ECCCos from the Black Box

Faithful Model Explanations through Energy-Based Conformal Counterfactuals

Patrick Altmeyer

Mojtaba Farmanbar Arie van Deursen Cynthia C. S. Liem

Delft University of Technology

2024-01-04

We propose *ECCCo*: a new way to generate faithful model explanations that are as plausible as the underlying model permits.

Summary

▶ Idea: generate counterfactuals that are consistent with what the model has learned about the data.

Pick your Poison?

```
Which one would you pick?

Factual Wachter (p=0.93) Schut (p=0.91) REVISE (p=1.0)
```

We propose *ECCCo*: a new way to generate faithful model explanations that are as plausible as the underlying model permits.

Summary

- ▶ Idea: generate counterfactuals that are consistent with what the model has learned about the data.
- ▶ **Method**: constrain the model's energy and predictive uncertainty for the counterfactual.

Pick your Poison?

```
Which one would you pick?

Factual Wachter (p=0.93) Schut (p=0.91) REVISE (p=1.0)
```

We propose *ECCCo*: a new way to generate faithful model explanations that are as plausible as the underlying model permits.

Summary

- ▶ Idea: generate counterfactuals that are consistent with what the model has learned about the data.
- ▶ **Method**: constrain the model's energy and predictive uncertainty for the counterfactual.
- ▶ Result: faithful counterfactuals that are as plausible as the model permits.

Pick your Poison?

```
Which one would you pick?

Factual Wachter (p=0.93) Schut (p=0.91) REVISE (p=1.0)
```

We propose *ECCCo*: a new way to generate faithful model explanations that are as plausible as the underlying model permits.

Summary

- ▶ Idea: generate counterfactuals that are consistent with what the model has learned about the data.
- ▶ **Method**: constrain the model's energy and predictive uncertainty for the counterfactual.
- ▶ Result: faithful counterfactuals that are as plausible as the model permits.
- **Benefits**: enable us to distinguish trustworthy from unreliable models.

Pick your Poison?

```
Which one would you pick?

Factual Wachter (p=0.93) Schut (p=0.91) REVISE (p=1.0)
```

Reconciling Faithfulness and Plausibility Counterfactual Explanations

Plausibility

There's no consensus on the exact definition of plausibility but we think about it as follows:

Definition (Plausible Counterfactuals)

Let $\mathcal{X}|\mathbf{y}^+ = p(\mathbf{x}|\mathbf{y}^+)$ denote the true conditional distribution of samples in the target class \mathbf{y}^+ . Then for \mathbf{x}' to be considered a plausible counterfactual, we need: $\mathbf{x}' \sim \mathcal{X}|\mathbf{y}^+$.

Plausibility has been linked to actionability, fairness and robustness.

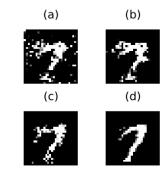
Faithfulness

Definition (Faithful Counterfactuals)

Let $\mathcal{X}_{\theta}|\mathbf{y}^+ = p_{\theta}(\mathbf{x}|\mathbf{y}^+)$ denote the conditional distribution of \mathbf{x} in the target class \mathbf{y}^+ , where θ denotes the parameters of model M_{θ} . Then for \mathbf{x}' to be considered a faithful counterfactual, we need:

Results

Visual Evidence



The Numbers

Questions?

With thanks to my co-authors Mojtaba Farmanbar, Arie van Deursen and Cynthia C. S. Liem.



Counterfactual Explanations

All the work presented today is powered by CounterfactualExplanations.jl .

There is also a corresponding paper, *Explaining Black-Box Models through Counterfactuals*, which has been published in JuliaCon Proceedings.

References