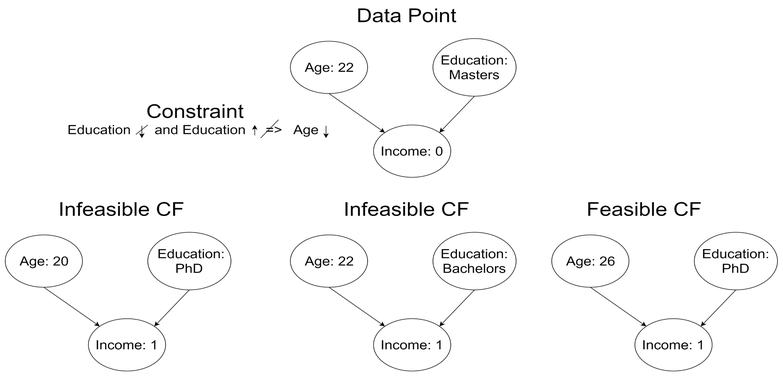
Preserving Causal Constraints in Counterfactual Explanations for Machine Learning Classifiers

Divyat Mahajan[1]; Chenhao Tan[2]; Amit Sharma[1]
[1]Microsoft Research, [2]University of Colorado Boulder

Feasibility of Counterfactual Explanations

- Counterfactual Explanations promise to provide faithful and actionable explanations for ML classifiers
- Actionability of counterfactual explanations rests on preserving certain feasibility constraints



Generative Modelling of CF Explanations

- Variational Inference based approach:
  - Encoder $q(z|x, y')$ embeds data point into the latent space
  - Decoder $p(x^{cf}|z, y')$ generates the counterfactual in class y' from the latent encoding
  - Learn the encoder and decoder by minimizing the following loss:

$$min\ \mathrm{E}_{q(z|x,y')}[Distance(x, x^{cf}) + \lambda * HingeLoss(f(x^{cf}), y', \beta)] + KL(q(z|x, y')||p(z))$$

Causal Connection to Feasibility

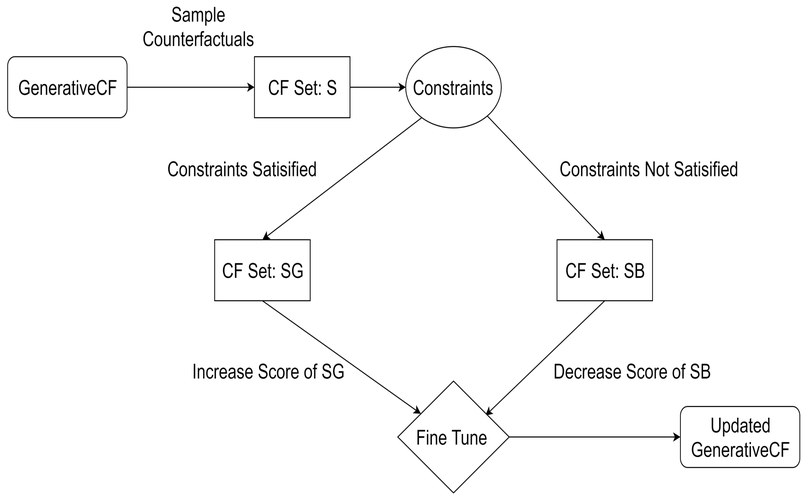
- **Global Feasibility**:
  A counterfactual explanation $< x_{cf}, y_{cf} >$ is globally feasible if it is valid ($y_{cf} = y'$) and changes from x to x_{cf} satisfies all the constraints given by the underlying causal model
- We can use the causal knowledge to define a better notion of Distance to preserve constraints (*SCMGenCF*)

$$DistCausal(x_v, {x_v}^{cf}) = Distance(x_v^{cf}, f({x_{v_{p1}}}^{cf}, \ldots, {x_{v_{pk}}}^{cf}))$$

where $v_{p1}, .., v_{pk}$ are the direct causes of v and f represents the ML Classifier to be explained

Preserving Feasibility via Oracle

- Knowledge of Structural Causal Models might be impractical in real life datasets
- Oracle implicitly models the constraint and provides a black box access via feasibility score
  - Oracle can represent user feedback to preserve user specific / local constraints
  - Oracle could be used to represent complex global constraints which are hard to optimize directly
- Score corresponding to Labelled CFs (q^{cf}) via Oracle:
  OracleGenCF: $e^{-(x^{cf} - q^{cf})^T (x^{cf} - q^{cf})}$



Our Approaches

BaseGenCF: Variational Inference based loss

AEGenCF: BaseGenCF + Reconstruction Loss on CF via Auto Encoder [1]

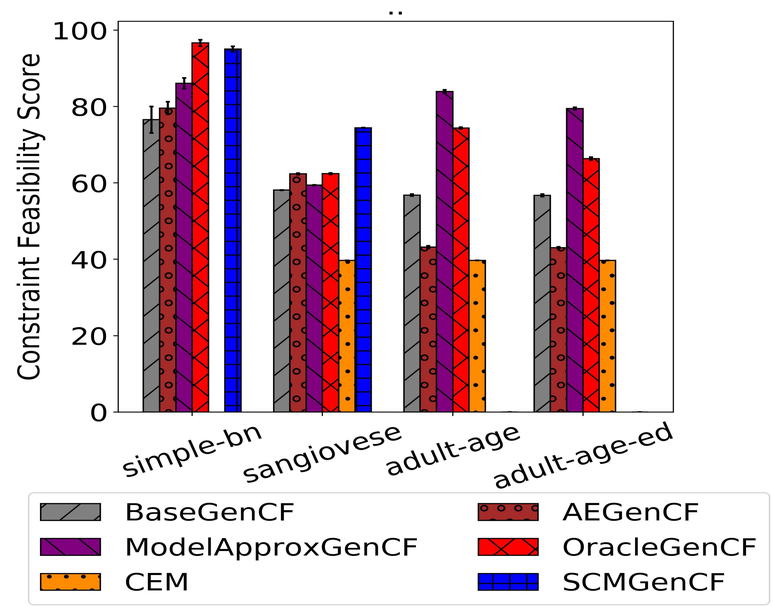
SCMGenCF: BaseGenCF + Causal Proximity Regularizer

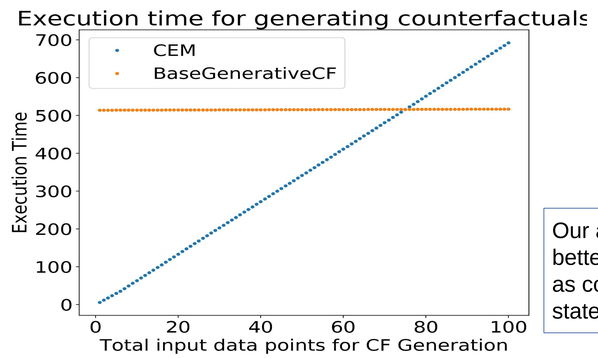
ModelApproxGenCF: BaseGenCF + Constraint Based Loss

OracleGenCF: BaseGenCF + Loss with CFs labelled via Oracle

Results

- Simple-BN: Synthetic dataset with monotonic constraint
- Sangiovese: Bayesian Network with monotonic constraint
- Adult: Real World dataset with unary and monotonic constraint
- Evaluation Metrics:
  - Validity, Proximity, Constraint Feasibility, Causal Edge Score, Causal Graph Score





Our approach scales better with data points as compared to the state of the art [1]

NeurIPS 2019 Workshop

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

Vancouver, Canada

References

[1] Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In Advances in Neural Information Processing Systems, pages 592–603, 2018