CS 763: Security and Privacy in Data Science **Fall 2019**

Presentation 8: Defining fairness: challenges and Causal Fairness

Presenters: Justin Hsu *Scribes: Zelin (Bobby) Lv*

8.1 Fairness (more)

- Individual fairness (Fairness through awareness)
- Group/ Subgroup fairness (Equal Opportunity)

  For Group/ Subgroup fairness setting, we have instance as a tuple, (A, X, Y, R), where A is the protected **attribute**, such as gender or race and A can have multiple feature. X is feature, Y is ground truth and R is prediction. R is typically defined to be a function of A and X.

  And we have talked different notations of fairness:

  - Independence: $A \perp\!\!\!\perp R$
  - Separation: $A \perp\!\!\!\perp R \mid Y$
  - Sufficiency: $A \perp\!\!\!\perp Y \mid R$

  And we also talked about some impossibility result that if $A \not\perp\!\!\!\perp Y$, then only one of these properties can hold.

8.2 50 Years of Test (Un)fairness: Lessons for Machine Learning

This paper gives [2] history of fairness research from different fields. To have more information about fairness, there is a big conference on fairness, called FAT * (ACM Conference on Fairness, Accountability, and Transparency). And this paper shows that new some results may have been discovered fifty years ago. And to know what have done can direct the future path of fairness research.

8.2.1 Test Fairness

The basic idea originates around 60s or 70s.

- Standardized testing (SAT, GRE...)
- Question: are these tests fair? (and what does that mean to be fair?)

We have a lot of tests given to students, scores and demographic information, including people's race, gender, their income and where they live or whatever. So we want to decide if these tests are fair for students or not.

8.2.2 1960s

In 1960s, people gave some interesting ideas on fairness:

- Guion (1966): "people with equal probability of success should have equal probability of being hired"
- Cleary (1966): "a subgroup does not have consistent error"

There are more ideas on the paper but these two are the essences. We can see the first idea is very similar to the individual fairness we have seen, and it doesn't tell us how to measure people's probability of success. And the second idea tells us that if a model, such as linear regression, has consistent error on one subgroup, this model underestimate one subgroup in some sense that can be formalized and it's not fair for that group.

8.2.3 1970s

In 70s, there was more research on test fairness

- Thorndike (1971):
  - Equalize:
    $$\frac{\#\text{ of predicted positives}}{\#\text{ of true positives}}$$
    for $A = 1, -1$. This is very similar to the *Equalize Odds* we saw.
  - Achieve this by setting different thresholds.
    So for each group, we may have a different predictor, and for each of them we can have a different threshold in order to achieve fairness, like achieve the equalization.
- Darlington (1971):
  - Unifies Cleary and Thorndike definitions
  - Plus 2 more definitions
  - Out of 4, can only have 1.
    The last one is very similar to the impossibility result we have seen last week.

8.2.4 1976: Special edition of Journal of Educational Measurement

This was one peak of research on test fairness.

- Peterson & Novick: Equalized odds + ...
- Progress stopped after this special edition
  - No single definition is good
  - No definition captured intuitive fairness

In order to solve these issues, people tried different things in 1980s, which also happen today.

8.2.5 1980s

- Try to incorporate ethics and sociology
- Formalize fairness in policy and legal system

This paper servers as a good survey on the history of fairness. By looking back we can see there are some interesting definition, people interesting in this topic may read this paper to see how things can work today, and on a machine learning setting.

8.3 Possible interesting directions

- Non-comparative definitions: Is process fair for $A = 1$?

  This means that we don't compare $A = 1$ with $A = -1$. Instead, for each group we define its notion of fairness.

- Focus on linear regression, and correlation not independence

  All the three definition we have used are based on Independence instead of correlation. But there are several advantages of correlation definition:

  - Correlations are easier to estimate from data

- Fair model vs. Fair use of model

  The difference between these two definitions is the first one focuses on the model prediction of different group and the second one on how we use model.

- Compromise definitions

  We have three definitions of fairness: independence, separation and sufficiency. Independence has problem of compatible but the other twos are compatible with each other. So the question is that whether we can have a definition that is not either separation nor sufficiency but somewhat in between to achieve fairness. This seems to be possible.

8.3.1 Compromise definition

Binary classifier is (λ_1, λ_2)-Thorndike fair if

$$1 : \frac{(\#\text{ True pos}) + \lambda_1(\#\text{ false pos})}{(\#\text{ True pos}) + \lambda_2(\#\text{ false pos})}, \text{ equal across groups}$$

$$1 : \frac{(\#\text{ True neg}) + \lambda_1(\#\text{ false neg})}{(\#\text{ True neg}) + \lambda_2(\#\text{ false neg})}, \text{ equal across groups}$$

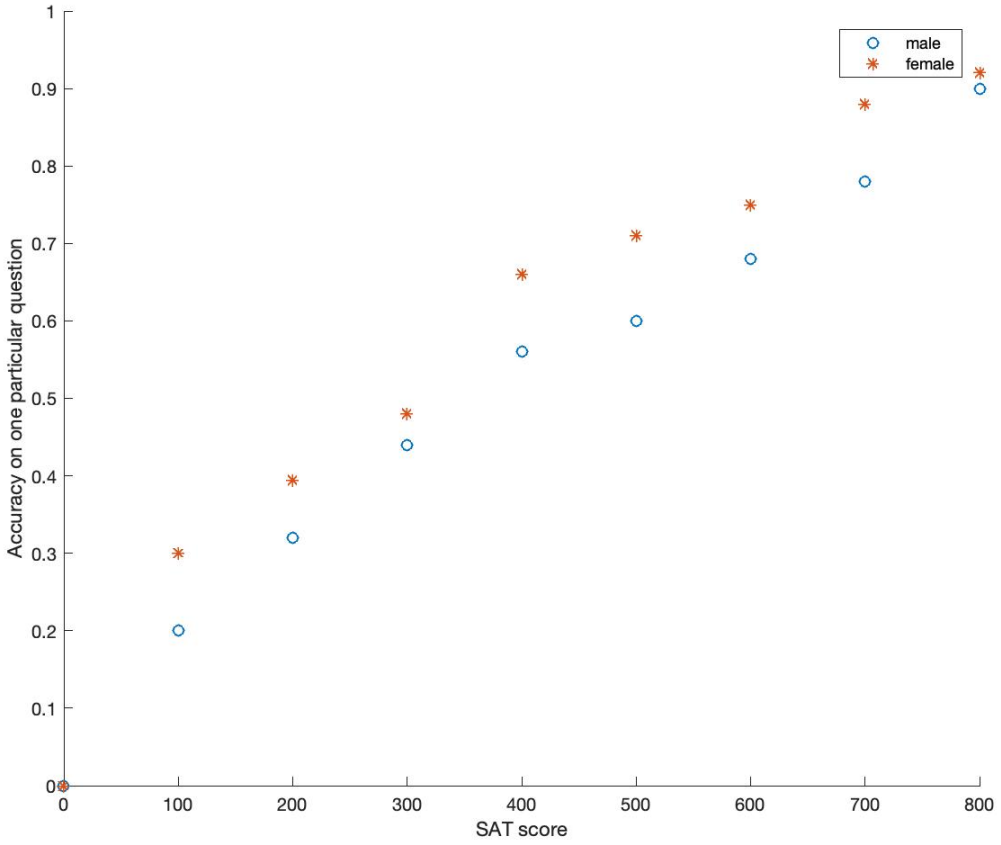
There few special cases of (λ_1, λ_2)-Thorndike fairness definition:

- $(1, 0)$-Thorndike is sufficiency
- $(0, 1)$-Thorndike is separation

8.3.2 Test fairness: Differential Item Function

This is not very popular to Machine Learning people today, because it's very test case specific.

- Evaluate fairness of particular questions



From the graph we can see that different group may have different performance on one particular question, so we might see there is something wrong with this particular question.

That's all on the paper of history of fairness included in this lecture. The rest of this lecture note is for the background on the casual stuffs.

8.4 Causality and Fairness

Causality is the body of research that formalizes one thing causes another thing. And there is a lot of ways we can use causality on fairness. More details can be found on the fourth chapter of [1].

- So far we assumed dataset (A, X, Y, R)
- "Observational" notion of fairness

This kind of fairness is that we only use the dataset distribution. But people showed that on the same data pattern, depended on how to interpret it, the result can be either fair or not fair.

- Many "intuitive" notions of fairness can't be captured just given these tuples (A, X, Y, R)

8.4.1 Example: UC Berkeley admissions

There are three pieces of dataset: A for gender, which is the protected attribute, Z for student choice, which is a department a student applies to and Y for admission decision. And we try see whether it's fair or not.

The investigation shows that

- Admissions rate for males was much higher
- In each department $z \in Z$,

$$\frac{\#\text{admitted to } z}{\#\text{applied to } z} \text{ is higher for females than males}$$

We can see that in the case of UC Berkeley admission, the case is reversed. If we look at all the data, we can see the admission is biased toward male students but for each department, it is biased toward female students. So just given the dataset, we can't say whether it's fair or not. This is "Simpon's paradox". This is kind of the starting point of using causality in fairness.

8.4.2 Causality

The main idea of causality is that we are going to have more assumption on our data, not just a distribution.

- Assume data generation process along with data

  Here are few examples of this assumption:

  1. For binary attributes $X = \textit{exercise}$, $X = \textit{overweight}$ and $H = \textit{heart disease}$. And on the top of these features, we can make an assumption of how they are generated, which we can interpret as a program.

$$X = B(\frac{1}{2})$$

$$W = 1 \textit{ if } X = 1 \textit{ then } 0 \textit{ else } B(\frac{1}{3})$$

$$H = 1 \textit{ if } X = 1 \textit{ then } 0 \textit{ else } B(\frac{1}{3})$$

  Where function $B(x)$ ranges on binary domain, with probability x getting a head. This is one example of data generation model. Intuitively, in this data generating model, being overweight and having heart disease are related to if having exercise or not.

  2. Here is a different possible data generation process.

$$W = B(\frac{1}{2})$$

$$X = 1 \textit{ if } W = 0 \textit{ then } B(\frac{1}{2}) \textit{ else } 0$$

$$H = 1 \text{ if } X = 1 \text{ then } 0 \text{ else } B(\frac{1}{3})$$

These two examples look similar regard how they are generated. It's possible that the joint distributions on X, W, H are the same on the two examples. But they tell different stories about causality.

Once we have the assumption about causality, we can start something like intervention.

8.4.3 Modeling interventions

Say we set $W = 1$. What happens to H?

In example 1: nothing happens to H.

$$X = B(\frac{1}{2})$$

$$W = 1$$

$$H = 1 \text{ if } X = 1 \text{ then } 0 \text{ else } B(\frac{1}{3})$$

In this model, we can see that weight change does nothing on H.

But in example 2: distribution of H changes.

$$W = 1$$

$$X = 1 \text{ if } W = 0 \text{ then } B(\frac{1}{2}) \text{ else } 0$$

$$H = 1 \text{ if } X = 1 \text{ then } 0 \text{ else } B(\frac{1}{3})$$

In this model, H changes. So we are encoding more assumption on the data.

8.4.4 Causal inference

- Given data, we can't prove or show causality
- We can only show correlation

  If two variable are correlated, we don't know their relation on causality. Say if X and Y are correlated, we don't know if X causes Y or Y causes X, maybe even there is no causation at all.

8.4.5 Structural casual models

A SCM has equations:

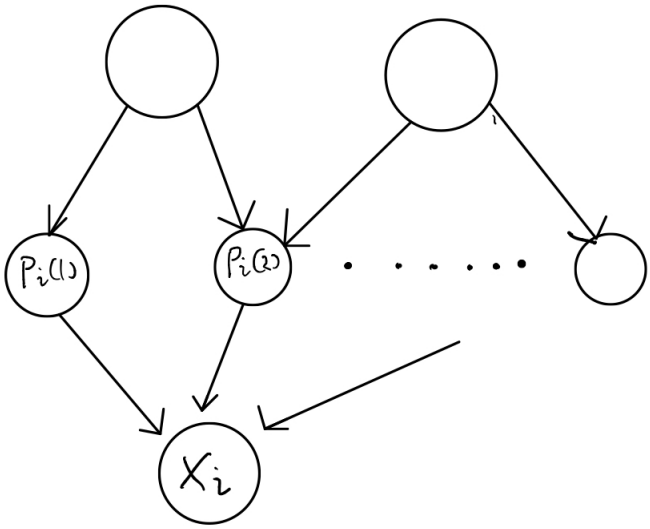
$$X_i = f_i(P_i, u_i)$$

where $P \subseteq \{X_1, X_2, ..., X_d\}$ are the parents and u_i are noise terms, assumed to be independent. And there is no 'cycles'.

This is also an assumption that sometimes we can check. Once it's assumed, we can perform various experiments.

8.4.6 Causal graphs/ graphical models

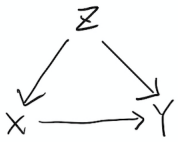
The idea is simple, given the SCM, we can draw its corresponding graph.
Example of Causal graphs:



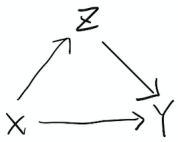
The graph is another way to represent the SCM and graph encodes which thing is dependent on the other thing.

By looking at these graphs, we can conclude several properties of how one variable is dependent on each other. There are few shapes we going to illustrate.

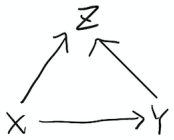
8.4.7 Examples of Graphs



This is called a 'fork'. In this graph, X and Y that we can't tell X causes Y and it may look like that X causes Y because Z causes both of them. To figure out if X causes Y, condition on Z.



This is called a 'mediator'. This is an indirect cause. For example, in UC Berkeley Admission, the gender X could cause different choice of department to apply Z and both of them influence the admission decision Y.



This is called a 'collider'. Don't condition on Z when analyzing if X causes Y.

These are the basic graphs that encode the causal information.

8.4.8 Conclusion on causal models

There are two things to be noted on causal model:

- Casual model is heavy assumption, which means it's hard to test.
- What does it mean for X to cause Y.

  The problem is hard to measure and it's close to philosophy. The measure usually used in science experiment is:

  - Randomized intervention on X results change in Y.

References

[1] Solon Barocas, Moritz Hardt, Arvind Narayanan. Fairness and machine learning *https://fairmlbook.org/*

[2] Ben Hutchinson and Margaret Mitchell. 50 Years of Test (Un)fairness: Lessons for Machine Learning FAT* 2019