What will it take to generate fairness-preserving explanations?

What is a "good" explanation in fairness-relevant applications? How can we find a "good" explanation?

Jessica Dai*, Sohini Upadhyay**, Stephen H. Bach*, Himabindu Lakkaraju** | *Brown University **Harvard University | XAI Workshop at ICML 2021

MOTIVATION

Explanations are meant to be used for high-stakes decisions: ensuring regulatory compliance, informing downstream actions....

When we care about fairness, what does the data look like?
- A meaningful difference in the distribution of **features** or **labels** across (demographic) groups
- Possibly causing performance disparities across groups

Explanations shouldn't **be misleading:** should not suggest a model is fair when it is unfair, or vice versa.

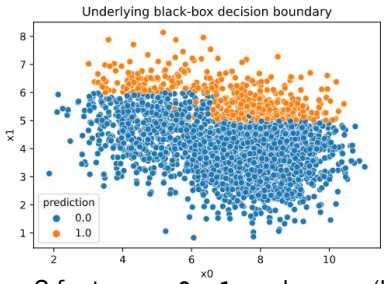
DIAGNOSING FAIRNESS MISMATCH

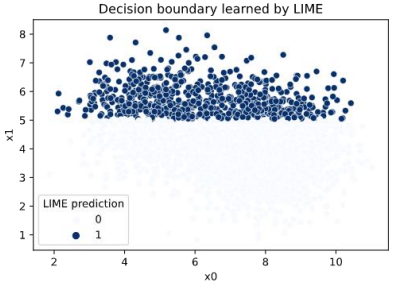
Group fairness mismatch – If the explanation provides an approximation of the decision boundary, then this approximated boundary should result in similar *group fairness metric values* as the black-box being explained.

Counterfactual fairness mismatch – if an explanation is generated for a specific datapoint, then *changing the sensitive feature value* of that datapoint should change the explanation in the same way that the black-box classification changes.

Additional open questions –
- How should the sensitive attribute feature importance be interpreted?
- How does this analysis change when evaluating *explanation methods,* vs a specific explanation for a given datapoint and model?

A SIMPLE EXAMPLE





- 3 features: `x0`, `x1`, and `group` (left or right cluster).
- Left cluster is 27% of population.
- Black-box algorithm:

```
if group == left:
        return x1 > 6
else:
        return x1 > 5
```

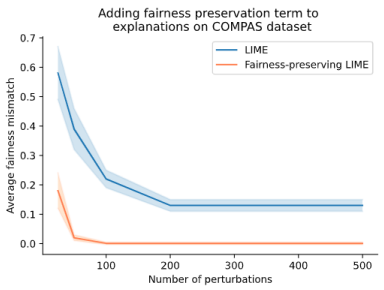
LIME's decision boundary is incorrect for both groups, but *closer* to the correct boundary for the right cluster.

Group fairness mismatch – The original black-box predicts Y=1 at equal rates; the LIME boundaries do not.

Counterfactual fairness mismatch – for a point in the left cluster with `x1=5.5`, switching group membership will change the prediction from `Y=0` to `Y=1`; the LIME explanation will not.

A PRELIMINARY APPROACH

Add an additional constraint to the LIME optimization problem, for *group fairness (demographic parity) mismatch*



Fairness mismatch constraint:

$$\psi = \left| DP(f(x)) - DP(E_f(x)) \right|$$

Original objective:

$$\xi(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Modified objective:

$$\xi_{fair}(x) = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \lambda_1 \Omega(g) + \lambda_2 \psi(f, g)$$

TAKEAWAYS

- We need more nuanced ways of evaluating explanations beyond "fidelity"
- We need to think about how to make clear what is and is not communicated about the model by a given explanation.