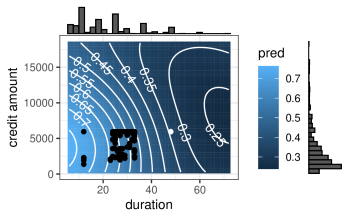
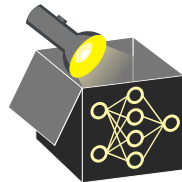


Interpretable Machine Learning

Counterfactual Explanations: Methods & Discussion of CEs



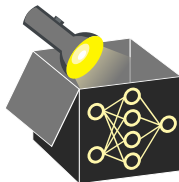
Learning goals

- See two strategies to generate CEs
- Know problems and limitations of CEs

OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

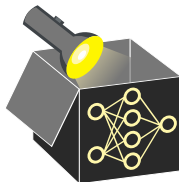
- **Target:** Most support classification; few extend to regression
 \rightsquigarrow Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

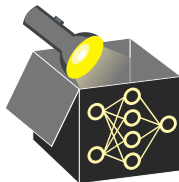
- **Target:** Most support classification; few extend to regression
~→ Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

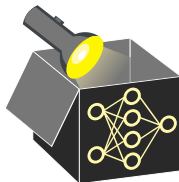
- **Target:** Most support classification; few extend to regression
~→ Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio
- **Feature space:** Some handle only numerical features; few support mixed data types



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

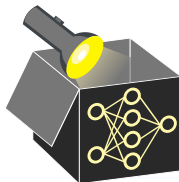
- **Target:** Most support classification; few extend to regression
~> Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio
- **Feature space:** Some handle only numerical features; few support mixed data types
- **Objectives:** From core goals like sparsity and plausibility to emerging aims such as fairness, personalization, and robustness



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

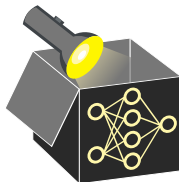
- **Target:** Most support classification; few extend to regression
~→ Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio
- **Feature space:** Some handle only numerical features; few support mixed data types
- **Objectives:** From core goals like sparsity and plausibility to emerging aims such as fairness, personalization, and robustness
- **Model access:** Methods range from model-specific (requiring access to model internals/gradients) to model-agnostic (using only prediction funcs)



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

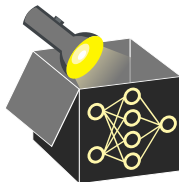
- **Target:** Most support classification; few extend to regression
~> Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio
- **Feature space:** Some handle only numerical features; few support mixed data types
- **Objectives:** From core goals like sparsity and plausibility to emerging aims such as fairness, personalization, and robustness
- **Model access:** Methods range from model-specific (requiring access to model internals/gradients) to model-agnostic (using only prediction funcs)
- **Optimization:** From gradient-based (differentiable models) and mixed-integer programming (linear models) to gradient-free methods (e.g., genetic algorithms)



OVERVIEW OF COUNTERFACTUAL METHODS

Many methods exist to generate counterfactuals, they mainly differ in:

- **Target:** Most support classification; few extend to regression
~> Recent work extends CEs to other ML tasks (un-, semi-, self-supervised)
- **Data type:** Focus is on tabular data; little work on text, vision, audio
- **Feature space:** Some handle only numerical features; few support mixed data types
- **Objectives:** From core goals like sparsity and plausibility to emerging aims such as fairness, personalization, and robustness
- **Model access:** Methods range from model-specific (requiring access to model internals/gradients) to model-agnostic (using only prediction funcs)
- **Optimization:** From gradient-based (differentiable models) and mixed-integer programming (linear models) to gradient-free methods (e.g., genetic algorithms)
- **Rashomon Effect:** Many methods return one CE, some diverse sets of CEs, others prioritize CEs, or let the user choose



FIRST OPTIMIZATION-BASED CE METHOD

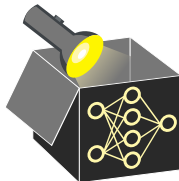
► WACHTER_2018

Introduced CEs in context of ML predictions by solving

$$\arg \min_{\mathbf{x}'} \max_{\lambda} \underbrace{\lambda (\hat{f}(\mathbf{x}') - y')^2}_{o_{target}(\hat{f}(\mathbf{x}'), y')} + \underbrace{\sum_{j=1}^p \frac{|x'_j - x_j|}{MAD_j}}_{o_{proximity}(\mathbf{x}', \mathbf{x})}$$

- o_{target} ensures prediction flips to y' (by increasing weight λ)
- $o_{proximity}$ penalizes deviations from \mathbf{x} , rescaled by median abs. deviation:

$$MAD_j = \text{med}_{i \in \{1, \dots, n\}} (|x_j^{(i)} - \text{med}_{k \in \{1, \dots, n\}} (x_j^{(k)})|)$$



FIRST OPTIMIZATION-BASED CE METHOD

► WACHTER_2018

Introduced CEs in context of ML predictions by solving

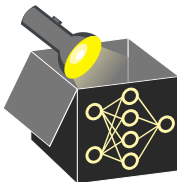
$$\arg \min_{\mathbf{x}'} \max_{\lambda} \underbrace{\lambda (\hat{f}(\mathbf{x}') - y')^2}_{o_{target}(\hat{f}(\mathbf{x}'), y')} + \underbrace{\sum_{j=1}^p \frac{|x'_j - x_j|}{MAD_j}}_{o_{proximity}(\mathbf{x}', \mathbf{x})}$$

- o_{target} ensures prediction flips to y' (by increasing weight λ)
- $o_{proximity}$ penalizes deviations from \mathbf{x} , rescaled by median abs. deviation:

$$MAD_j = \text{med}_{i \in \{1, \dots, n\}} (|x_j^{(i)} - \text{med}_{k \in \{1, \dots, n\}} (x_j^{(k)})|)$$

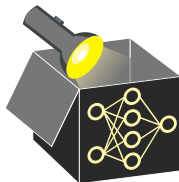
Approach: Alternating optimization over \mathbf{x}' and λ

- Start with an initial λ (controls emphasis on o_{target} vs. $o_{proximity}$)
- Use a gradient-free optimizer (e.g., Nelder-Mead) to minimize over \mathbf{x}'
- If prediction constraint not satisfied ($\hat{f}(\mathbf{x}') \neq y'$), increase λ and repeat
 $\rightsquigarrow \lambda$ serves as soft constraint, gradually enforcing prediction validity
 $\hat{f}(\mathbf{x}') = y'$
- Iteratively shift focus: 1. achieve prediction validity, 2. minimize proximity



LIMITATIONS OF WACHTER'S APPROACH

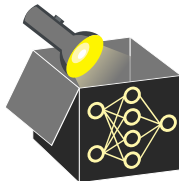
- **Manual tuning:** No principled way to set λ ; requires iterative increase
- **Asymmetric focus:** Early iterations dominated by minimizing target loss
- **Limited feature support:** Proximity term defined only for numerical feats
- **No additional objectives:** Ignores sparsity, plausibility, fairness, diversity
- **Single solution:** Returns one CE; no support for diverse or ranked CEs



- **Multi-Objective Counterfactual Explanations (MOC):** Instead of collapsing objectives into a single obj., optimize all 4 obj. simultaneously

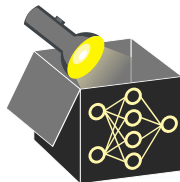
$$\arg \min_{\mathbf{x}'} \left(o_{target}(\hat{f}(\mathbf{x}'), y'), o_{proximity}(\mathbf{x}', \mathbf{x}), o_{sparse}(\mathbf{x}', \mathbf{x}), o_{plausible}(\mathbf{x}', \mathbf{X}) \right).$$

- Avoids using/tuning of weights (e.g., λ); returns Pareto-optimal set
- Uses an adjusted multi-objective genetic algo. (NSGA-II) for mixed feats
- Outputs diverse CEs representing different trade-offs between objectives



EXAMPLE: CREDIT DATA

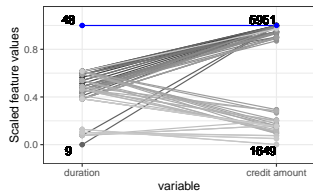
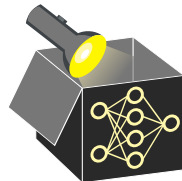
- Model: SVM with RBF kernel
- \mathbf{x} : First data point of credit data with $\mathbb{P}(y = \textit{good}) = 0.34$
- Goal: Increase the probability to desired outcome $[0.5, 1]$
- MOC (with default parameters) returned 69 valid CEs after 200 iterations
- All CEs modified credit duration; many also adjusted credit amount



EXAMPLE: CREDIT DATA

► DANDL_2020

- Feature changes can be visualized using parallel and 2D surface plots
- Parallel plot: All CEs had values equal to or smaller than the values of \mathbf{x}

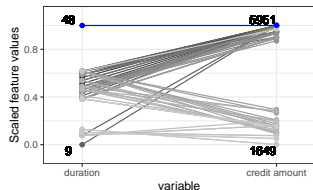
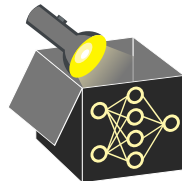


Parallel plot: Grey lines = CEs \mathbf{x}' , blue line = \mathbf{x} .
Features without changes omitted.
Bold numbers denote numeric ranges.

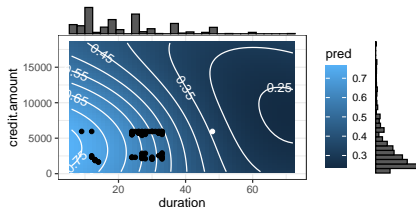
EXAMPLE: CREDIT DATA

► DANDL_2020

- Feature changes can be visualized using parallel and 2D surface plots
- Parallel plot: All CEs had values equal to or smaller than the values of \mathbf{x}
- Surface plot: CEs in lower-left appear distant, but lie in high-density regions near training data (as shown by histograms)



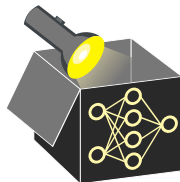
Parallel plot: Grey lines = CEs \mathbf{x}' , blue line = \mathbf{x} .
Features without changes omitted.
Bold numbers denote numeric ranges.



Surface plot: White dot = \mathbf{x} , black dots = CEs \mathbf{x}' .
Histograms: Marginal distribution of training data \mathbf{X} .

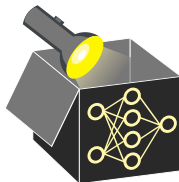
PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives, reducing complexity but offering limited explanatory power
 - ~> Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged



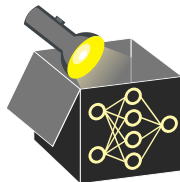
PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives, reducing complexity but offering limited explanatory power
 - ↪ Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
 - ↪ e.g., L_1 can be reasonable for tabular data but not for image data
 - ↪ sparsity desirable for end-users but not for auditors searching for model bias

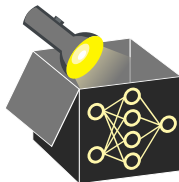


PROBLEMS, PITFALLS, & LIMITATIONS

- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives, reducing complexity but offering limited explanatory power
 - ~> Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
 - ~> e.g., L_1 can be reasonable for tabular data but not for image data
 - ~> sparsity desirable for end-users but not for auditors searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
 - ~> End-users must know that CEs explain the model, not the real world



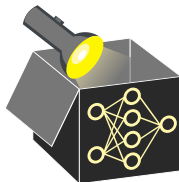
PROBLEMS, PITFALLS, & LIMITATIONS



- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives, reducing complexity but offering limited explanatory power
 - ~> Psychologists have shown that although perceived model understanding of end-users increases, the objective model understanding remains unchanged
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
 - ~> e.g., L_1 can be reasonable for tabular data but not for image data
 - ~> sparsity desirable for end-users but not for auditors searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
 - ~> End-users must know that CEs explain the model, not the real world
- **Disclosing too much information:** CEs can reveal too much information about the model and help potential attackers

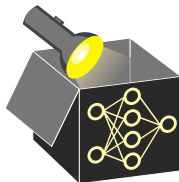
PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
~> No universal answer; depends on user goals, cognitive load, and resources



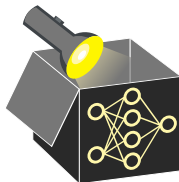
PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
 - ~> No universal answer; depends on user goals, cognitive load, and resources
- **Actionability vs. fairness:** Focusing on actionable changes may hinder contestability
 - ~> E.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model



PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
 - ~> No universal answer; depends on user goals, cognitive load, and resources
- **Actionability vs. fairness:** Focusing on actionable changes may hinder contestability
 - ~> E.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model
- **Assumption of constant model:** To provide guidance for the future, CEs assume that their underlying model does not change in the future
 - ~> in reality this assumption is often violated making CEs unreliable



PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
 - ~> No universal answer; depends on user goals, cognitive load, and resources
- **Actionability vs. fairness:** Focusing on actionable changes may hinder contestability
 - ~> E.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model
- **Assumption of constant model:** To provide guidance for the future, CEs assume that their underlying model does not change in the future
 - ~> in reality this assumption is often violated making CEs unreliable
- **Attacking CEs:** Researchers can create models with great performance, which generate arbitrary explanations specified by the ML developer
 - ~> how faithful are CEs to the models underlying mechanism?

