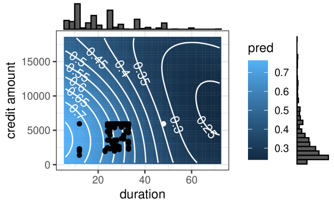
Interpretable Machine Learning

Methods & Discussion of CEs

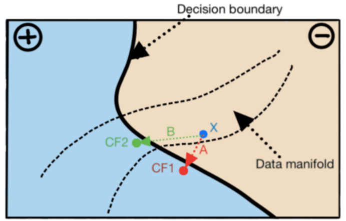


Learning goals

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

Interpretable Machine Learning

Counterfactual Explanations: Optimization Problem and Objectives



Learning goals

- Formulate CEs as optimization problem
- Identify key objectives (proximity, sparsity)
- Understand trade-offs in CE generation

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)

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y' =$ "grant credit") or interval ($y' = [1000, \infty[$)

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

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y' =$ "grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE's prediction $\hat{f}(\mathbf{x}')$ is equal to the desired pred. y'
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

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 types

MATHEMATICAL PERSPECTIVE

Terminology:

- $\mathbf{x}$: original/factual data point whose prediction we want to explain
- $y' \subset \mathbb{R}^g$: desired predi. ($y' =$ "grant credit") or interval ($y' = [1000, \infty[$)

A **valid** counterfactual $\mathbf{x}'$ satisfies two criteria:

1. **Prediction validity:** CE's prediction $\hat{f}(\mathbf{x}')$ is equal to the desired pred. y'
2. **Proximity:** CE $\mathbf{x}'$ is as close as possible to the original input $\mathbf{x}$

Reformulate these two objectives as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_{target}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{proximity}(\mathbf{x}', \mathbf{x})$$

- λ_1 and λ_2 balance the two objectives
- o_{target}: distance in target space
- $o_{proximity}$: distance in feature space

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 types
- **Objectives:** From core goals like sparsity and plausibility to emerging aims such as fairness, personalization, and robustness

OBJECTIVE FUNCTIONS ▸ DANDL_2020

Distance in target space o_{target}:

- **Regression:** L_1 distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

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 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 model internals or access to gradients) to model-agnostic (using only prediction functions)

OBJECTIVE FUNCTIONS ▸ DANDL_2020

Distance in target space o_{target}:

- **Regression:** L_1 distance $o_{target}(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- **Classification:**
  - For predicted probabilities: $o_{target} = |\hat{f}(\mathbf{x}') - y'|$
  - For predicted hard labels: $o_{target} = \mathbb{I}\{\hat{f}(\mathbf{x}') \neq y'\}$

Distance in input space $o_{proximity}$: Gower distance (mixed feature types)

$$o_{proximity}(\mathbf{x}', \mathbf{x}) = (\mathbf{x}', \mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j', x_j) \in [0, 1], \text{ where}$$

- $\delta_G(x_j', x_j) = \mathbb{I}\left\{x_j' \neq x_j\right\}$ if x_j is categorical
- $\delta_G(x_j', x_j) = \frac{1}{\widehat{R}_j} |x_j' - x_j|$ if x_j is numerical
  $\rightsquigarrow$ $\widehat{R}_j$: range of feature j in the training set to ensure $\delta_G(x_j', x_j) \in [0, 1]$

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 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 model internals or access to gradients) to model-agnostic (using only prediction functions)
- **Optimization:** From gradient-based (differentiable models) and mixed-integer programming (linear models) to gradient-free methods (e.g., genetic algorithms)

FURTHER OBJECTIVES: SPARSITY

Additional constraints can improve the explanation quality of the corresponding CEs
$\leadsto$ popular constraints include **sparsity** and **plausibility**

Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations

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 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 model internals or access to gradients) to model-agnostic (using only prediction functions)
- **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

FURTHER OBJECTIVES: SPARSITY

Additional constraints can improve the explanation quality of the corresponding CEs
⇝ popular constraints include **sparsity** and **plausibility**

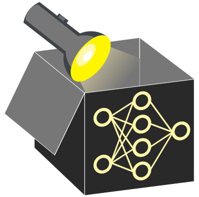
Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using L_0-norm (number of changed features) or L_1-norm (LASSO)

FIRST OPTIMIZATION-BASED CE METHOD ‣ Wachter et. al (2018)

Introduced CEs in context of ML predictions by solving

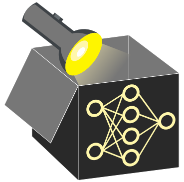
$$\arg\min_{\mathbf{x}'}\max_{\lambda} \lambda \underbrace{(\hat{f}(\mathbf{x}') - y')^2}_{o_{target}(\hat{f}(\mathbf{x}'), y')} + \underbrace{\sum_{j=1}^{p} \frac{|x_j' - x_j|}{MAD_j}}_{o_{proximity}(\mathbf{x}', \mathbf{x})}$$



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

FURTHER OBJECTIVES: SPARSITY

Additional constraints can improve the explanation quality of the corresponding CEs
$\leadsto$ popular constraints include **sparsity** and **plausibility**



Sparsity Favor explanations that change few features

- End-users often prefer short over long explanations
- Sparsity could be integrated into $o_{proximity}$
  e.g., using L_0-norm (number of changed features) or L_1-norm (LASSO)
- Alternative: Include separate objective measuring sparsity, e.g., via L_0-norm

$$o_{sparse}(\mathbf{x}', \mathbf{x}) = \sum_{j=1}^{p} \mathcal{I}_{\{x_j' \neq x_j\}}$$

FIRST OPTIMIZATION-BASED CE METHOD ▸ Wachter et. al (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 absolute deviation: $MAD_j = \text{med}_{i \in \{1,\ldots,n\}}(|x_j^{(i)} - \text{med}_{k \in \{1,\ldots,n\}}(x_j^{(k)})|))$

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: first achieve prediction validity, then minimize proximity

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed

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 features
- **No additional objectives:** Ignores sparsity, plausibility, fairness, diversity
- **Single solution:** Returns one CE; no support for diverse or ranked CEs

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values

MULTI-OBJECTIVE CE
Dandl et al. (2020)

- **Multi-Objective Counterfactual Explanations (MOC):** Instead of collapsing objectives into a single objective, optimize all four objectives 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 algorithm (NSGA-II) for mixed features
- Outputs diverse CEs representing different trade-offs between objectives

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\rightsquigarrow$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

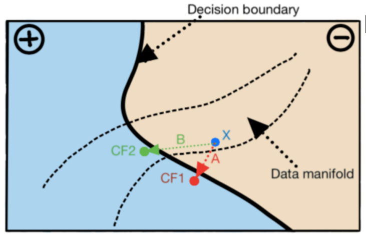
EXAMPLE: CREDIT DATA

- 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

FURTHER OBJECTIVES: PLAUSIBILITY

Plausibility:

- CEs should suggest realistic (i.e., plausible) alternatives
  $\rightsquigarrow$ Implausible: increase income *and* become unemployed
- CEs should adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ Avoid unrealistic combinations of feature values
- Estimating joint distribution is hard, especially for mixed feature spaces
  $\rightsquigarrow$ Common proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$



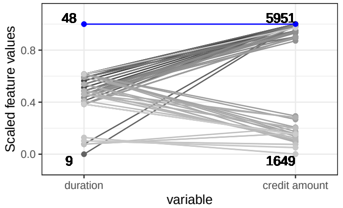
Example from ▸ Verma 2020

- Input **x** originally classified as ⊖
- Two valid CEs in class ⊕: CF1 and CF2
- Path **A** (CF1) is shorter (but unrealistic)
- Path **B** (CF2) is longer but in data manifold

EXAMPLE: CREDIT DATA

Dandl et al. (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.

FURTHER OBJECTIVES

Plausibility term: Encourage counterfactuals close to observed data.

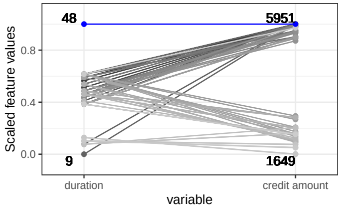
- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

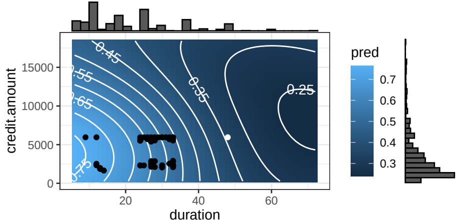
$$o_{plausibe}(\mathbf{x}', \mathbf{X}) = (\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p}\sum_{j=1}^{p} \delta_G(x_j', x_j^{[1]})$$

EXAMPLE: CREDIT DATA

Dandl et al. (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 $\mathbf{X}$.

FURTHER OBJECTIVES

Plausibility term: Encourage counterfactuals close to observed data.

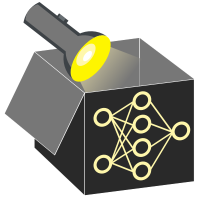
- Define $\mathbf{x}^{[1]}$ as the nearest neighbor of $\mathbf{x}'$ in the training set $\mathbf{X}$
- Use Gower distance between $\mathbf{x}'$ and $\mathbf{x}^{[1]}$ to define plausibility objective:

$$o_{plausibe}(\mathbf{x}', \mathbf{X}) = (\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p}\sum_{j=1}^{p} \delta_G(x_j', x_j^{[1]})$$

Extended optimization: Add sparsity and plausibility terms to the objective

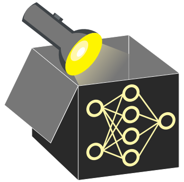
$$\underset{\mathbf{x}'}{\arg\min}\ \lambda_1 o_{\text{target}}(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_{\text{proximity}}(\mathbf{x}', \mathbf{x}) + \lambda_3 o_{\text{sparse}}(\mathbf{x}', \mathbf{x}) + \lambda_4 o_{\text{plausible}}(\mathbf{x}', \mathbf{X})$$

PROBLEMS, PITFALLS, & LIMITATIONS



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

REMARKS: THE RASHOMON EFFECT



Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

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
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
  $\rightsquigarrow$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\rightsquigarrow$ sparsity desirable for end-users but not for auditors searching for model bias

REMARKS: THE RASHOMON EFFECT

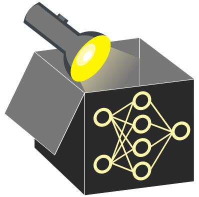
Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

Possible solutions:

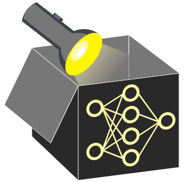
- Present all CEs for $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one/few CEs (but: by which criterion should guide this choice?)

PROBLEMS, PITFALLS, & LIMITATIONS



- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives, reducing complexity but offering limited explanatory power
  $\leadsto$ 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)
  $\leadsto$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\leadsto$ 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
  $\leadsto$ End-users must know that CEs provide insights into a model, not real world

REMARKS: THE RASHOMON EFFECT



Issue (Rashomon effect):

- Solution to the optimization problem might not be unique
- Many equally close CE might exist that obtain the desired prediction
  $\Rightarrow$ Many different equally good explanations for the same decision exist

Possible solutions:

- Present all CEs for $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one/few CEs (but: by which criterion should guide this choice?)

Note:

- Nonlinear models can produce diverse and inconsistent CEs
  $\leadsto$ suggest both increasing and decreasing credit duration (confusing for users)
- Handling this **Rashomon effect** remains an open problem in interpretable ML

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
- **Right metric:** Similarity measures are crucial to find good CEs (depends on context/domain)
  $\rightsquigarrow$ e.g., L_1 can be reasonable for tabular data but not for image data
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ End-users must know that CEs provide insights into a model, not real world
- **Disclosing too much information:** CEs can reveal too much information about the model and help potential attackers

REMARKS: MODEL OR REAL-WORLD

- CEs explain model predictions, but may seem to explain real-world users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  $\rightsquigarrow$ The applicant waits 5 years and reapplies

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

REMARKS: MODEL OR REAL-WORLD

- CEs explain model predictions, but may seem to explain real-world users
  ⇝ Transfer of model explanations to explain real-world is generally not permitted
- **Example:** CE suggests increasing age by 5 to receive a loan
  ⇝ The applicant waits 5 years and reapplies
- **Problem:** Other features may change in the meantime (e.g., job status, income) ⇝ ▸ Karimi 2020 propose CEs that respect causal structure
- **Model drift:** Bank's algorithm itself may change over time
  ⇝ Past CEs may become invalid