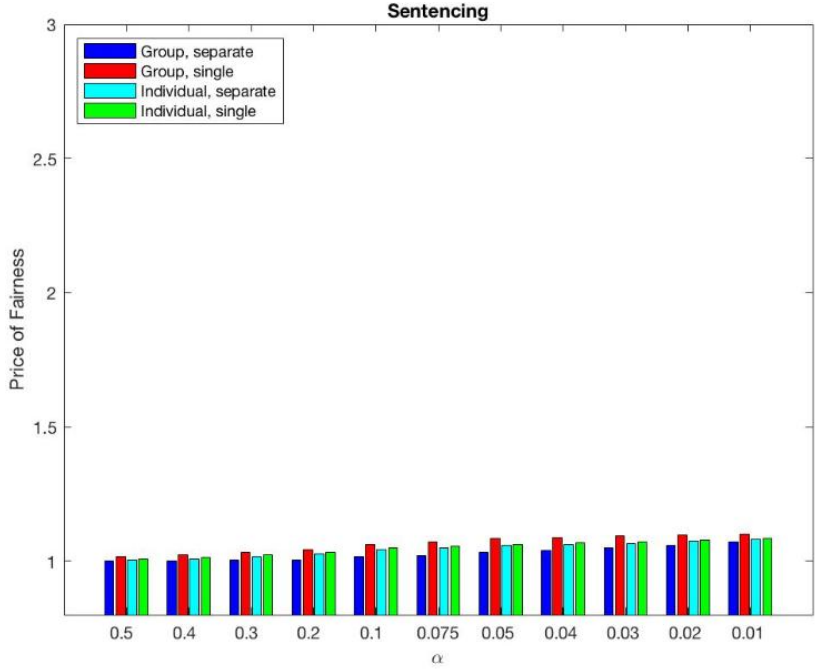
The price of fairness: how does the MSE of a fair model compare to that of an unconstrained model?

$$\mathrm{PoF}(\alpha) = \frac{\min_{\mathbf{w}} \ell_{\mathcal{P}}(\mathbf{w}) \text{ subject to } f_{\mathcal{P}}(\mathbf{w}) \leq \alpha f_{\mathcal{P}}(\mathbf{w}^*)}{\ell_{\mathcal{P}}(\mathbf{w}^*)}$$

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression





from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

A convex framework for fair regression

A few more things from the paper:

- The same idea can also be used as an intervention to train classifiers
- They also describe a “two model” variant, where each of the two groups is treated via a separate model

from https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf

Other strategies

Again, the above are just some representative approaches, usually just the “cleanest” way to show how methods in the category work

Other strategies

See also e.g. “Fair Variational Autoencoders” (link in footnote):

from https://arxiv.org/pdf/1511.00830

Food for thought

Any thoughts about these compared to (e.g.) dataset-based interventions?

- Although optimization should be "easy", in experiments (we'll see some later) implementations may involve approximations, whose suboptimal behavior may be hard to understand
- Is it actually more "generalizable"? What categories of models will these interventions work for?
- It's worth thinking about the difficulty of implementing these models (among the more complex ones we'll see in this course); compared to datase-based interventions, what is actually likely to be deployed?

Study points & take-homes

- Implementing these models is fairly difficult... but try to get a sense of how they work; implementations are available
- The methods we saw over very direct control in terms of fairness *guarantees* (i.e., in the form of a constraint); think about their advantages/disadvantages compared to dataset-based techniques (both technical merits as well as “softer” concerns)

Fairness and bias interventions

3.4: Debiasing by post-processing

This section

- Explore post-processing schemes to achieve fairness goals on model outputs, i.e., using "pre-trained" models

Why post processing?

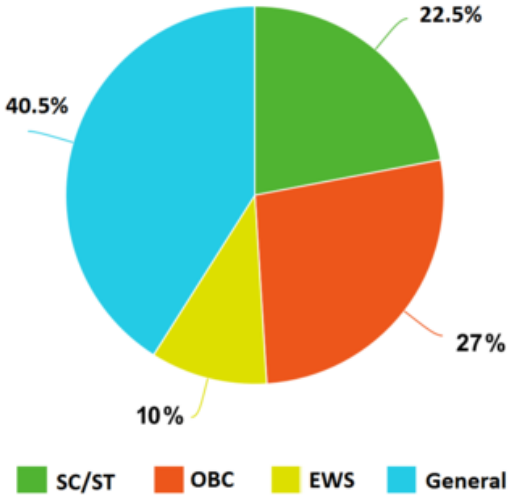
- Might want methods that we can apply to models that already exist, without needing to retrain the model – that is, which work directly with the *outputs of the model*
- (Possibly) these techniques can apply to a wide variety of models, to the extent that they operate over the predictions directly (i.e., they can be *model agnostic*)

see e.g. https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Affirmative action

Colloquially, ***Affirmative Action*** refers to a policy explicitly designed to benefit members of marginalized groups

E.g. quota systems, various examples of policies in India, South Africa, Norway, etc.



Caste reservations in universities and government jobs (from wikipedia)

Affirmative action

How would we implement such a policy for an (ML) algorithm?

Pretty simple:

- Train a classifier $f(x) \rightarrow y$ (e.g. to classify candidates as qualified for a job)
- **Optionally:** train *different* classifiers for members of each group

Affirmative action

- Select best (e.g.) female applicants according to:
- Select best male (any other grouping) applicants according to:
- Draw a number of candidates from each list in order to achieve any desired proportional representation

Affirmative action

Equivalently, set per-group thresholds:

Affirmative action

What's good about this algorithm?

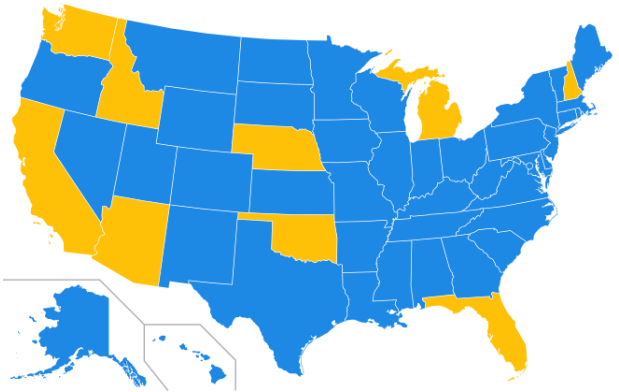
- **Trivial** to implement
- **Controllable:** can guarantee the exact proportions that we want
- Doesn't compromise **accuracy** for the sake of achieving fairness: we can use our best possible algorithm, without needing to "hide" attributes
- Later (from case study): ***cannot be beaten*** in terms of its accuracy / fairness tradeoff

Affirmative action

What's not good about this algorithm?

It's (often) illegal!

In practice, it's sometimes ***allowed***, sometimes ***required*** (e.g. a Canadian act requires employers in certain industries to give preferential treatment to certain groups). But it's often **illegal**



Status of affirmative action in the US:
Yellow: banned
Blue: not banned (*but still depends on setting in which it's applied)
(from wikipedia)

Affirmative action

A few points:

- Of course, there are other objections to affirmative action besides it being illegal
- But in the context of this class, we *want* some sort of intervention to correct disparities, so affirmative action should presumably be “on the table”, were it legal
- There are also many contexts where it *isn’t* illegal – not everything is about hiring! If you want to balance (e.g.) author gender in a book recommender system, *affirmative action* is (probably) the algorithm you should apply!

Score functions, calibration and fairness

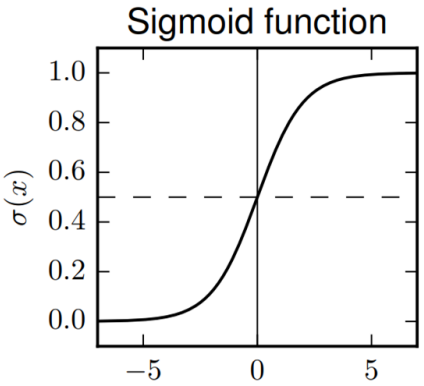
In any case, what can we do in terms of post-processing, at least assuming it's allowed? (Which it often is – not all settings are legally fraught!)

- Based on notes from https://afraenkel.github.io/fairness-book/content/07-score-functions.html
- Which is itself based on the *Fairness and Machine Learning* (Barocas *et al.*)

Score functions, calibration and fairness

(Most of) the classifiers we've looked at have a **score function** associated with their predictions, e.g. a sigmoid function in the case of logistic regression

which we can interpret as a probability:



see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Score functions, calibration and fairness

More generally, other classifiers may be associated with a score function $S(x)$ and an arbitrary threshold t:

Question: how should we choose t?

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Score functions, calibration and fairness

Recall: ROC curves (from Module 1)

Could choose a classifier by:

- Selecting the threshold that achieves the highest accuracy (if False Positives and False Negatives are equally bad)
- Selecting the threshold that achieves a desired balance between FPs and FNs

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Score functions, calibration and fairness

Q: Can we choose threshold values to **optimize fairness objectives?** In other words, can we make an unfair classifier more fair by modifying its thresholds?

Doing so is a **post-processing** intervention, in the sense that we don't need to modify the model or know anything about it

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Score functions, calibration and fairness

Formally, we'll build a new "derived classifier" $Y = F(S, z)$ that applies some operation to the score function S; e.g.:

- Apply different thresholds to different groups
- Apply some randomization procedure e.g. to select among thresholds

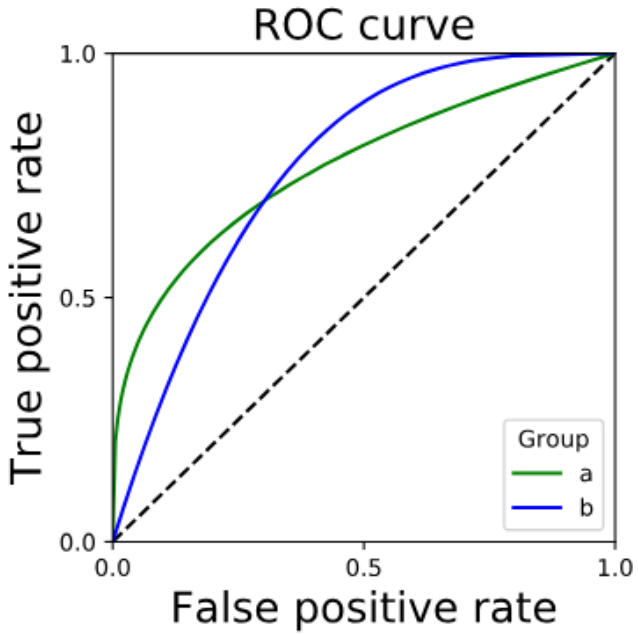
see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds

E.g. can we build a derived classifier that satisfies **equalized odds?**

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds



(from http://www.fairmlbook.org)

As we vary our threshold, “unfair” classifiers will have different FPRs/FNRs across different groups

What point on this curve satisfies the notion of **equalized odds** across the two groups?

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds

The point of intersection will generally correspond to a different threshold value for the two groups (*a* and *b*), so we'll need two thresholds *t_a* and *t_b* to build our derived classifier:

$$F(x, c) = \begin{cases} 0 \text{ if } (S|_a(x) < t_a) \text{ and } c = a \\ 0 \text{ if } (S|_b(x) < t_b) \text{ and } c = b \\ 1 \text{ if } (S|_a(x) \geq t_a \text{ and } c = a \\ 1 \text{ if } (S|_b(x) \geq t_b) \text{ and } c = b \end{cases}$$

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds

But what if the two curves never intersect?

(alternately, might have specific requirements in terms of TPR/FPR not satisfied by the intersecting point)

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds

Can choose a model that's **under** both ROC curves!

Procedure:

- Consider two classifiers *f* and *g*
- Classifiers on the *line between them* are realized by:

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Equalized odds

Note: any such classifier that's *under* the ROC curve will be suboptimal (in terms of TPR/FPR etc.); thus we are sacrificing utility for accuracy

Also note: the procedure above is a form of **affirmative action:** we looked explicitly at the group labels (in this case, at inference time) to achieve specific fairness goals

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Food for thought

Consider the second case above (one group has strictly worse results for all FPR/TPR values)

Is it justifiable to *deliberately* decrease performance for the “advantaged” group to achieve equality, even though performance for the “disadvantaged” group won’t be improved?

Calibration

A “calibrated” classifier (Module 2) is one that for all scores s exhibits

For example for COMPAS, among defendants receiving the lowest risk score (of 0.1), those that re-offend should make up 10% of that group; the highest risk score (0.9) should consist of 90% of people who did re-offend

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Calibration: Platt scaling

To calibrate any classifier (scoring function) $S(x)$, we can replace our classifier by a new one of the form:

A and B are parameters; they must be chosen so that the classifier is calibrated; see Platt's slides (https://www.cs.cornell.edu/courses/cs678/2007sp/platt.pdf) for details

see https://afraenkel.github.io/fairness-book/content/07-score-functions.html

Learning to defer

One more post-processing intervention, just to get a different take:

- Consider a scenario where a decision maker and an algorithm are working together to make decisions
- Further assume that the model has already been trained using some fairness intervention such that *it is fair, but possibly at the cost of accuracy*
- The (human) decision maker is (relatively) high accuracy, but has biases

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

Q: How can we combine (fair but inaccurate) machine predictions with (accurate but biased) human predictions?

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

Idea 0: Let the model *reject* making predictions in cases where it is not sufficiently confident (assuming the model is capable of outputting "confidence" estimates)

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

Idea 0: Let the model *reject* making predictions in cases where it is not sufficiently confident (assuming the model is capable of outputting "confidence" estimates)

Problem: both the model and the decision-maker act *independently* of one another; but, the decision to reject should depend on the model's confidence, *as well as the decision-maker's expertise and weaknesses*.

E.g.: the model might be uncertain about some subgroup, but the expert may be biased against that subgroup: we might prefer the model's predictions in spite of its uncertainty

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

Learning to defer is somewhat similar to a Mixture of Experts framework: roughly speaking, each expert gets to "vote" on the outcome, and we also predict a confidence score associated with each expert

In this way, we can simultaneously train both the confidence function, and the parameters of the experts themselves, *such that the experts only need to be accurate about those instances where they are confident*

(main difference is that we can't train the "parameters" of the human decision makers, though we can still make confidence estimates for them)

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Learning to defer

(see paper for more training details)

Mostly just wanted to mention this one to give a sense of other types of possibilities: e.g. our goal needn't be to "replace" humans by algorithms, but might instead be to reduce biases in a setting with humans in the loop

We'll look at a **case study** later in which we algorithmically manipulate a user interface such that humans make less biased decisions

see https://arxiv.org/abs/1711.06664 and https://hci.stanford.edu/courses/cs335/2020/sp/lec9.pdf

Study points & take-homes

- By now, try to understand when dataset-based, (pre-processing), model-based (in-processing) interventions, or post-processing algorithms might be preferred
- Understand the constraints involved in terms of:
  - What is actually allowed (e.g. in terms of using the protected attribute)
  - What is controllable or predictable
  - What is practical to implement, and perhaps most likely to be adopted

Fairness and bias interventions

3.5: Limits of bias and fairness

This section

- Conclude by thinking about broader limitations of fairness research, and their connection to issues in society

Algorithmic fairness

(we saw this slide at the beginning of the module):

- Conducting “race blind” (for example) college admissions might have potential benefits, e.g. by helping to eliminate preconceived biases
- E.g. resumes with white-sounding names receive 50% more callbacks (see study below which explores randomly swapping names on resumes); symphony orchestras went from < 5% female to over 30% female after blinding auditions (among other things)
- Arguably, this type of blinding helps to reduce *implicit discrimination* simply by removing access to biased variables

see: Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination (2004)
https://www.theguardian.com/women-in-leadership/2013/oct/14/blind-auditions-orchestras-gender-bias

Algorithmic fairness

- So: humans make less biased decisions when denied access to the sensitive attribute
- Does this imply that sensitive attributes should be unavailable to algorithms?

Algorithmic fairness

- The paper *Algorithmic Fairness* (Kleinberg *et al.* https://www.cs.cornell.edu/home/kleinber/aer18-fairness.pdf) argues that while it may be desirable to set different *thresholds* per group (to achieve specific fairness outcomes), trying to blind algorithms to sensitive characteristics will ultimately lead to less fair decisions (to say nothing of its legality)
- (We'll explore a similar result from a different case study later)

Fairness vs model performance

Most of the fairness interventions we've described impact model performance

- In some cases (e.g. building fair recommender systems), there may be many nearly-equivalent outputs (in terms of some value function) that we could choose, and choosing a "fairer" one costs us very little; in others, achieving fairness will lead to a serious performance gap
- Who will pay for that performance gap?
- How will any company be incentivized to even *uncover* fairness issues, given that fixing them will likely be costly (in terms of model performance)
- Little understanding in fairness literature about the real "cost" of implementing fairness interventions in high-stakes settings (compare to e.g. research on differential privacy)

see https://arxiv.org/pdf/2010.04053

Incompatible notions of fairness

(Already covered this one a fair bit)

- Other than the definitions being incompatible, there's no real consensus about what definitions should be used in a particular context
- Fairness definitions from ML do not necessarily map exactly to legal, social, or economic understanding of the same issues

see https://arxiv.org/pdf/2010.04053

Tensions with context and policy

- ML researchers generally use "convenience" datasets to study fairness, which might have very little connection to current issues faced by industrial ML practitioners, or broader societal issues
- E.g. most datasets are dated, small, and focused on a very limited set of sensitive attributes (race or gender), which might mask all sorts of other biases
- On the other hand, industry practitioners have very little incentive to share data, *or even measure whether bias exists in the first place*

see https://arxiv.org/pdf/2010.04053

Fairness and bias interventions

Case study: Does mitigating ML's impact disparity require treatment disparity?

Recall: impact disparity and treatment disparity

Treatment disparity: Algorithms exhibit *treatment disparity* if they formally treat members of protected subgroups differently

Impact disparity: Algorithms exhibit *impact disparity* when outcomes differ across subgroups (even unintentionally)

link to paper: https://arxiv.org/pdf/1711.07076

Impact disparity and treatment disparity

One can achieve *impact parity* through deliberate *treatment disparity*

Affirmative action is an example of such a process (albeit not a legal one in some contexts!): two groups are deliberately *treated* differently, in order to achieve equal *outcomes* (impact) across the two groups.

Affirmative action achieves this pretty straightforwardly: just use the sensitive attribute to set different decision thresholds for each group

(see: earlier in the module)

link to paper: https://arxiv.org/pdf/1711.07076

Impact disparity and treatment disparity

Disparate learning processes (which we also saw earlier in the module) try to achieve impact parity *without* treatment disparity, that is without using the sensitive attribute *at inference time*

They do this by using the sensitive attribute at *training time* in order to perturb model weights, but do not use the sensitive attribute at *test time*

Food for thought: avoiding the sensitive attribute at test time is mostly done to overcome legal barriers; but is using it at training time (legal or otherwise) really "better"?

link to paper: https://arxiv.org/pdf/1711.07076

Impact disparity and treatment disparity

This paper (“Does mitigating ML's impact disparity require treatment disparity?”) studies what is the “cost” of implementing a disparate learning process, especially compared to just implementing affirmative action

- Obviously, depriving ourselves of the sensitive attribute at inference time removes information from the process (whereas with affirmative action we know everything)
- On some level, removing information from the process might be harmful to the decisions made by the algorithm
- Which groups (or subgroups) are harmed?
- *Is it worth applying these types of intervention?*

link to paper: https://arxiv.org/pdf/1711.07076

Disparate learning processes

Quick reminder:

At training time, find the best possible (most accurate) model among models that have *little correlation between the prediction and the sensitive attribute*

link to paper: https://arxiv.org/pdf/1711.07076

Disparate learning processes

But! The algorithm must still make use of available attributes; if those attributes are themselves correlated with the protected attribute, the decision function may indirectly encode the sensitive attribute

And because it does so indirectly, it may make worse (we'll see how) decisions

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

Consider a synthetic dataset with a few available attributes:

- Years of work experience
- Gender (sensitive attribute)
- label (whether the person is qualified)
- Hair length – an *irrelevant attribute* (i.e., no relation to qualification), which is *correlated with* the sensitive attribute

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

Data is generated randomly:

$$
\begin{aligned}
z_i &\sim \text{Bernoulli}(0.5) \\
\text{hair\_length}_i \mid z_i = 1 &\sim 35 \cdot \text{Beta}(2, 2) \\
\text{hair\_length}_i \mid z_i = 0 &\sim 35 \cdot \text{Beta}(2, 7) \\
\text{work\_exp}_i \mid z_i &\sim \text{Poisson}(25 + 6z_i) - \text{Normal}(20, \sigma = 0.2) \\
y_i \mid \text{work\_exp} &\sim 2 \cdot \text{Bernoulli}(p_i) - 1, \\
\text{where } p_i &= 1/\left(1 + \exp[-(-25.5 + 2.5\text{work\_exp})]\right)
\end{aligned}
$$

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

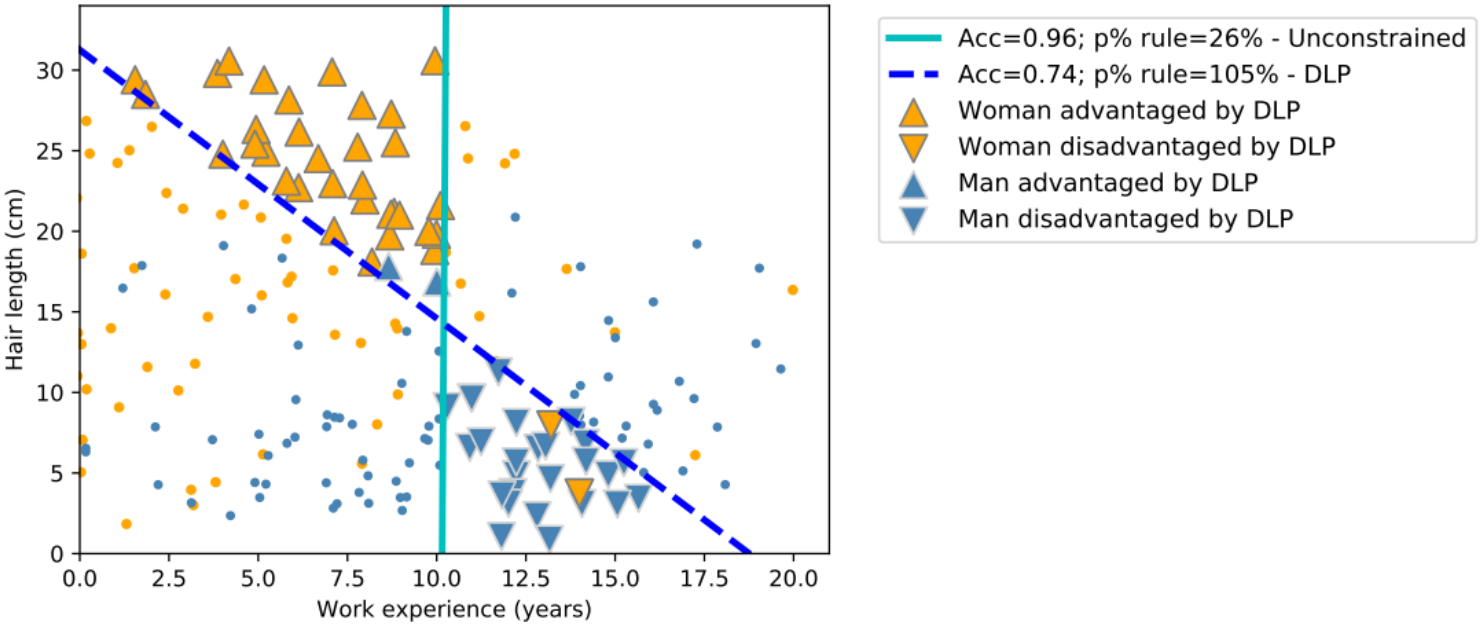
Summary of randomly generated data:

1. Historical hiring is based *solely on years of work experience;*
2. Women have fewer years of work experience (5 vs 11 on average), causing men to have been hired at a much higher rate than women
3. Women have longer hair than men, though this is irrelevant to historical hiring practice

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

What's the outcome of applying a DLP to this data?



link to paper: https://arxiv.org/pdf/1711.07076

(recall) p-% rule

Impact disparity is sometimes measured using a quantity known as the p-% rule which measures ratio between the probability of being assigned to the positive class for the advantaged versus disadvantaged group

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

What's the outcome of applying a DLP to this data?

- The unconstrained classifier hires based on work experience, as expected, though it has a low p%-rule
- The DLP does indeed achieve near-parity (i.e., it has a p%-rule of close to 100%), i.e., it fulfills its desired objective
- However, it does so by differentiating based on an irrelevant attribute (hair length)
- **Note:** the process hurts some short-haired women, and helps some long-haired men

link to paper: https://arxiv.org/pdf/1711.07076

(Synthetic) example

- This synthetic example is, well, pretty synthetic; real datasets wouldn't have a feature for hair length, for example
- **But:** it shows that fairness interventions will make use of attributes that are correlated with the sensitive attribute
- **Q:** Who is harmed by this intervention?
- **Q:** Who benefits from this intervention?

link to paper: https://arxiv.org/pdf/1711.07076

Another (slightly less synthetic) example

UCSD graduate admission data!

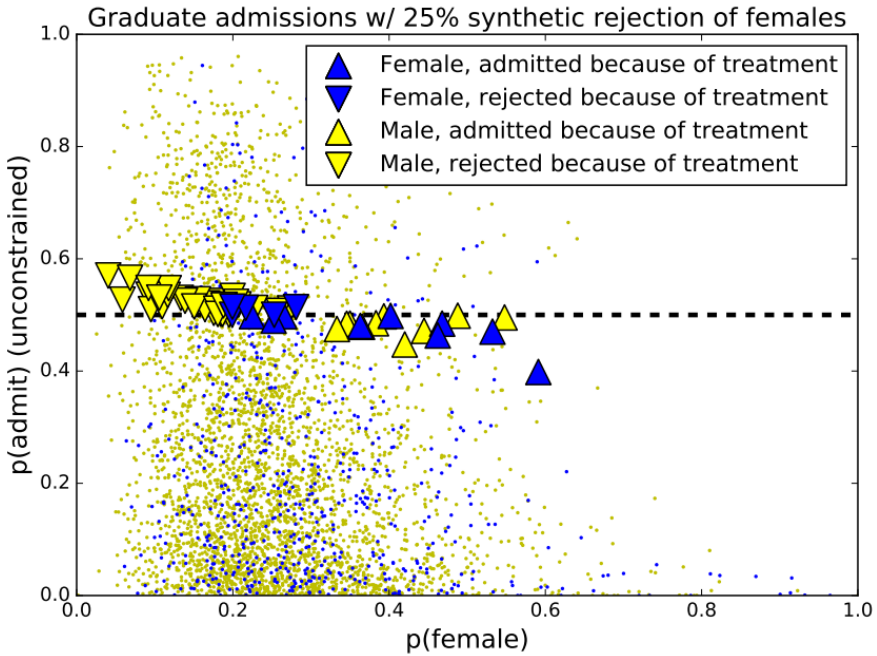
- Real data of our graduate admissions process in CSE (IRB approved!)
- Features include schools, GPA, GREs, letter percentiles, etc.
- Labels are admissions outcomes
- Sensitive attribute is still gender

The historical data doesn't exhibit significant gender bias, *so female admits have their labels flipped to "reject" 25% of the time*

What does the fairness intervention (DLP) do on this data?

link to paper: https://arxiv.org/pdf/1711.07076

Another (slightly less synthetic) example

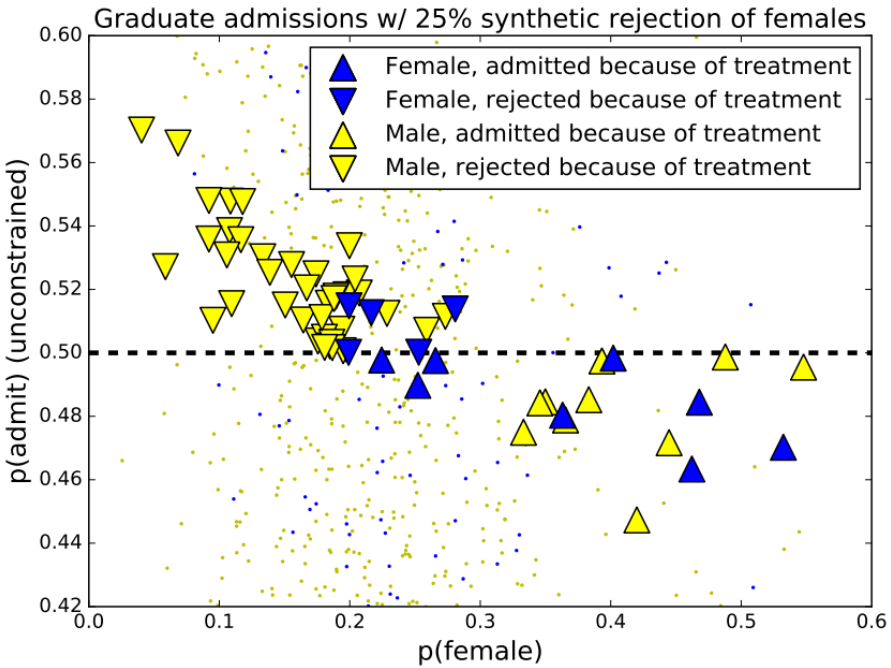


y-axis: unconstrained classifier (logistic regression)

x-axis: logistic regressor to predict gender from features

link to paper: https://arxiv.org/pdf/1711.07076

Another (slightly less synthetic) example



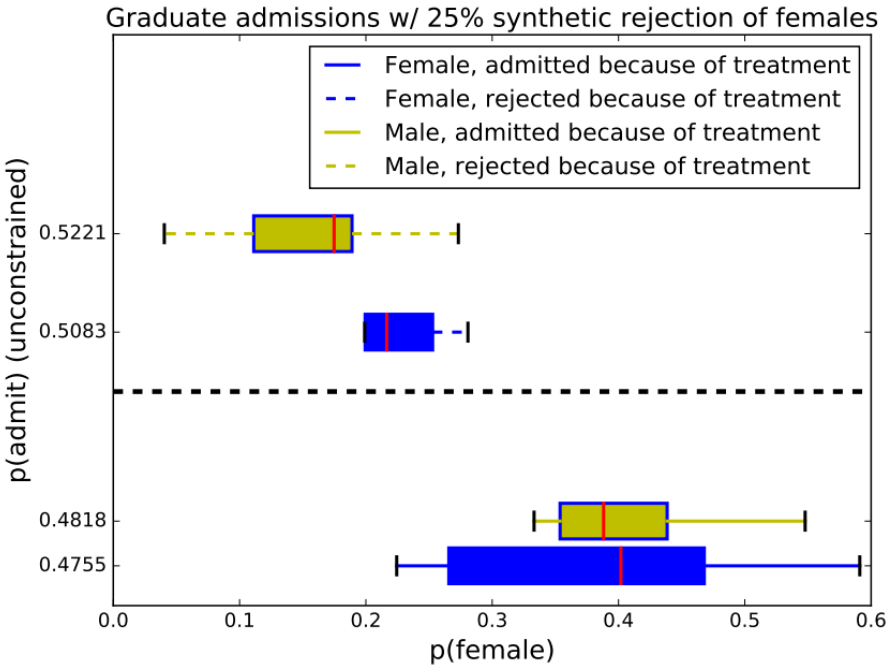
y-axis: unconstrained classifier (logistic regression)

x-axis: logistic regressor to predict gender from features

(detail of same plot)

link to paper: https://arxiv.org/pdf/1711.07076

Another (slightly less synthetic) example



y-axis: unconstrained classifier (logistic regression)

x-axis: logistic regressor to predict gender from features

(summary statistics of same plot)

link to paper: https://arxiv.org/pdf/1711.07076

Another (slightly less synthetic) example

Findings:

- Attributes encode gender to some extent, since the gender classifier is much better than random (**Q:** what attributes might predict gender?)
- Like our hair length example:
  - Students benefiting from the DLP are males who ‘look like’ females based on other features
  - Females who ‘look like’ males are hurt by the DLP

link to paper: https://arxiv.org/pdf/1711.07076

So what?

The fairness intervention violates the *do no harm* principle (one of our desired goals of a fairness intervention), that is, it disadvantages some women (members of the protected group) who, but for the DLP, would have been admitted

link to paper: https://arxiv.org/pdf/1711.07076

So what?

Q: But is it possible to do any better?

Well sure – none of these problems apply to an “affirmative action”-type intervention!

Recall: affirmative action ranks groups separately based on the sensitive attribute (male or female) and admits the top candidates from each list

link to paper: https://arxiv.org/pdf/1711.07076

So what?

Note: can design an “affirmative” action protocol that is optimal in terms of the p%-rule. Rough outline:

- Start with the accuracy-maximizing classifications
- Our affirmative action policy will admit some females (disadvantaged group) and reject some males (advantaged group) to achieve a desired p%-rule
- Changing any label will reduce the accuracy; we want the ***smallest reduction in accuracy*** for the ***biggest increase in p%-rule***

Algorithm:

- Assign each *{reject, female}* or *{accept, male}* example a score *c* equal to the increase in p%-rule divided by the reduction in accuracy
- Flip examples in descending order until the desired p%-rule is achieved

link to paper: https://arxiv.org/pdf/1711.07076

So what?

(exact algorithm from paper)

1. Assign each example with $\{\tilde{y}_i = 0, z_i = b\}$ or $\{\tilde{y}_i = 1, z_i = a\}$, a score c_i equal to the reduction in the p-gap divided by the reduction in accuracy:
   (a) For each example in group a with initial $\hat{y}_i = 1$, $c_i = \frac{p}{100 n_a (2\hat{p}_i - 1)}$.
   (b) For each example in group b with initial $\hat{y}_i = 0$, $c_i = \frac{1}{n_b (1 - 2\hat{p}_i)}$.
2. Flip examples in descending order according to this score until the desired CV-score is reached.

link to paper: https://arxiv.org/pdf/1711.07076

So what?

(in case you got lost)

Nothing remotely deep is going on here: of course (?), if we're willing to look at the sensitive attribute, we can make the fairest decisions

Again, the issue is that we're trying to avoid doing something illegal!

But note (less obvious): ***no DLP can do better***

link to paper: https://arxiv.org/pdf/1711.07076

(More about) affirmative action

Take-homes from this paper:

- (Potentially troubling?) reminder that in some sense, fairness interventions seem intended to mimic affirmative action while getting around a legal constraint
- Are fairness interventions *ever* preferable in settings where affirmative action is allowed? (this paper argues that they are not)

link to paper: https://arxiv.org/pdf/1711.07076

(More about) affirmative action

Food for thought:

To what extent are fairness interventions just a “loophole”? If affirmative action is illegal, which of these interventions should also be illegal?

Assuming that using the sensitive attribute is illegal in the decision-making process, should it be legal to:

- Implement a fairness regularizer that involves the attribute?
- Implement a “balanced” learning objective “as if” we had equal proportions of each group?
- Etc. (e.g. any other interventions from this module?)

link to paper: https://arxiv.org/pdf/1711.07076

(More about) affirmative action

Finally...

Remember that “affirmative action" is *just fine* in many settings – we're not always dealing with race/gender and job applicants; if we're just trying to recommend movies on *Netflix* and want more diversity in terms certain groups (for e.g.), there's no reason why you shouldn't use an "affirmative action"-style policy

(or more simply: when legally allowable, post-processing seems always preferable to in-processing!)

link to paper: https://arxiv.org/pdf/1711.07076

Fairness and bias interventions

Case study: AI-moderated decision-making: capturing and balancing anchoring bias in sequential decision tasks

AI-moderated decision making

Why discuss this paper?

- Look at a real-world scenario where a particular type of bias impedes decisions **but which is a bit different from any of the standard classification settings seen so far (also, not every paper is about gender!)**
- Give a sense of how to design interventions for a new setting

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

This paper looks at **anchoring bias:** here, users' decisions are *anchored* by evidence they've seen recently, e.g.:

- If you see a $1,000 t-shirt, and then a $100 t-shirt, you may perceive the second as inexpensive (or, you see a higher old price on a "sale" item)
- In human-AI decision-making, showing the AI's output to the user may bias the user (see e.g. COMPAS case study)
- Users' opinions may be biased by seeing the reviews already left by other users
- (etc.)

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

This paper specifically studies **self-anchoring:** here, *a user's decisions might be biased by their own recent decisions*

Consider the following scenarios: a human evaluator is reviewing dozens of files for admission to a competitive graduate program:

A) They see five strong files (for which they recommend acceptance) *followed by a borderline file*
B) They see five weak files (for which they recommend rejection) *followed by a borderline file*

Are they likely to make the same decision in either case?

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

(Real-world) motivating example

- Collect all of the (CSE-MS) admissions data from UCSD (**note:** we got IRB approval!)
- Data contains a reviewer decision (“admit” / “deny”), as well as a (potentially less biased?) final admissions decision made by the admissions committee
- We also know the *order* in which files were reviewed: reviewers tend to review lots of files in a single session
- Anonymize the data and extract a feature vector (GPA, school, letter percentiles, etc.)
- Train a (SVM) classifier to predict reviewer decisions

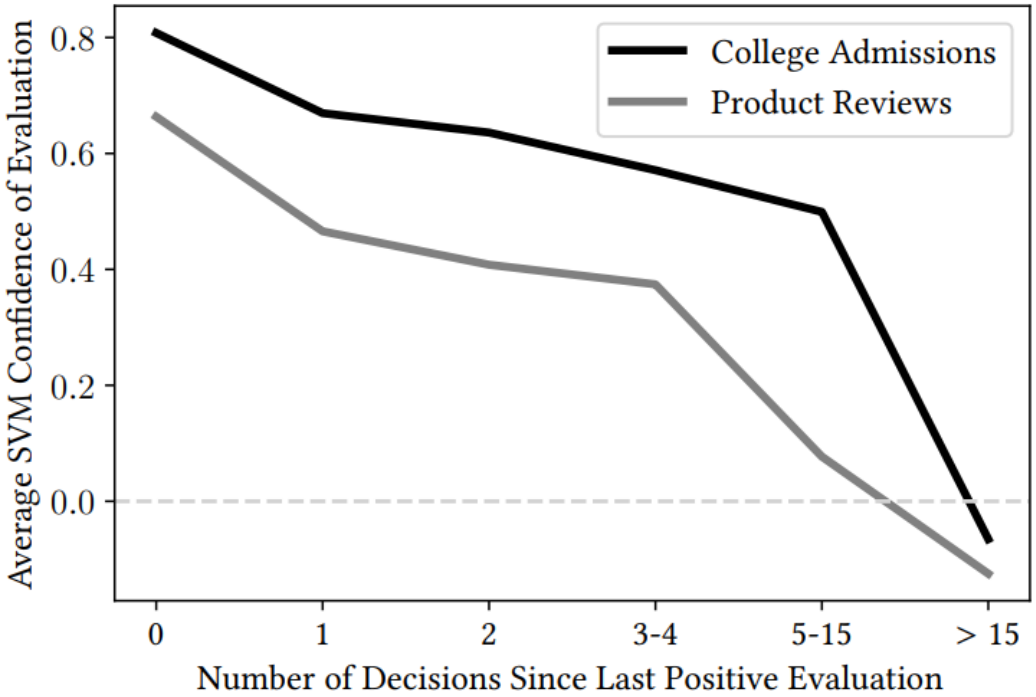
Note: the classifier *cannot suffer from (this type of) bias*, since it makes decisions *independently!* So, how do its decisions compare to human decisions?

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making



link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

What happened?

- The longer it has been since a reviewer admitted a student (i.e., the number of consecutive rejections), the more likely they become to accept the next student they see (e.g. if you haven't admitted the last 10 students you reviewed, you're likely to admit the 11th student *even if they are very weak*)
- If your file happens to come after a bunch of weak competitors, lucky you!
- If your file comes after a lot of strong students, bad luck!

(**Note:** plot also shows a similar experiment conducted on Amazon Mechanical Turk)

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

What does this tell us?

- **Humans** struggle to make lots of independent decisions *in sequence:* their recent decisions (and possibly their associated features) bias their decisions
- **Models** can easily make independent decisions in sequence, but may be inaccurate

Somewhat similar to “Learning to Defer” paper: humans are accurate but biased, versus models that are biased but inaccurate

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

What should we do about it?

- In some sense, the bias arises as a function of the *order in which files are shown to the user*. E.g. if a borderline file follows several stellar applicants, that might be a something we could flag automatically as a case where a decision is likely to be biased (compare to $1000/$100 t-shirt example)

Perhaps we can design algorithms that perturb the order in which files are shown to reviewers. *Can we do so in a way that will minimize bias?*

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Perhaps we can design algorithms that perturb the order in which files are shown to reviewers. *Can we do so in a way that will minimize bias?*

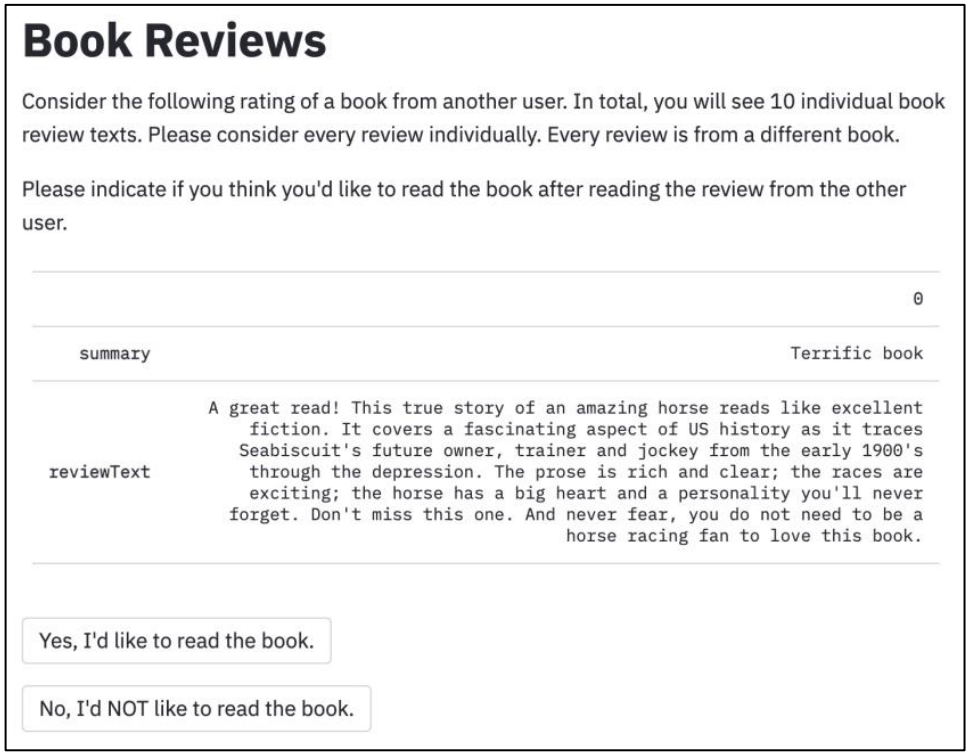
Three possible strategies:

1. **Static (or "non-adaptive"):** use the features associated with the files to design a (predetermined) ordering that will lead to less-biased decisions (compared to e.g. a random ordering)
2. **Co-operative (or "retrospective"):** use a model to estimate whether a decision was biased, and (sometimes) adjust the decision
3. **Dynamic (or "adaptive"):** every time a user makes a decision, choose what file they should be shown next based on the decision they just made

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Note: cannot actually deploy this to make decisions about MS applicants (don't have IRB approval for this!) so set up a similar experiment on Mechanical Turk in which users are shown text from book reviews and have to render binary decisions in sequence



link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Strategies – non-adaptive:

Random: surface files in a random order (or just “no intervention” in the case of college admissions)

Heuristic: alternate between “strong” and “weak” files (as predicted by a pre-trained classifier) when showing them to the user)

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Strategies – cooperative:

Probabilistic Adaptation (PA): Let the human make their (potentially biased) decisions as normal, but simultaneouly model the amount of “bias” in their decision (i.e., are they likely to have been “anchored” by previous decisions)

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Strategies – adaptive:

1. Build a model to estimate the user's "anchoring state" at each step; basically, has to estimate the human's decision using a combination of item features plus a latent "anchoring" valiable; the model is based on an LSTM
2. Use a reinforcement learning framework to (learn how to) surface files to users in an order that will minimize the value of this anchoring state

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

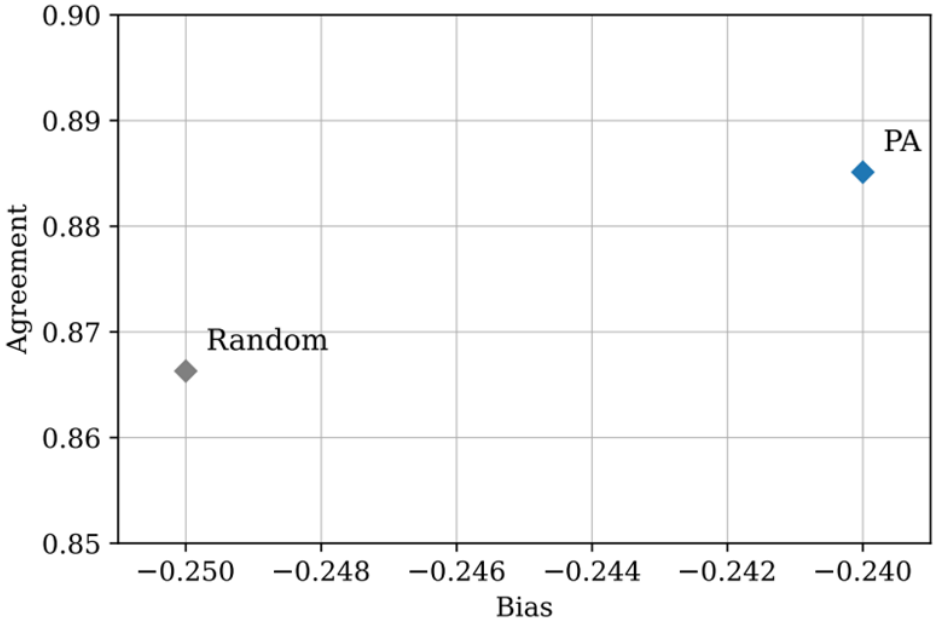
Finally, evaluate the performance of the various models:

Bias: how much are model decisions correlated with previous decisions?

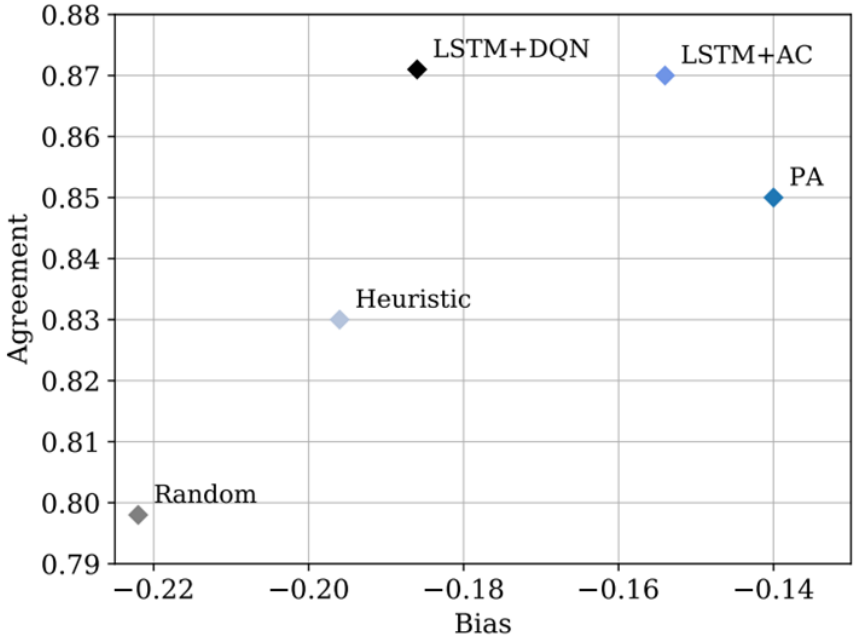
Alignment: how well do model decisions agree with human decisions?

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making



MS admissions



product reviews

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Findings:

- There isn't really a trade-off between bias and accuracy! Reviewer decisions can be made less biased (i.e., less correlated with previous decisions) ***and*** more aligned with committee outcomes (i.e., final admissions decisions)
- Even trivial heuristics substantially beat random ordering

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Food for thought: You (probably?) wouldn't want ML models deciding the outcome of your application (in exchange for your $100 admissions fee).

Do you see any possible concerns with these types of interventions, given that decisions are still ultimately made by humans?

Some "admit" decisions would be changed to "deny" as a result of this intervention; do they have any specific characteristics?

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Main message: *the order in which users are shown items* influences their decisions. This can lead to bad outcomes! But can be corrected by changing the order in which items are shown.

What else could we do with this?

- An adversary could perturb the order of files to try and get a weak student admitted, or prevent a good student from being admitted, even if reviewers themselves are not adversarial
- In a different context, this might be (somewhat) less adversarial, e.g. an e-Commerce site might perturb item ordering to promote certain items

link to paper: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf

AI-moderated decision making

Follow-up paper:

- Users are browsing items sequentially, e.g. choosing a movie to watch on *Netflix*
- Many sessions result in users never choosing any item!
- By perturbing the order in which items are shown to a user, can we make them *quickly* decide upon an item they like?

link to paper: https://dl.acm.org/doi/fullHtml/10.1145/3627043.3659569

AI-moderated decision making

Lots of potential strategies!

Though roughly the same as with fairness interventions:

- Use a heuristic to select item ordering; or
- Learn a strategy to dynamically reorder items

Acronym	Strategy	Explanation
Random-SGB	Heuristic	Equal number of popular and unpopular items shown in a sequence
Random-MGTB	Heuristic	More popular than unpopular items shown in a sequence
Random-MBTG	Heuristic	More unpopular than popular items shown in a sequence
Random-All	Heuristic	Any random sequence, independent of the number of (un)popular items
SVD	Item-Similarity	Shows similar items based on best user rating (rating ≥ 6)
DMN-Trim	RL	Learns which ordering of items leads to short decision sequences
DMN	RL + Item-Similarity	Learns which ordering of items leads to short decision sequences while including item similarity information

link to paper: https://dl.acm.org/doi/fullHtml/10.1145/3627043.3659569

AI-moderated decision making

Worth making the broad point: for “conventional” fairness problems (e.g. gender bias in hiring), there seem to be fundamental trade-offs between fairness objectives, where historical differences in outcomes can’t simply be “brushed away” by interventions

But plenty of settings don’t have this issue! There are plenty of “biases” that go beyond sensitive attributes, and plenty of tasks other than training classifiers; for some, lowering bias *and* improving performance can be achieved simultaneously

(although, what *was* the fairness issue in this case; how precisely would you characterize it?)

link to paper: https://dl.acm.org/doi/fullHtml/10.1145/3627043.3659569

References for Module 3

- Fairness & Algorithmic Decision Making: https://afraenkel.github.io/fairness-book/
- A Survey on Bias and Fairness in Machine Learning: https://dl.acm.org/doi/pdf/10.1145/3457607
- Fairness in Machine Learning: https://arxiv.org/pdf/2010.04053
- Fairness without Harm: Decoupled Classifiers with Preference Guarantees: https://proceedings.mlr.press/v97/ustun19a/ustun19a.pdf
- Data Pre-Processing Techniques for Classification without Discrimination: https://link.springer.com/article/10.1007/s10115-011-0463-8
- Mechanisms for Fair Classification: https://arxiv.org/pdf/1507.05259
- A Convex Framework for Fair Regression: https://www.fatml.org/media/documents/convex_framework_for_fair_regression.pdf
- Fair, Accountable, and Transparent (FAccT) Deep Learning: https://hci.stanford.edu/courses/cs335/2020/sp/
- Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer: https://arxiv.org/abs/1711.06664
- Does mitigating ML's impact disparity require treatment disparity? https://arxiv.org/pdf/1711.07076
- AI-moderated decision making: https://cseweb.ucsd.edu/~jmcauley/pdfs/chi22.pdf