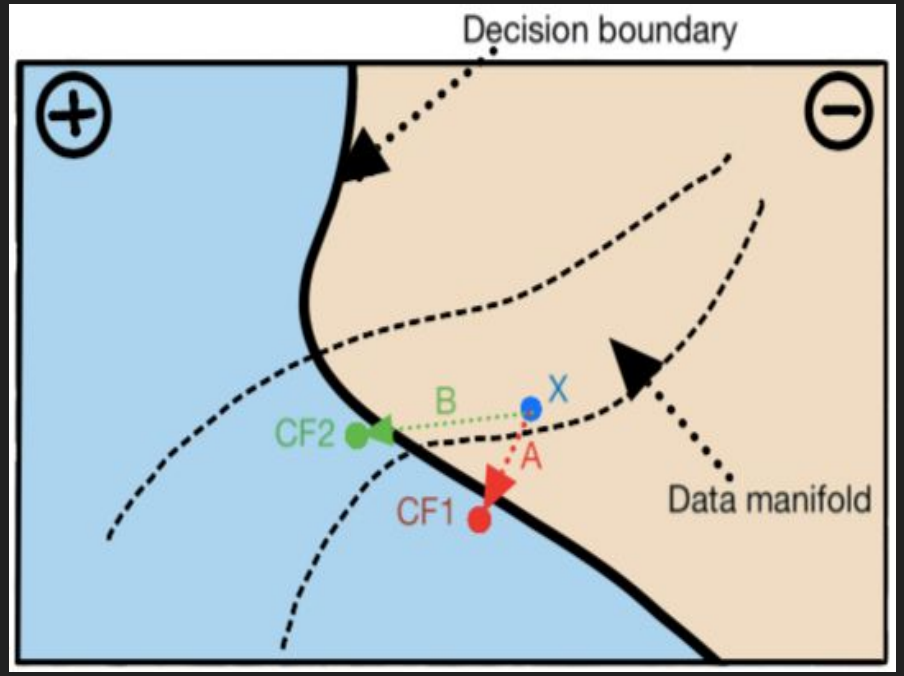
Counterfactual Explanations for Machine Learning: A Review

Sahil Verma, John Dickerson, Keegan Hines

NeurIPS 2020 ML-RSA Workshop

What are Counterfactual Explanations?

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



Contributions

- We examine a set of 39 recent papers and develop a rubric for easy comparison and comprehension of difference approaches.
- We provide a comprehensive and lucid introduction to beginners.
- We identify future research directions and justify their importance.

Desiderata of Counterfactual Explanations

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

Actionability / Mutability of Features

- An effective counterfactual must not change immutable and unactionable features, e.g., citizenship, marital status.
- Only 35% of previous papers consider actionability among features.

Causal Relations

- An effective counterfactual must respect the causal relations among features, e.g., getting a new degree would increase the age, age can’t decrease.
- Only 3 of previous papers consider causal relations among features, out of which only one can work with partially specified causal graphs.

Counterfactual vs. Contrastive explanations

- Recent discussion has been around the terminological differences between counterfactual and contrastive explanations.
- “Counterfactual” term reserved for the explanation using a structural causal model.
- Other explanations which change the model’s prediction are contrastive.

Adherence to Data Manifold

- Assuming training data is a representative distribution of features, an attainable counterfactual should be close to it.
- Adding a VAE loss term, constraining distance from k-nearest datapoints, and sampling from latent space are some ways of attaining it.
- Only 35% of the papers consider closeness to data manifold.

Model-Agnostic Approach

- Many approaches have restriction on the class of models they can handle.
- Gradient based approaches require differentiable models. Solver based approaches require piece-wise linear models.

Black-box Approach

- Most approaches require access to full model internals or gradients.
- Proprietary models can not be used with such approaches.
- Only 35% of the approaches are model-agnostic and work in black-box fashion.

Sparsity

- Shorter explanations are more comprehensible [Miller et al.] .
- Preferable to alter a few features, rather than small changes in all features.
- Using L0/L1 norm, changing features iteratively, or post-hoc sparsity are a few methods for inducing sparsity.
- 62% of the papers consider sparsity.

Amortized Inference or Fast Counterfactuals

- Most approaches need to perform separate optimization to generate counterfactual(s) for every input datapoint.
- Only one previous work learns a strategy to generate counterfactuals which can be later used for any datapoint from the data distribution.

Table 1: Evaluation of the collected papers on the same set of properties, which are important for readily comparing and comprehending the differences and limitations of different counterfactual algorithms. Papers are sorted chronologically. Full table is given in appendix A.

	Preconditions		Optimization amortization		CF attributes			CF opt. problem attributes	
Paper	Model access	Model domain	Amortized Inference	Multiple CF	Sparsity	Data manifold	Causal relation	Feature preference	Categorical dist. func
[70]	Black-box	Agnostic	No	No	Changes iteratively	No	No	Yes	-
[108]	Gradients	Differentiable	No	No	L1	No	No	No	-
[101]	Complete	Tree ensemble	No	No	No	No	No	No	-
[72]	Black-box	Agnostic	No	No	L0 and post-hoc	No	No	No	-
[55]	Black-box	Agnostic	No	Yes	Flips min. split nodes	No	No	No	Indicator
[29]	Gradients	Differentiable	No	No	L1	Yes	No	No	-
[54]	Black-box	Agnostic	No	No	No	No	No	No[2]	-
[94]	Complete	Linear	No	Yes	L1	No	No	No	N.A.[3]
[104]	Complete	Linear	No	No	Hard constraint	No	No	Yes	-
[96]	Black-box	Agnostic	No	Yes	No	No	No	Yes	Indicator
[30]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	-
[90]	Black-box	Agnostic	No	No	No	No	No	No	-
[59]	Gradients	Differentiable	No	No	No	Yes	No	No	-
[89]	Gradients	Differentiable	No	No	No	No	No	No	-
[110]	Black-box	Agnostic	No	No	Changes one feature	No	No	No	-
[83]	Gradients	Differentiable	No	Yes	L1 and post-hoc	No	No	No	Indicator
[88]	Black-box	Agnostic	No	No	No	Yes[4]	No	No	-
[105]	Black-box or gradient	Differentiable	No	No	L1	Yes	No	No	Embedding
[79]	Gradients	Differentiable	Yes	Yes	No	Yes	Yes	Yes	-
[62]	Complete	Linear	No	Yes	Hard constraint	No	No	Yes	Indicator
[85]	Gradients	Differentiable	No	No	No	Yes	No	Yes	N.A.[5]
[66]	Black-box	Agnostic	No	No	Yes	Yes	No	No	-
[63]	Complete	Linear and causal graph	No	No	L1	No	Yes	Yes	-
[64]	Gradients	Differentiable	No	No	No	No	Yes	Yes	-
[74]	Gradients	Differentiable	No	No	Changes iteratively	Yes	No	No[6]	-
[26]	Black-box	Agnostic	No	Yes	L0	Yes	No	Yes	Indicator
[61]	Complete	Linear and tree ensemble	No	No	No	Yes	No	Yes	-
[46]	Complete	Random Forest	No	Yes	L1	No	No	No	-
[77]	Complete	Tree ensemble	No	No	L1	No	No	No	-

Evaluation Metrics

- Validity
- Proximity
- Sparsity
- Diversity (if applicable)
- Causal relation satisfaction
- Closeness to data manifold
- Time taken

Open Questions

We identified several open questions, like:

- Considering bias while generating counterfactuals
- Generating robust counterfactuals
- Counterfactuals as an interactive service
- Counterfactuals as discrete and sequential steps of actions

Conclusions

- We collected papers in counterfactual explanations and tied them in a lucid manner to serve as an introduction as well as platform for reckoning current research.
- We developed a rubric to evaluate past and future research on counterfactual explanations.
- We pose several open questions that can influence future research directions.

Thank you!

Questions?