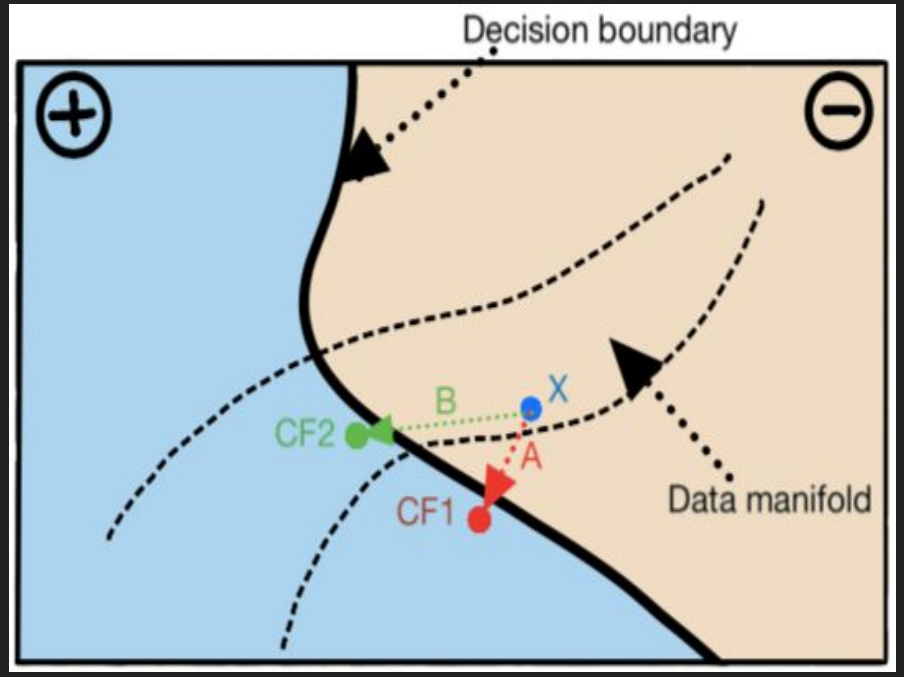
Counterfactual Explanations for Machine Learning: Challenges Revisited

Sahil Verma, John Dickerson, Keegan Hines

CHI 2021 HCXAI Workshop

What are Counterfactual Explanations?

- Emerging explainability technique in Machine Learning.
- For a datapoint and a model, counter factual explanation is a datapoint in the vicinity, but with a different prediction (additional desiderata).
- Helpful in answering questions about the changes required to be brought for getting to other side of decision boundary.



Verma, S., Dickerson, J., & Hines, K. (2020). Counterfactual Explanations for Machine Learning: A Review.

Karimi, A. H., Barthe, G., Schölkopf, B., & Valera, I. (2020). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects.

Desiderata of Counterfactual Explanations

- Actionability / Mutability of features
- Causal Relations
- Adherence to Data Manifold
- Sparsity
- Model-agnosticity and Black-box approach
- Amortized Inference

Challenges in industry adoption

Challenges with causality:

- Lack of structured causal models (SCMs).
- Lack of interventional data.

Challenges with real world dynamics:

- ML models are not static.
- Individuals might not be able to precisely follow previous advice.

Challenges in industry adoption

Challenges with biases:

- CFEs do not consider the demographic bias in the ML models.
- CFEs do not yet adhere to the personal biases of individuals.

Challenges with community action:

- Lack of unified vocabulary within the XAI research community.
- Lack of acceptance by regulatory bodies (legal community).
- Lack of visualization of CFEs (HCI community).
- Missing applicability in regression problems (ML community).

Thank You!

Counterfactual Explanations for Machine Learning: Challenges Revisited

Arthur

Sahil Verma

sahil.verma@arthur.ai

@Sahil1V