HW3: Detecting and Mitigating Algorithmic Bias – Part 4: Reflection

In this project, Kyle and I explored both the technical and ethical sides of algorithmic fairness but creating and analyzing models based on the Adult Income dataset. We started with bias detection, went into model training and evaluation fairness, and finished with mitigating bias. We applied two different mitigation strategies, pre-processing resampling and in-processing.

Which Mitigation Worked Best and Why?

Both mitigation techniques edited the behavior of the model in different ways, but we have determined that resampling worked the best overall. Resampling balanced the training data so that the different possible combinations of sex/income appeared equally. This technique had the greatest effect on equal opportunity, and the equal opportunity difference improved from 0.114 to 0.041. This means that the model was more consistent at finding and recognizing actual high-income individuals.

The fairness constrained model was able to effectively equalize the position prediction rates, 0.176 to 0.007. But with this, it sacrificed recall and hurt equal opportunity, so it was less reliable at identifying the qualified individuals we are looking for.

Which fairness-accuracy trade-offs did you observe?

We were able to observe that fairness does not come for free. Both techniques for mitigation reduced our accuracy, an each method had its own sacrifices toward performance. Resampling lowered accuracy and precision, but improved recall and equal opportunity statistics. The fairness-constrained method kept precision high but reduced accurate recall and had worse equal opportunity.

The baseline model reflected biases that were in the data, when fairness constraints were inputted, the model moved away from an accuracy optimized result and moved towards the fairness goals. Improving one fairness metric harmed another. Because the data is imbalanced across groups (Demographic Parity and Equal opportunity), fully fixing one issue creates another.

How might domain knowledge influence fairness goals?

Different domains have different fairness definitions. Fairness depends on the domain in which the model is being used. For example, when diagnosing a medical issue, missing a possible positive by diagnosing a false negative is much more harmful than a false positive professionals can schedule follow up tests to confirm. In this medical setting, equal opportunity is very important because if group a and group b both have the same condition, a fair model should detect it equally well in both groups.

Domain knowledge can influence what type of errors are most meaningful. As shown, if false negatives are more harmful, then fairness techniques that have improved recall are much more preferred. If false positives are more important, then you should use techniques that enhance/protect the precision of your model. Knowing the domain in which you're working in, and knowing whats important can completely alter your fairness goals.

How would you communicate these results to non-technical stakeholders?

To communicate results well to non-technical stakeholders we would make sure that we avoid high level technology or mathematical terminology. We would first summarize what the baseline model does well, and focus on the conclusion generated rather than how it works. We would also make sure that when explaining the mitigation methods they were explained simply and in laymen terms. Stating things such as "We adjusted the training data so the model sees men and women equally often" when talking about resampling. Or "We told the model to give men and women the same rate of positive predictions." when talking about a fairness-constrained modeling technique.

We would also present results visually, as most people understand data through visuals rather than intricate explanations of our conclusions and how they were derived. Showing charts that have acceptance rates and accuracy would show the non-technical stakeholders the differences more clearly.