How does Predictive Uncertainty Quantification Correlate with the Plausibility of Counterfactual Explanations

Author: Dimitar Nikolov

Email: d.n.nikolov-1@student.tudelft.nl

Responsible Professor: Cynthia Liem

Supervisor: Patrick Altmeyer

1. Introduction

Counterfactual explanations can be applied to algorithmic recourse, which is concerned with helping individuals in the real world overturn undesirable algorithmic decisions

Factual

Plausible Unfaithful

Implausible Faithful

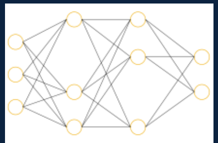
~ Plausible Faithful

Predictive uncertainty quantification measures the degree of certainty a model has in its predictions. We will measure it locally with Predictive entropy [1] and globally with Expected calibration error [2]

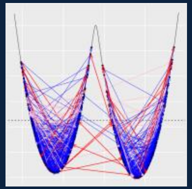
Bayesian Deep Learning treats the parameters of the neural network as random variables. The predictions become a distribution which is intractable.

$$P(\hat{y}|\hat{x}) = E_{P(w|D)}[P(\hat{y}|\hat{x}, w)]$$

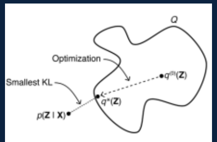
Dropout



Langevin Dynamics



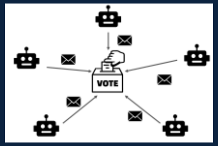
Source [3]

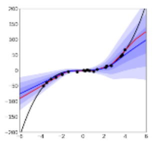


Source [4]

Variational Inference

Deep Ensemble





Source [5]

Laplace Approximation

2. Research Question

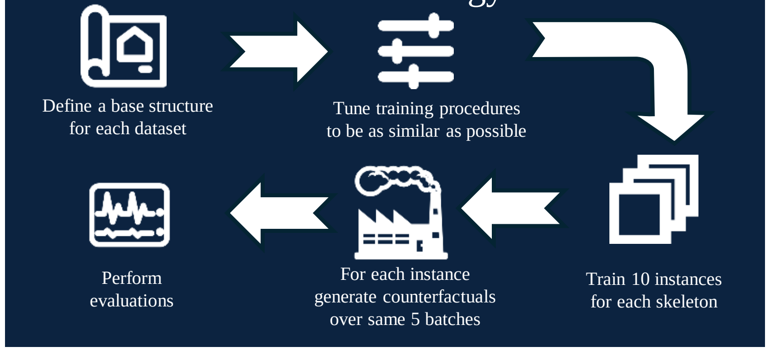
How do counterfactual explanations correlate with predictive uncertainty quantification?

Are models that provide uncertainty quantification more explainable?

How to make the different models comparable?

How modalities of the data influence the plausibility of the counterfactual explanations?

3. Methodology



Define a base structure for each dataset

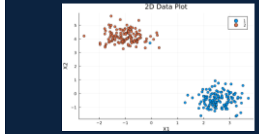
Tune training procedures to be as similar as possible

Train 10 instances for each skeleton

For each instance generate counterfactuals over same 5 batches

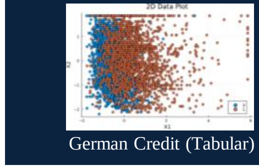
Perform evaluations

4. Datasets

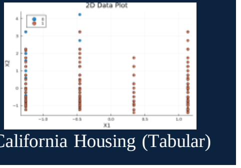


Linearly Separable (Synthetic)

MNIST (Visual)



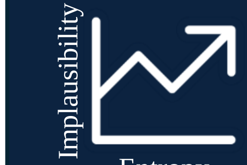
German Credit (Tabular)



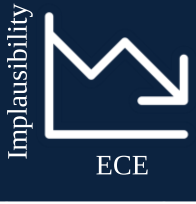
California Housing (Tabular)

5. Results

Quantitative



Entropy

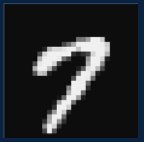


ECE

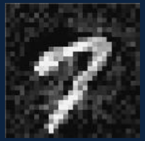
Linearly Separable	California Housing	German Credit	MNIST
ANN	ANN	Deep Ensemble	Deep Ensemble

Dataset and the model produced the least implausible counterfactuals

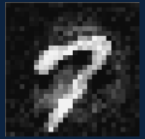
Qualitative



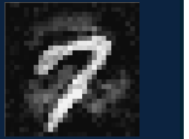
Factual



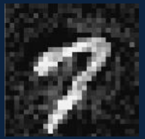
ANN – Generic



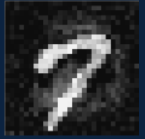
Dropout – Generic



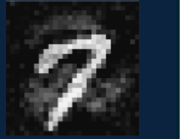
Ensemble – Generic



ANN – ECCo



Dropout – ECCo



Ensemble – ECCo

6. Future Research

Perform more extensive hyper-parameter tuning on the predictive uncertainty models

Identify more suitable metric for evaluation of plausibility

7. Conclusions

Predictive uncertainty models are more capable to learn the visual data than the tabular data

Predictive uncertainty by itself seems necessary but not sufficient to guarantee plausibility of the counterfactuals

References

[1] C. E. Shannon, "A Mathematical Theory of Communication," Bell Syst. Tech. J., vol. 27, no. 3, pp. 379–423, Jul. 1948.
[2] M. H. DeGroot and S. E. Fienberg, "The Comparison and Evaluation of Forecasters," The Statistician, vol. 32, no. 1/2, p. 12, Mar. 1983.
[3] W. Deng, G. Lin, and F. Liang, "A Contour Stochastic Gradient Langevin Dynamics Algorithm for Simulations of Multi-modal Distributions."
[4] "The ELBO in Variational Inference," gregorygundersen.com. https://gregorygundersen.com/blog/2021/04/16/variational-inference/
[5] H. Ritter, A. Botev, and D. Barber, "A SCALABLE LAPLACE APPROXIMATION FOR NEURAL NETWORKS," Int. Conf. Learn. Represent., 2018.