Bias and Fairness

Bias and Fairness and Social Data Science

Bias and Fairness within Machine Learning (and Data Science generally) **is unique in that it is very much a domain of the social sciences.**

Note: Machine learning is a **tool** and like any tool, it can be used for good or bad.

Important to remember: Machine learning is good at making predictions **with respect to a specific performance measure**!

- A hammer is good at hitting nails. But you need to make sure the situation requires hitting nails, not gluing or using a screwdriver or welding.

Bias and Fairness in ML

Note: The techniques we are discussing deal with **Bias and Fairness in Machine Learning model-building**. There are many sources of bias that **we won’t be equipped to deal with**.

One example case: Biased data will always lead to biased predictions. This is because we aren’t able to detect the pattern that actually exists.

Takeaway from Predictive Policing Example

The first step in addressing possible bias within Machine Learning models is to make sure that we have **a good understanding of what our data represents**.

Note: This does not mean that we must make sure it is completely unbiased and fully representative!

When to use these techniques

Remember how we defined machine learning:

A computer is said to learn from experience E with respect to some tasks T and performance measure P, if its performance in T as measured by P improves with E.

Addressing bias and fairness in ML is essentially **making sure that our performance measures match with what we actually want**.

- *Example*: We don't just want overall best precision – we want the best precision balanced by race and gender.

When to use these techniques

Bias and Fairness is where **context matters the most.**

May not always be motivated by “fairness” either.

- Balancing performance by categories can be preferable in some cases. For example, may want to make sure we are predicting well in each region of the country because the resources are spread out rather than concentrated in one area.

Context

Questions to think about when addressing bias and fairness:

- What is the **exact thing** you are trying to predict? How well does it match up with the outcome that you have?
- What is the **goal** of the prediction? What are you using the predictions for?
- Is there an **intervention**? Is it positive or negative?
- What are the **consequences** of being incorrect?

Strategies

Main strategies to consider:

- Think carefully about where the data comes from and if there are any possible sources of bias
- Compare performance measures for different groups and choose different models based on existence of bias
- Use Interpretable ML methods to understand the models better.
- Note: These are model-agnostic

“Best” Model

When building machine learning models, we typically pick the one that performs the best according to some performance metric.

- For example, we optimize on precision, so we **choose model that has the lowest precision**.

This doesn’t need to be the case!

- We can optimize on precision AND fairness by **choosing the model that has the lowest precision out of all models that are fair**.

Managing Bias

Generally, you won’t be able to complete remove any bias from models. There will always be some small differences between groups.

Goal is to mitigate bias and find the best model that has the best predictions with the least amount of bias.