Towards Unifying Feature Attribution and Counterfactual Explanations: Different Means to the Same End

Ramaravind Kommiya Mothilal
**Microsoft Research India
raam.arvind93@gmail.com

Amit Sharma
Microsoft Research India
amshar@microsoft.com

Divyat Mahajan
Microsoft Research India
t-dimaha@microsoft.com

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

**Work done during stay at MSR
Local Explanation Methods Convey Different Pictures

Feature Attributions and Counterfactuals often disagree even for simple linear models

\[ f(x_1, x_2) = I(0.45x_1 + 0.1x_2 \geq 0.5), \quad x_1, x_2 \in [0, 1] \]

| Feature Attributions      | LIME (Ribeiro et al., 2016) | SHAP (Lundberg et al., 2017) | WachterCF (Wachter et al., 2017) | DiCE (Mothilal et al., 2020) |
|---------------------------|-----------------------------|-------------------------------|----------------------------------|-----------------------------|
|                           |                             |                               |                                  |                             |
| Counterfactual examples   |                             |                               |                                  |                             |
|                           |                             |                               |                                  |                             |
| Importance Scores         |                             |                               |                                  |                             |
| \( x_1 \)                 | 0.34                        | 0.69                          | 0.98                             | 1.00                        |
| \( x_2 \)                 | 0.07                        | 0.28                          | 0.97                             | 0.98                        |
Complementarity of Local Explanation Methods

Contributions:

• A unifying framework based on Actual Causality (Halpern, 2016) to interpret Feature Attributions and Counterfactual Explanations

• Evaluate attribution-based methods on the necessity and sufficiency of their top-ranked features using Counterfactual Explanations
Actual Causality and Model Explanations

\((\alpha, \beta)\) goodness of an explanation

**Necessity:** 
\[
\alpha = \Pr(x_j \text{ is a cause of } y^* | x_j = a, y = y^*)
\]

“is a cause” \(\Rightarrow x_j = a \) satisfies the definition of actual causality

**Sufficiency:** 
\[
\beta = \Pr(y = y^* | x_j \leftarrow a)
\]
Counterfactuals Measure Necessity and Feature Attributions Measure Sufficiency

**Counterfactual explanation (\(\alpha_{CF}\))**

- Optimizes Necessity
- Perturbed feature subset \(x_j\) is a but-for cause of the original output
- \(\alpha_{CF}\) summarizes the outcomes of all such perturbations and ranks any feature subset for their necessity

\[
\alpha_{CF} = \Pr((x_j \leftarrow a' \Rightarrow y \neq y^*)| x_j = a, x_{-j} = b, y = y^*)
\]

**Attribution-based explanations (\(\beta\))**

- Optimizes Sufficiency
- Importance of \(x_j\) can be interpreted as its sufficiency
- \(\beta\) provides the fraction of all contexts where \(x_j \leftarrow a\) leads to \(y = y^*\)

\[
\beta = \Pr(y = y^*| x_j \leftarrow a)
\]
Building Blocks of Explanations: Necessity and Sufficiency

Counterfactual Explanations to evaluate Feature Attribution Methods

\[
\text{Necessity} = \frac{\sum_{i, x_j \neq a} 1(CF_i)}{nCF \times N}
\]

\[
\text{Sufficiency} = \frac{\sum_{i} 1(CF_i)}{nCF \times N} - \frac{\sum_{i, x_j \leftarrow a} 1(CF_i)}{nCF \times N}
\]

Steps:
- Generate CFs by changing only \( x_j \)
- Compute the fraction of times that changing \( x_j \) leads to a valid counterfactual example

Steps:
- Generate CFs by fixing only \( x_j \)
- Compare the fraction of unique CFs generated using all features to that generated while keeping \( x_j \) constant
Results: Evaluating Necessity and Sufficiency

Data: Adult-Income, LendingClub, German-Credit, HospitalTriage (222 features)

Methods: LIME, SHAP, DiCE, WachterCF
Results: Evaluating Necessity and Sufficiency

Data: Adult-Income, LendingClub, German-Credit, HospitalTriage (222 features)

Methods: LIME, SHAP, DiCE, WachterCF

Key Results:

• Highly ranked features may often neither be necessary nor sufficient explanations of a model’s predictions

• Necessity and Sufficiency become weaker for top-ranked features as the number of features in a dataset increases
Summary

- **Unifying framework** for attribute-based and counterfactual examples using actual causality
- **Evaluate** attribution-based methods on the **necessity** and **sufficiency** of their top-ranked features using counterfactual explanations
- **Generate necessity-inspired** feature attributions