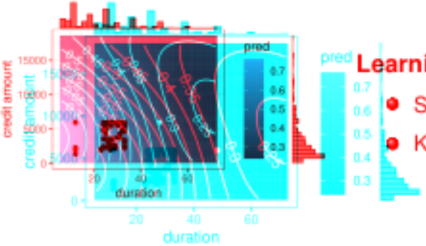
Interpretable Machine Learning

Methods & Discussion of CEs



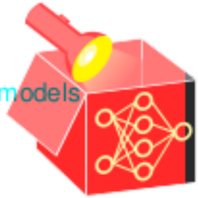
Learning goals

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

Learning goals

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

OVERVIEW OF METHODS

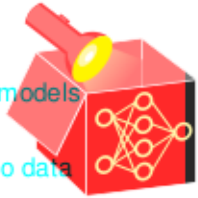


Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  ⇝ so far, all methods remain in the supervised learning paradigm

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS



Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\leadsto$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data

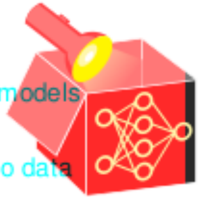
OVERVIEW OF METHODS

Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\leadsto$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS

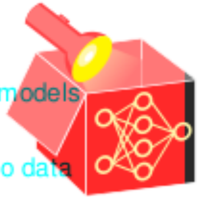


Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  ⇝ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS

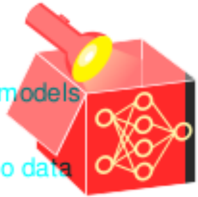


Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\leadsto$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS

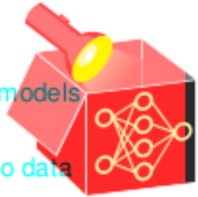


Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions ⇒ Model-agnostic and model-specific methods exist

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS

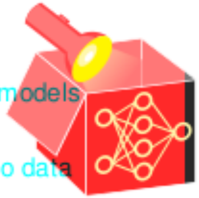


Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\leadsto$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions $\Rightarrow$ Model-agnostic and model-specific methods exist
- **Optimization tool:** Gradient-based algorithms (only for differentiable models), mixed-integer programming (only linear), or gradient-free algorithms e.g. Nelder-Mead, genetic algorithm

Interpretable Machine Learning – 1 / 7

OVERVIEW OF METHODS



Currently, multiple methods exist to calculate counterfactuals. They mainly differ in:

- **Targets:** Most methods focus on classification models, only few cover regression models
  $\leadsto$ so far, all methods remain in the supervised learning paradigm
- **Data:** Methods mainly focus on tabular data, few on visual/text data, none on audio data
- **Feature space:** Some methods can only handle numerical features, few can process mixed (numerical and discrete) feature spaces
- **Objectives:** Many methods focus on action guidance, plausibility and sparsity, few on other objectives like fairness or individual preferences
- **Model access:** Methods either require access to complete model internals, access to gradients, or only to prediction functions $\Rightarrow$ Model-agnostic and model-specific methods exist
- **Optimization tool:** Gradient-based algorithms (only for differentiable models), mixed-integer programming (only linear), or gradient-free algorithms e.g. Nelder-Mead, genetic algorithm
- **Rashomon Effect:** Many methods return a single counterfactual per run, some multiple counterfactuals, others prioritize CEs or let the user choose

FIRST OPTIMIZATION METHOD ▸ Wachter et. al (2018)

Introduced counterfactual explanations in the context of ML predictions by solving

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

MAD_j is the median absolute deviation of feature j. In each iteration, optimizers like Nelder-Mead solve the equation for $\mathbf{x}'$ and then λ is increased until a sufficiently close solution is found

This optimization problem has several shortcomings:

- We do not know how to choose λ a priori
- Due to the maximization of λ, we focus primarily on the minimization of o_p
  $\rightsquigarrow$ only if $\hat{f}(\mathbf{x}') = y'$, we focus on minimizing o_f
- Definition of o_f only covers numerical features
- Other objectives such as sparsity and plausibility of counterfactuals are neglected

MULTI-OBJECTIVE COUNTERFACTUAL EXPLANATIONS ▸ Dandl et al. (2020)

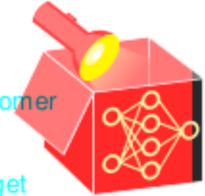
- **Multi-Objective Counterfactual Explanations (MOC):** Instead of collapsing objectives into a single objective, we could optimize all four objectives simultaneously

$$\arg\min_{\mathbf{x}'} \left(o_p(\hat{f}(\mathbf{x}'), y'), o_f(\mathbf{x}', \mathbf{x}), o_s(\mathbf{x}', \mathbf{x}), o_4(\mathbf{x}', \mathbf{X}) \right).$$

- Note that weighting parameters like λ are not necessary anymore
- Uses an adjusted multi-objective genetic algorithm (NSGA-II) to produce a set of diverse counterfactuals for mixed discrete and continuous feature spaces
- Instead of one, MOC returns multiple counterfactuals that represents different trade-offs between the objectives and are constructed to be diverse in feature space

MULTI-OBJECTIVE COUNTERFACTUAL EXPLANATIONS ▸ Dandl et al. (2020)
Interpretable Machine Learning – 3 / 7

EXAMPLE: CREDIT DATA



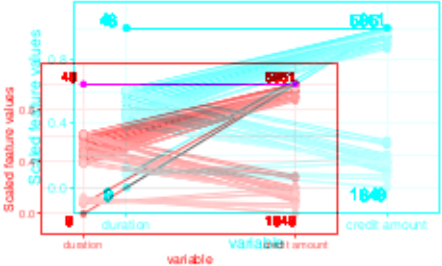
- Model: SVM with RBF kernel
- **x**: First data point of credit data with $\mathbb{P}(y = good) = 0.34$ of being a "good" customer
- Goal: Increase the probability to $[0.5, 1]$
- MOC (with default parameters) found 69 CEs after 200 iterations that met the target
- All counterfactuals proposed changes to credit duration and many of them to credit amount

Interpretable Machine Learning – 4 / 7

EXAMPLE: CREDIT DATA

Dandl et al. (2020)

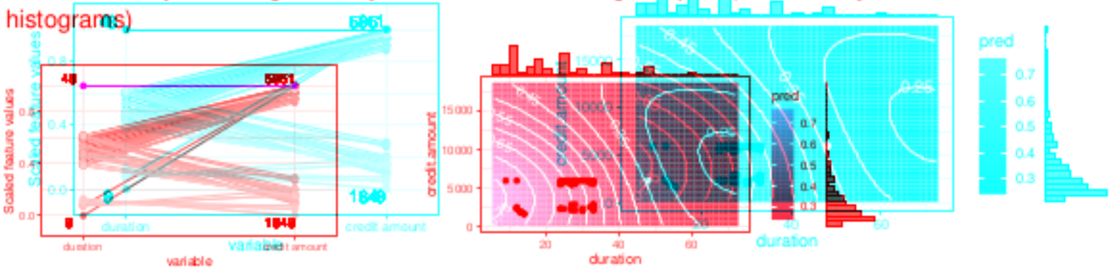
- We can visualize feature changes with a parallel plot and 2-dim surface plot
- Parallel plot reveals that all counterfactuals had values equal to or smaller than the values of $\mathbf{x}$



Parallel plot: Grey lines show feature values of CEs $\mathbf{x}'$, blue line are values of $\mathbf{x}$. Features without proposed changes are omitted. Bold numbers refer to range of numeric features.

 Dandl et al. (2020)

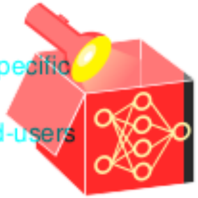
- We can visualize feature changes with a parallel plot and 2-dim surface plot
- Parallel plot reveals that all counterfactuals had values equal to or smaller than the values of **x**
- Surface plot illustrates why these feature changes are recommended
- Counterfactuals in the lower left corner seem to be in a less favorable region far from **x**, but they are in high density areas close to training samples (indicated by histograms)



Parallel plot: Grey lines show feature values of CEs $\mathbf{x}'$, blue line are values of **x**. Features without proposed changes are omitted. Bold numbers refer to range of numeric features.

Surface plot: White dot is **x**, black dots are CEs $\mathbf{x}'$. Histograms show marginal distribution of training data **X**.

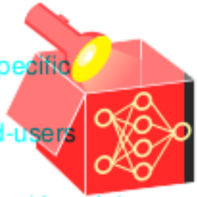
PROBLEMS, PITFALLS, & LIMITATIONS



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

Interpretable Machine Learning – 6 / 7

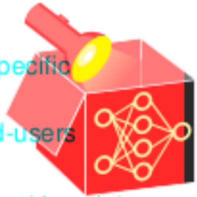
PROBLEMS, PITFALLS, & LIMITATIONS



- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in 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 can be desirable for end-users but not for data scientists searching for model bias

Interpretable Machine Learning – 6 / 7

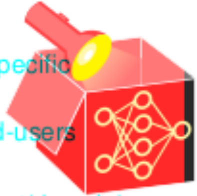
PROBLEMS, PITFALLS, & LIMITATIONS



- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in 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 can be desirable for end-users but not for data scientists searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
  ⇝ End-users need to be aware that CE provide insights into a model not the real world

Interpretable Machine Learning – 6 / 7

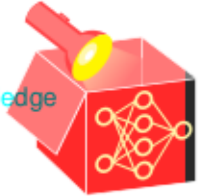
PROBLEMS, PITFALLS, & LIMITATIONS



- **Illusion of model understanding:** CEs explain ML decisions by pointing to few specific alternatives which reduces complexity, but is limited in 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 can be desirable for end-users but not for data scientists searching for model bias
- **Confusing Model and Real-World:** Model explanations are not easily transferable to reality
  ⇝ End-users need to be aware that CE provide insights into a model not the real world
- **Disclosing too much information:**
  CEs can reveal too much information about the model and help potential attackers

Interpretable Machine Learning – 6 / 7

PROBLEMS, PITFALLS, & LIMITATIONS



- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  ⇝ No perfect solution, depends on end-users computational resources and knowledge

Interpretable Machine Learning – 7 / 7

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  ⇝ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  ⇝ Counteract contestability, e.g., if ethnicity is not changed in a CE since it is not actionable, this could hide racial biases in the model

Interpretable Machine Learning – 7 / 7

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  ⇝ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  ⇝ Counteract 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 and CEs are not reliable anymore

Interpretable Machine Learning – 7 / 7

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  ⇝ No perfect solution, depends on end-users computational resources and knowledge
- **Actionability vs. fairness:** Some authors suggest to focus only on the actionability of CEs
  ⇝ Counteract 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 and CEs are not reliable anymore
- **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?

Interpretable Machine Learning – 7 / 7