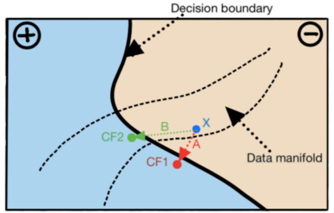
Interpretable Machine Learning

Counterfactual Explanations (CEs): Motivation

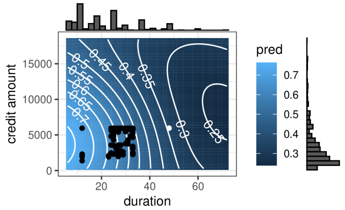


Learning goals

- Understand the motivation behind CEs
- Know why and how CEs are used
- Recognize the philosophical foundations of counterfactual reasoning

Interpretable Machine Learning

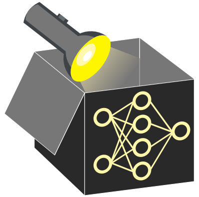
Counterfactual Explanations: Methods & Discussion of CEs



Learning goals

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

MOTIVATING EXAMPLE: CREDIT RISK & CE



x: customer and credit information $\quad$ y: grant or reject credit

Age 52

Gender m

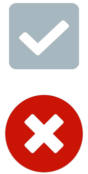
Job unskilled

Amount 10T

Duration 24

Purpose TV

→ **BLACK BOX** →



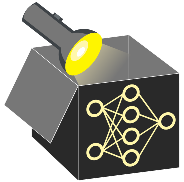
Grant

Reject

Potential questions:

- Why was the credit rejected?
- Is this decision fair compared with similar applicants?
- **How should x be changed so that the credit is accepted?**

OVERVIEW OF COUNTERFACTUAL METHODS

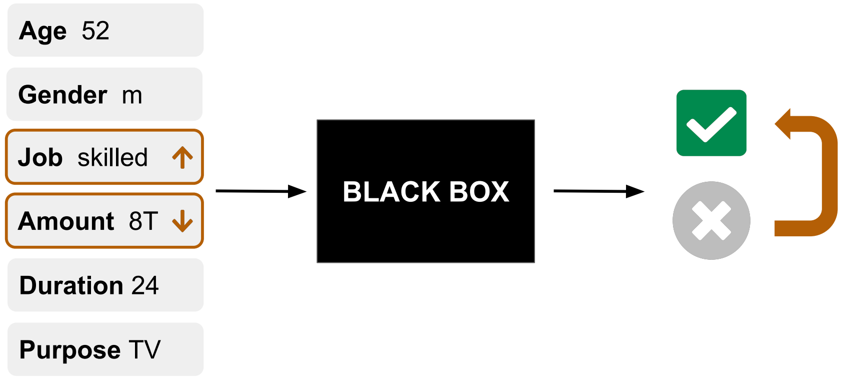


Many methods exist to generate counterfactuals, they mainly differ in:

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

MOTIVATING EXAMPLE: CREDIT RISK & CE

Counterfactual Explanations provide answers in the form of "What-If"-scenarios.



“If the applicant had higher skills and the credit amount had been reduced to $8.000, the loan would have been granted."

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)
- **Data type:** Focus is on tabular data; little work on text, vision, audio

CORE DEFINITION AND PURPOSE OF CE

- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'

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)
- **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

CORE DEFINITION AND PURPOSE OF CE

- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

$$\text{Find} \quad \mathbf{x}' \approx \mathbf{x} \quad \text{such that} \quad f(\mathbf{x}') = y' \quad \text{and distance} \quad d(\mathbf{x}, \mathbf{x}') \quad \text{is minimal}$$

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)
- **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

CORE DEFINITION AND PURPOSE OF CE

- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

$$\text{Find} \quad \mathbf{x}' \approx \mathbf{x} \quad \text{such that} \quad f(\mathbf{x}') = y' \quad \text{and distance} \quad d(\mathbf{x}, \mathbf{x}') \quad \text{is minimal}$$

- **Minimal actionable changes:** Difference $\mathbf{x}' - \mathbf{x}$ shows the smallest feature change a user could realize in practice

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)
- **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)

CORE DEFINITION AND PURPOSE OF CE

- **Counterfactual explanation (CE):** Hypothetical input $\mathbf{x}'$ close to the data point of interest $\mathbf{x}$ whose prediction equals a user-defined desired outcome y'
- **Proximity constraint:**

$$\text{Find} \quad \mathbf{x}' \approx \mathbf{x} \quad \text{such that} \quad f(\mathbf{x}') = y' \quad \text{and distance} \quad d(\mathbf{x}, \mathbf{x}') \quad \text{is minimal}$$

- **Minimal actionable changes:** Difference $\mathbf{x}' - \mathbf{x}$ shows the smallest feature change a user could realize in practice
- **Primary audience:**
  - Individuals aiming to alter model predictions
  - ML engineers exploring model behavior under adversarial conditions
    $\leadsto$ how small text changes in email flip prediction from "spam" to "no spam"

OVERVIEW OF COUNTERFACTUAL METHODS

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

- **Target:** Most support classification; few extend to regression
  $\leadsto$ 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)

INTERPRETIVE AIMS & ROLES

CEs can serve various purposes; the user can decide what to learn from them, e.g.:

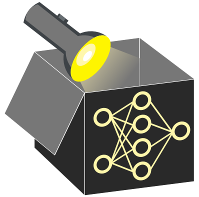
“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

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

INTERPRETIVE AIMS & ROLES



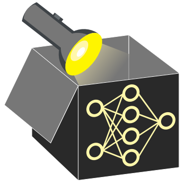
CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to \$12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

FIRST OPTIMIZATION-BASED CE METHOD

▸ WACHTER_2018



Introduced CEs in context of ML predictions by solving

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

- o_{target} ensures prediction flips to y' (by increasing weight λ)
- $o_{proximity}$ penalizes deviations from $\mathbf{x}$, rescaled by median abs. deviation:
  $MAD_j = \text{med}_{i\in\{1,\ldots,n\}}(|x_j^{(i)} - \text{med}_{k\in\{1,\ldots,n\}}(x_j^{(k)})|)$

INTERPRETIVE AIMS & ROLES

CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

FIRST OPTIMIZATION-BASED CE METHOD

▸ WACHTER_2018

Introduced CEs in context of ML predictions by solving

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

- 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
  $\rightsquigarrow$ λ serves as soft constraint, gradually enforcing prediction validity $\hat{f}(\mathbf{x}') = y'$
- Iteratively shift focus: 1. achieve prediction validity, 2. minimize proximity

INTERPRETIVE AIMS & ROLES

CEs can serve various purposes; the user can decide what to learn from them, e.g.:

“If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.

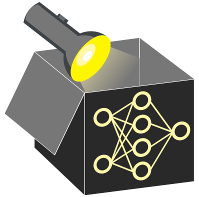
- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.

- **Provide grounds to contest the decision:**
  How dare you, I do not want to be discriminated for my age in an application.

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

INTERPRETIVE AIMS & ROLES

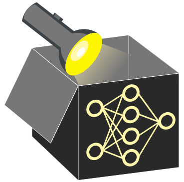


CEs can serve various purposes; the user can decide what to learn from them, e.g.:

"If the person had been **one year older** and the **credit amount had been increased** to $12.000, the credit would have been granted."

- **Guidance for future actions:**
  Ok, I will apply again next year for the higher amount.
- **Provide reasons:**
  Interesting, I did not know that age plays a role in loan applications.
- **Provide grounds to contest the decision:**
  How dare you, I do not want to be discriminated for my age in an application.
- **Detect model biases:**
  There is a bug, an increase in amount should not increase approval rates.

MULTI-OBJECTIVE CE ▸ DANDL_2020



- **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

PHILOSOPHICAL FOUNDATIONS ▸ Lewis (1973)

Counterfactuals have a long tradition in analytic philosophy
$\rightsquigarrow$ A **counterfactual conditional** takes the form:

"If S had occurred, Q would have occurred."

- *S*: past event that never happened $\rightsquigarrow$ CE run contrary to fact
- Statement is true iff *Q* holds in all **closest** worlds where *S* is true
- Closest worlds preserve laws and change as few facts as possible (related to *S*)

EXAMPLE: CREDIT DATA

- Model: SVM with RBF kernel
- **x**: First data point of credit data with $\mathbb{P}(y = good) = 0.34$
- 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

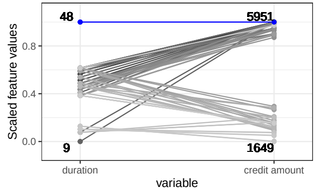
PHILOSOPHICAL FOUNDATIONS

- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\rightsquigarrow$ Good CEs point to critical causal factors that drove the algorithmic decision
  $\rightsquigarrow$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features

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 $\mathbf{x}$



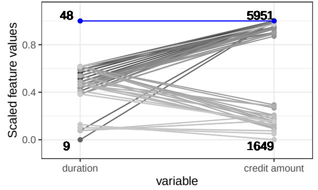
Parallel plot: Grey lines = CEs $\mathbf{x}'$, blue line = $\mathbf{x}$.
Features without changes omitted.
Bold numbers denote numeric ranges.

PHILOSOPHICAL FOUNDATIONS

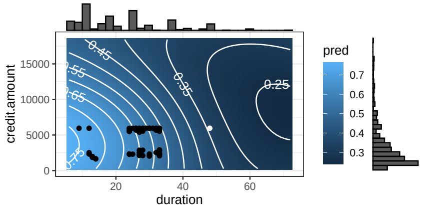
- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\rightsquigarrow$ Good CEs point to critical causal factors that drove the algorithmic decision
  $\rightsquigarrow$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features
- Relaxing closeness may add causally irrelevant edits to the explanation
  $\rightsquigarrow$ e.g., suggest to lower loan *and* increase age (but only loan matters)

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 $\mathbf{X}$.

PHILOSOPHICAL FOUNDATIONS

- CEs have largely been studied to explain causal dependence
- **Causal dependence**: Q depends on $S \Leftrightarrow$ without S, no Q
  $\rightsquigarrow$ Good CEs point to critical causal factors that drove the algorithmic decision
  $\rightsquigarrow$ **CE objective**: find $\mathbf{x}' \approx \mathbf{x}$ with $f(\mathbf{x}') = y'$ to expose causal features
- Relaxing closeness may add causally irrelevant edits to the explanation
  $\rightsquigarrow$ e.g., suggest to lower loan *and* increase age (but only loan matters)
- CEs are contrastive: Explain a decision by comparing it to a different outcome
  $\rightsquigarrow$ If age were 30 instead of 60, loan would have been \$9k instead of rejected
  $\rightsquigarrow$ Answers contrastive question: "Why Q' instead of Q?" (preferred by humans)

PROBLEMS, PITFALLS, & LIMITATIONS

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