Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers

“Do the right thing”: machine learning and causal inference for improved decision making

Divyat Mahajan (Microsoft Research India)
Chenhao Tan (University of Colorado Boulder)
Amit Sharma (Microsoft Research India)

Explainability in ML

- Critical decision-making situations in which the predictions of a black box ML model would not be sufficient
  - Healthcare: You have a 90% risk of heart disease in the next 2 years
  - Finance: You have been denied a loan due to a high risk prediction

- Explanations should be **interpretable** and can serve a dual purpose:
  - Shed more light on the bias of the model
  - Should have some meaningful value for the user

- We focus on Local Explanations for Machine Learning Classifiers

Counterfactual Explanations

Loan Application Scenario

- Explanations based on Feature Importance
  - Fidelity-Interpretability Tradeoff
  - No Actionable Advice
- Counterfactual (CF) Explanations
  - Perturbations in the original feature that could have led to change in the prediction of the model
- CF generation generic formulation:

$$\arg\min \; Loss(\, f(x^{cf}), y') + \; Distance(x, x^{cf})$$

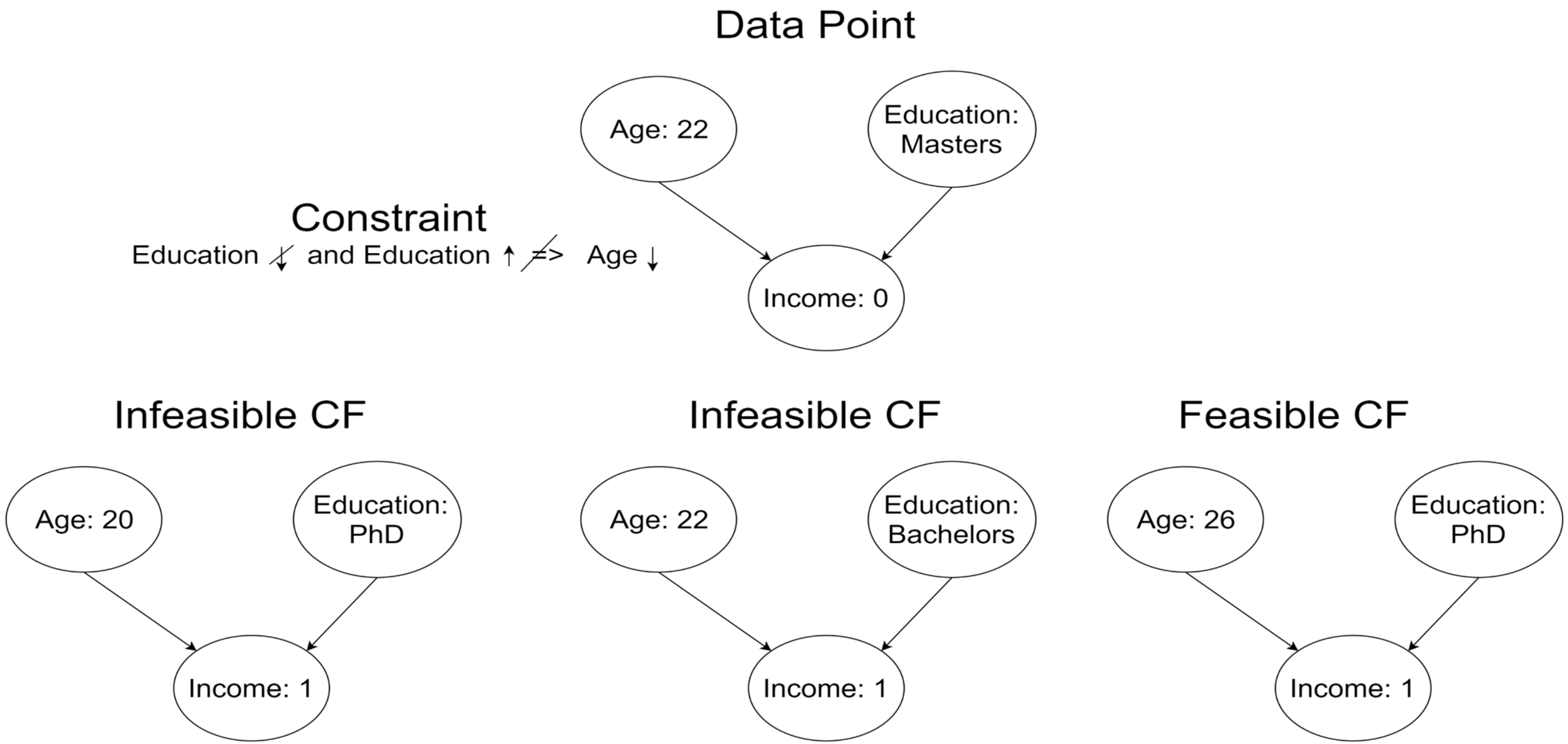
We cannot offer you loan currently
Contact us in few weeks

Most common reasons for the rejection
1. Credit Score
2. Educational Qualification

Counterfactual Explanation
You would have received the loan if your Education was PhD

Issues with Counterfactual Explanations

- Independent feature perturbation lead to infeasible CF explanations



Feasibility of Counterfactual Explanations

- Notation
  - Machine Learning Classifier f: $X -> \ Y$
  - x in X are features, y in Y is categorical output
  - (x^{cf}, y^{cf}) represents the counterfactual explanation for data point x under classifier f
- Structural Causal Models (SCM) M: $< U, V, F >$
  - U are the endogenous variables, V are the exogenous variables
  - U and V do not contain the outcome Y
- Global Feasibility of CF Explanations:
  - Validity: $y^{cf} = y'$, where y'represents the target class
  - Changes from x to x^{cf} satisfies all the constraints given by SCM M
  - Exogenous variables ${x^{cf}}_{exog}$ in U constrained within their input domain

Preserving Feasibility

- Causal Proximity Regulariser:
  - Perturbation in feature v should be causally related to perturbations in other features instead of just being proximal
  - Given the knowledge of SCM, we can preserve global feasibility with a better notion of Distance for endogenous nodes
    - $DistCausal(\, x_v, {x_v}^{cf} \,) = \, Distance(\, {x_v}^{cf}, f({x_{v_{p1}}}^{cf}, \ldots.., {x_{v_{pk}}}^{cf})\,)$
- Learning from Oracle/Expert:
  - Modelling the constraint implicitly via Oracle which provides access to feasibility score
  - Oracle may represent user/human feedback
  - Learn to mimic the Oracle using fixed number of queries (q^{cf})
    - $Orac\mathrm{leScore}$: $\mathrm{e}^{-(x^{cf} - q^{cf})^T (x^{cf} - q^{cf})}$
  - Maximise OracleScore for queries that received higher feasibility score via Oracle

Conclusion

- Poor performance of state of the art method on feasibility of CF Explanations
- Generative framework for CF Explanations
  - Computational advantage
  - Easy extensions to preserve constraints
- Visit our Poster ***#57*** to learn more