2nd Reading for COMP6321 Machine Learning

Prior to the Midterm 2 you must the read and understand key paragraphs from the following recent paper:

- Kusner, Loftus, Russell, Silva (2017). **Counterfactual fairness.** *Neural Information Processing Systems (NeurIPS)* 2017. (**PDF**)

The goal if this reading is to expose you to literature on fairness in AI and how it relates to causality.

The above paper is about three things:

1. **Fairness in machine learning, in general.** The paper provides a short but good introduction to issues of making predictors that are "fair" according to different definitions of fairness (Definitions 1–4). Choosing a fairness criterion can impact, for example, a university admissions program that aims to be equitable across race and gender. The paper also discusses *example scenarios*.
2. **A fairness criterion they call "counterfactual fairness."** The authors introduce a *new* "fairness" criterion (Definition 5) based on the framework of "causal inference."
3. **A learning algorithm that respects "counterfactual fairness."** Having introduced a new notion of fairness, they proceed to describe a learning algorithm that attempts to ensure this type of "fairness" in the resulting model.

Your job is to understand **only the first** of the three topics above:

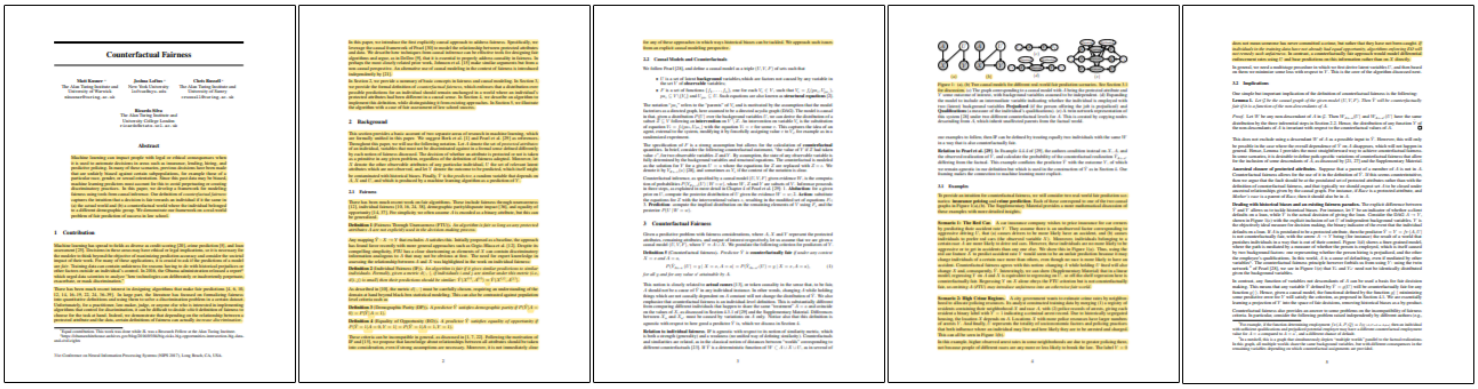
- You should understand the fairness criteria reviewed in Definitions 1–4 and the formulas *mean*. If all you do is memorize the formulas verbatim without understanding what they mean, it will not help you much.
- You should understand the concrete example scenarios discussed in the paper where fairness is a concern, and the potential accuracy trade-off of enforcing fairness.
- You should understand how postulated causal relationships can be represented as a directed graph, specifically Figure 1 (a) and (b) and why they correspond to the scenarios described in the referring text.

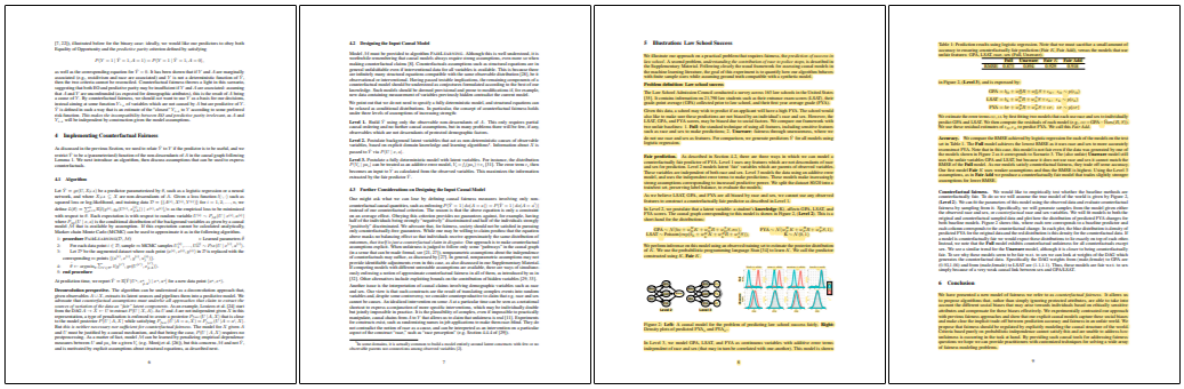
Things you do *not* need to understand:

- You do *not* need to understand causal inference on a technical level for this course. For example you do not need to know how to calculate specific probabilities, nor what the specialized notation such as $\hat{Y}_{A \leftarrow a}(U)$ means. This is beyond the scope of the course.
- You do *not* need to understand the "counterfactual fairness" criterion proposed by this paper (Defn 5).
- You do *not* need to understand the learning algorithm proposed by this paper.

Requirements

You must read and understand the sections highlighted in yellow below.





You are not expected to read cited papers.

Example midterm 2 questions related to this reading

1. Suppose you want to build a predictor $\hat{Y}$ that predicts whether a future PhD applicant would pass a comprehensive exam. You have historical training cases containing three attributes: G (their undergraduate grades), S (their biological sex), and Y (whether they passed or failed). For each of the following, give your answer in words, *without* using any probabilistic notation $P(\cdot \mid \cdot)$.
   - Describe the **fairness through unawareness** criterion applied to your predictor.
   - Describe the **individual fairness** criterion applied to your predictor.
   - Describe the **demographic parity** criterion applied to your predictor.
   - Describe the **equality of opportunity** criterion applied to your predictor.
2. Does the *fairness through unawareness* criterion guarantee that protected attributes will not influence the prediction? Explain.
3. The paper describes a "red car" scenario (fairness in insurance prices) and a "high crime" scenario (fairness in law enforcement). Why in the "red car" scenario did the postulated causal graph have no arc from X (car is red) to Y (accident rate), whereas in the "high crime" scenario the postulate causal graph *did* have an arc from X (resident location) to Y (arrest rate)?
4. In the "high crime" scenario:
   - Why did the postulated causal graph have an arc from A (race) to X (residential location)?
   - If a resident of neighbourhood X is observed to have $Y = 0$ arrests, what can we say about whether they have committed a crime?
5. In the "law school success" scenario:
   - What were the specific unprotected features and protected features used to make a prediction? List them separately.
   - What specific outcome was being predicted, and how was prediction error measured?

- Which model had lowest prediction error, and why?
- Which model had the highest prediction error, and why?
- The authors propose a "level 2" model that postulates a certain unobserved variable as having causal influence on the observable features. What was this latent variable, and how did the authors choose to model its influence on the observed features? Be precise.

Supplementary discussion, for those who are interested (optional)

The paper proposes "counterfactual fairness" (Definition 5) but it is difficult to understand without a background in causal inference. **You are not expected to understand this definition**, but for those interested in "fairness in AI" the text below tries to clarify what the paper proposes.

Assume the following general scenario:

- we wish to predict some outcome variable $\hat{Y}$ (e.g. future grades) based on observed features;
- we observe protected features A (race, gender, etc);
- we observe unprotected features X (past grades, publications, awards, etc); and
- we cannot directly observe features U (*e.g.* knowledge, intelligence), but we can postulate the form by which they may cause (or not cause) the observed features.

The paper defines "counterfactual fairness" as follows.

- Predictor $\hat{Y} = f(X, A)$ is "fair" if under every condition $X = x, A = a$ the probability of predicting any $\hat{Y} = y$ would remain unchanged had we forced the protected features to be otherwise (forced $A = a'$).

In other words, for any case $X = x, A = a$, the particular setting of A should never *cause* prediction $\hat{Y} = y$. Note that this is subtly different than the "demographic parity" (DP) criterion, which essentially says "when we observe $A = 0$ the prediction $\hat{Y}$ should be the same as for when we observe $A = 1$." The DP criterion does not try to explicitly model causation.

Counterfactual statements can seem contradictory. For example, *if we already observed $A = a$ how can we assume that we've 'intervened' to also make $A = a'$??* Yet such statements can be well defined and the probabilities of counterfactual events can be computed with respect to a postulated causal structure. For those who are interested, read Chapter 7.1 of the book below, in particular equation (7.6):

- Pearl, J. **Causality: Models, Reasoning, and Inference** (2nd Ed). Cambridge University Press, 2009.

Again, reading Pearl's book is obviously not required for this course.

Students may also find this online introduction to fairness by Ziyuan Zhong to be clarifying, although it is not necessary to refer to this blog post.