Fairness, bias, and transparency in Machine Learning

Course outline

What are Fairness, Bias, and Transparency?

Some good sources to get a sense of the types of things that will be covered in this class:

- 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
- (lots of others in slides)

What are Fairness, Bias, and Transparency?

And some examples of similar courses from other universities:

- https://interpretable-ml-class.github.io/ (Harvard, interpretability)
- https://fairmlclass.github.io/ (Berkeley)
- https://hci.stanford.edu/courses/cs335/2020/sp/schedule.html (Stanford)

Example: who should get a ventilator?

- People with severe cases of coronavirus might require being on a ventilator for weeks at a time
- During the height of the pandemic, many hospitals filled beyond capacity and experienced a shortage of ventilators

Q: who should get a ventilator?

from https://afraenkel.github.io/fairness-book/content/02-frameworks.html

Example: who should get a ventilator?

A: give ventilators to those with the highest chance of recovery

Any potential objections?

- Those with “highest chance of recovery” might be those who’d already have the highest chance of recovery *without* a ventilator
- “Chance of recovery” is probably correlated with demographic attributes, or wealth (e.g. access to healthcare)

from https://afraenkel.github.io/fairness-book/content/02-frameworks.html

Example: who should get a ventilator?

A: give ventilators to those with the highest chance of recovery

Any other suggestions?

- First-come first-serve?
- Should some ventilators be preserved for future patients?
- Random allocation? (everyone has an equal chance!)

from https://afraenkel.github.io/fairness-book/content/02-frameworks.html

Example: mortgages

“Algorithms” are just sets of rules used to calculate some output (whether machine learning or otherwise)

As far back as 1934 (predating computers!), the Federal Housing Administration (FHA) created maps of cities that classified neighborhoods by “estimated riskiness” of mortgage loans and distributed them to lending organizations as guidelines for making loaning decisions

from https://afraenkel.github.io/fairness-book/content/01-introduction.html

Example: mortgages

Are such algorithms “fair”?

- Are the “riskiness” scores accurate in the first place, or are there systematic errors?
- Even if they are accurate, are they actually measuring historical prejudice against certain groups?
- Even if they are accurate, which groups will benefit (or be harmed) by algorithmic decisions?
- What sort of “feedback loops” will occur as decisions from such a system are used to make future measurements?
- Do potential accuracy gains of such “evidence-based” systems outweigh any fairness concerns?

from https://afraenkel.github.io/fairness-book/content/01-introduction.html

Course outline

This course explores questions like these *in the context of machine learning*:

- How might the the way our **datasets** are collected reflect historical biases?
- How might **algorithmic choices** (e.g. the choice of a certain loss function or error metric) disproportionately impact certain groups?
- What can we **do about it** to make “fairer” algorithms?
- What **should** we do about it (or what is even **legal** to do about it)?
- (later) How can we understand or **interpret** model decisions in order to understand or audit their behavior (and how does this relate to fairness)?

Course outline

The course is broken into 5 “modules”

- **Module 1:** Regression and classification (mostly revision, but with a focus on fairness-specific examples and model interpretation)
- **Module 2:** Intro to bias and fairness
- **Module 3:** Bias and fairness interventions
- **Module 4:** Fairness and bias in application domains
- **Module 5:** Interpretable and explainable AI

(more detailed descriptions to follow)

Assessment

(in case of any inconsistency:
ask on Piazza; **trust the course webpage and dates posted on gradescope**)

- Homework: **50%**
  - Each module (roughly) will be associated with one homework assignment, worth 10% of your final grade. Your lowest homework grade will be dropped (or you can skip one); most people will probably skip the last. Homework assignments are due in **weeks 2, 4, 6, 8, and 10**
- ~~Midterm: **20%**~~
  - ~~**Week 6;** covering content from Modules 1-3~~
  - **Not planning to run a midterm while this is a topics course**
- First assignment: **25%**
  - **Week 7;** implementation focused, focused on various fairness interventions
- Second assignment: **25%**
  - **Week 10; presenting** a report of bias of bias/fairness/explainability issues using datasets and models of your choice
  - Groups can be 1-4 students
  - Presentations might be recorded, or presented in weeks 9-10 (for bonus marks), depending on schedule

Assessment

- **Homework** will be **autograded:** you will submit your code and generated outputs and receive grades immediately; you can submit an unlimited number of times
- The **midterm** will be a **take-home**, 12 hour format, mostly to accommodate remote students; the submission process will be the same as for the homework
- The **first assignment** will be mostly graded based on your ability to pass performance thresholds for the given task; a *tiny fraction* of your grade will be based on performance relevant to your peers
- The **second assignment** will be peer graded

Assessment

- **Homework** is meant to be fairly “easy” and just to *make sure everyone is staying on top of the material*
- The **midterm** is intended to make sure everyone has properly *synthesized the course materials* (up to that point); that being said it’s fairly similar in format to homeworks, subject to a time constraint
- The **first assignment** is intended to test your ability to *just get something practical working* (regardless of whether your solution is elegant or not!)
- The **second assignment** is intended to test your ability to *apply what you’ve learned creatively*

Expected knowledge

- **Basic** data processing
  - Text manipulation: count instances of a word in a string, remove punctuation, etc.
  - Process formatted data, e.g. JSON, html, CSV files etc.
  - Install and run libraries to process structured data formats
- **Basic** mathematics
  - Some linear algebra
  - Some optimization
  - Some statistics

The expectation is not that everyone comes in knowing all these things, but that if you don't know them already, you will self-study them; links/resources will be made available

Expected knowledge

No prior experience in machine learning is required

But! Many people in the class (probably?) have some ML background, so I don't want the revision to go too slow for them.

This is (for now) a "topics" class, meaning the material is somewhat advanced, though the assessment is relatively lightweight.

Course outline in detail

Module 1: Regression and classification (~2 weeks)

- **Linear regression**
- **Linear classification**
- Some **feature engineering**
- Brief exploration of **more "interpretable" classifiers**
- Introduction to the notion of **sensitive attributes**
- Informal introduction to **bias** and **bias-reducing interventions**
- **Case studies:** Recidivism prediction

*If you've taken other ML classes (especially mine!) this will be mostly revision, but will at least revisit these topics **with bias and fairness in mind** – so try to work through the examples and exercises even if the methods are not new*

Course outline in detail

Module 2: Intro to bias and fairness (~1 week)

- Definitions and examples of different types of **bias**
- Study of sources of bias: **datasets, algorithms,** and **users**
- Definitions and examples of different notions of **fairness**
- Notions of sensitive attributes, protected groups
- **Case study:** Are fairness goals mutually compatible?

Course outline in detail

Module 3: Fairness and bias interventions (~1.5 weeks)

- What is desired from a fairness intervention?
- Algorithmic debiasing:
  - **Pre-processing:** how can we make algorithms less biased by modifying my **data?**
  - **In-processing:** how can we make algorithms less biased by modifying the **training process?**
  - **Post-processing:** how can we make algorithms less biased by modifying the **outputs?**
- **Case studies:** Exploring limits of fairness interventions, and mitigating bias in decision-making tasks

Course outline in detail

Module 4: Fairness and bias in application domains (~2 weeks)

- Brief introduction to bias in **language models**
- Exploration of bias in **word embeddings** (mostly a case study)
- Bias in **retrieval and recommendation**:
  - Content **diversity** and **filter bubbles**
  - **Concentration** effects
- **Case study:** bias in conversational recommenders

Course outline in detail

Midterm: Week 7 (Tuesday; skip Thursday lecture)

- Modules 1-3
- (Module 4, while still on related topics, won't be included as I'm uncertain how much we'll be through by then)

Course outline in detail

Module 5: Interpretable and explainable AI (~2 weeks)

- Discussion of relationship between **bias/fairness** and **interpretability**
- Discussion of desiderata for interprepretable models
- Explainability techniques for **linear models**
- **Sparse models** and **variable selection** techniques
- **Evaluation** of explainability techniques
- Explainability in **language models**
- **Case study:** the mythos of model interpretability

Fairness, bias, and transparency in Machine Learning

Course introduction

Why might humans be unfair?

(e.g. when evaluating job candidates)

Why might humans be unfair?

- Explicit bias against certain groups (race, gender, religion, sexuality)
- *Implicit* bias against certain groups (e.g. people may positively evaluate taller candidates, even though they likely don't consider that feature consciously)
- **Lots** of other forms of implicit bias besides obvious ones: e.g. humans may evaluate candidates differently due to social pressure, "anchoring", or because they are hungry
- Inability to consider large amounts of evidence (compared to statistical methods)
- Different value functions from different groups of people that are hard to reconcile
- etc.

Why might algorithms be unfair?

Do *algorithms* generate unfair outcomes for different reasons?

Why might algorithms be unfair?

- Reproducing human biases already reflected in datasets
- Simplifying assumptions in models that lead to systematic biases
- *Amplifying* those datasets via feedback loops
- Objective functions that implicitly focus on the “head” of the distribution (i.e., the minority population)
- Error metrics that implicitly make some types of errors “cheaper,” causing those errors to be concentrated on particular groups
- etc.

Some (recent) examples

“Linkedin profile picture of X professor” (x in CS, philosophy, chemistry, biology)

- Results are predominantly white males (all with glasses)
- Even if this is the “mode” of the distribution, the results do not seem to capture the shape of the distribution accurately

example from https://danluu.com/ai-bias/

Some examples

“Linkedin profile picture of X professor” (x in veterinary science, nursing, gender studies, Chinese history)

- When non-white people are represented, they are often reduced to stereotypes

example from https://danluu.com/ai-bias/

Some examples

“1943 German soldiers”

- Intervening to enforce diversity in the results is (probably) not the right solution (in this specific context!)

example from https://www.theguardian.com/technology/2024/feb/22/google-pauses-ai-generated-images-of-people-after-ethnicity-criticism

Some examples

"Diverse" beer recommendations (from a later module, and from my other class):

Low diversity	Medium diversity	High diversity
Founders KBS (Kentucky Breakfast Stout)	Founders KBS (Kentucky Breakfast Stout)	Founders KBS (Kentucky Breakfast Stout)
Two Hearted Ale	Samuel Smith's Nut Brown Ale	Samuel Smith's Nut Brown Ale
Bell's Hopslam Ale	Two Hearted Ale	Salvator Doppel Bock
Pliny The Elder	Bell's Hopslam Ale	Oil Of Aphrodite - Rum Barrel Aged
Samuel Smith's Oatmeal Stout	Kolsch	Great Lakes Grassroots Ale
Blind Pig IPA	Drax Beer	Blue Dot Double India Pale Ale
Stone Ruination IPA	A Little Sumpin' Extra! Ale	Calistoga Wheat
Schneider Aventinus	Odell Cutthroat Porter	Dogwood Decadent Ale
The Abyss	Miner's Daughter Oatmeal Stout	Traquair Jacobite
Northern Hemisphere Harvest Wet Hop Ale	Rare Bourbon County Stout	Cantillon Gueuze 100% Lambic

Food for thought

Humans are sometimes prevented from using sensitive attributes when making decisions, e.g. race-blind college admissions

You're entitled to your own opinion as to whether this is a good policy, but the basic idea is that simply forbidding the use of the sensitive attribute in decision-making will lead to "fairer" or somehow "less biased" outcomes

Should algorithms also be forbidden from looking at sensitive attributes?

Food for thought

Algorithms exhibit biases, but so do people!

To what standard should algorithms be held?

It is enough that they're "less biased" or "less unfair" than their human counterparts, or should they be held to a higher standard?

Food for thought

Generally, algorithms are trained to be as “accurate” as possible

If we modify an algorithm (or dataset etc.) to make it fairer, this will (usually) come at the expense of accuracy

Who will pay for this loss in accuracy?

How much of a tradeoff is “acceptable”? Are our notions of accuracy the right ones in the first place

Food for thought

This course will mostly explore algorithms and methodology (i.e., how to *measure* and *correct* bias), but we'll also try to think critically about questions like those above, e.g.:

- What are the real-world harms that algorithms actually cause?
- Are these harms fundamentally resolvable?
- What incentives are there to correct them?
- How do definitions from machine learning differ from societal values

Food for thought

But rest assured that assessment will be methodological stuff and ***will not require you to have the "right" opinions about anything!***

Interpretability

What about **interpretability?** What does it mean for a model to be interpretable?

Interpretability

What about **interpretability?** What does it mean for a model to be interpretable?

- A user should be able to understand the model's "reasoning" process
- Every parameter of the model should correspond to a specific, meaningful concept
- The user of a model should know what inputs would need to change to get a different output
- The model should be able to provide examples or supporting evidence for its predictions
- The model should be able to explain its predictions using natural language, much like a human can explain their reasoning

(**note:** these goals are certainly not mutually compatible!)

Interpretability

What does interpretability have to do with fairness?

- An important part of assessing model bias and fairness consists of reasoning about how those models make decisions
- Understanding a model's decision-making process helps users to trust the model and helps developers to improve the model
- Interpreting models' decision-making processes is important even for "black box" models!

Questions

That’s it!

Any questions about the course before starting with actual material?