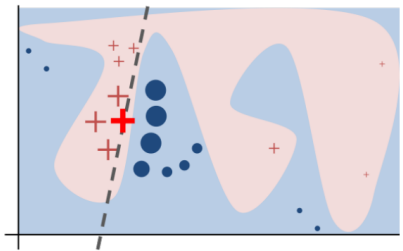
Interpretable Machine Learning

Local Explanations



Learning goals

- Understand motivation for local explanations
- Develop an intuition for possible use-cases
- Know characteristics of local explanation methods

MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular decision**
  - Understand the ML model's decisions in a **local neighborhood** of a given input (e.g., feature vector)

MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular decision**
  - Understand the ML model's decisions in a **local neighborhood** of a given input (e.g., feature vector)

- Local methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input **x**?
  - **How** does the model decide for observations that are similar to **x**?
  - **What** would the ML model have decided if **x** its values in $\mathcal{X}$ were different?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

MOTIVATION

- Purpose of local explanations:
  - Insight into the driving factors for a **particular decision**
  - Understand the ML model's decisions in a **local neighborhood** of a given input (e.g., feature vector)

- Local methods can address questions such as:
  - **Why** did the model decide to predict $\hat{y}$ for input **x**?
  - **How** does the model decide for observations that are similar to **x**?
  - **What** would the ML model have decided if **x** its values in $\mathcal{X}$ were different?
  - **Where** (in which regions in $\mathcal{X}$) does the model fail?

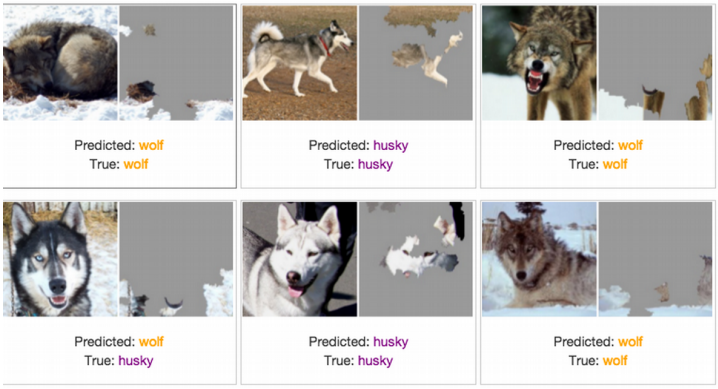
- Local methods can be particularly useful for laypersons, e.g.:
  - Why was my loan application rejected?
  - Why do I have a high predicted disease risk?

EXAMPLE: HUSKY OR WOLF? ▸ Sameer Singh 2018

- We trained a model to predict if an image shows a wolf or a husky
- Below the predictions on six test images are given
- Do you trust our predictor?

- Sometimes the ML model is wrong
- Can you guess the pattern the ML model learned to identify a wolf?

EXAMPLE: HUSKY OR WOLF? USING LIME

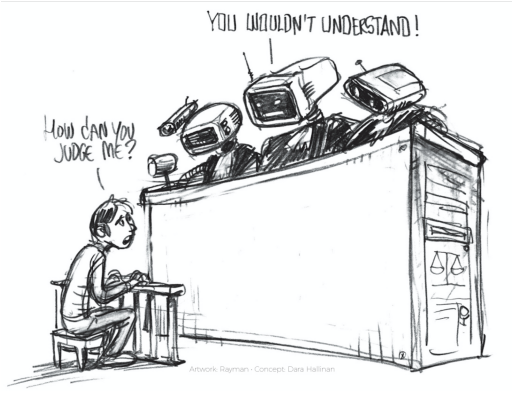


- Local explanations highlight the parts of an image which led to the prediction

$\rightsquigarrow$ our predictor is actually a snow detector

EXAMPLE: LOAN APPLICATION

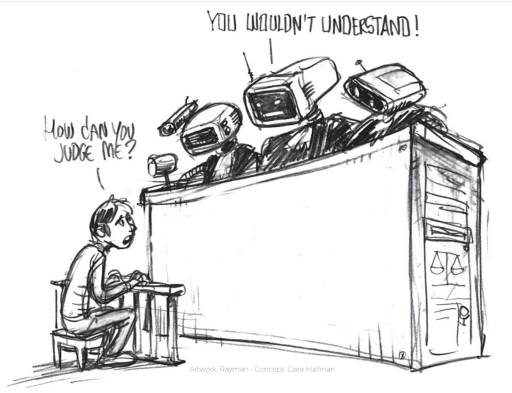
https://www.elte.hu



- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons

EXAMPLE: LOAN APPLICATION

https://www.elte.hu

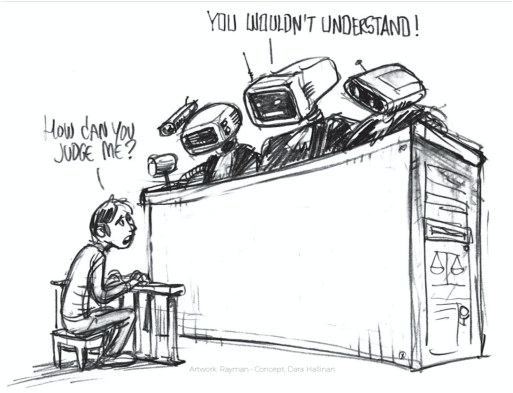


- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  > "If you were older than 21, your loan application would have been accepted."

EXAMPLE: LOAN APPLICATION

▸ https://www.elte.hu



- Imagine: You apply for a loan at an online bank and are immediately rejected without reasons
- Bank could e.g. provide a counterfactual explanation using local explanation methods:

  > "If you were older than 21, your loan application would have been accepted."

$\rightsquigarrow$ helps to understand the decision and to take actions for recourse (if req.)

CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment

CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models

CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons

CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons
  $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data

CHARACTERISTICS OF LOCAL EXPLANATIONS

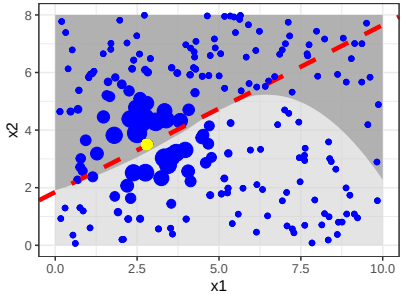
- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data
- **Methods:** Many, most prominent are counterfactual explanations, shapley values, local interpretable model-agnostic explanations (LIME), adversarial examples, single ICE curve

CHARACTERISTICS OF LOCAL EXPLANATIONS

- **Explanation scope:** Specific prediction, local environment
- **Model classes:** Model-agnostic by definition, model-specific for computational reasons $\rightsquigarrow$ very popular also for deep learning models
- **Audience:** ML modelers and laypersons
- **Data types:** Often agnostic, including tabular, image, text and audio data
- **Methods:** Many, most prominent are counterfactual explanations, shapley values, local interpretable model-agnostic explanations (LIME), adversarial examples, single ICE curve
- **Special:** Due to audience, strong interactions with social sciences and strong connections to cognitive science and neurosciences due to data types

Interpretable Machine Learning

LIME



Learning goals

- Understand motivation for LIME
- Develop a mathematical intuition

LIME

- Local Interpretable Model-agnostic Explanations (LIME) assume that even if a ML model is very complex, the local prediction can be described with a simpler model

LIME

- Local Interpretable Model-agnostic Explanations (LIME) assume that even if a ML model is very complex, the local prediction can be described with a simpler model
- LIME explains **individual** predictions of **any** black-box model by approximating the model **locally** with an interpretable model

LIME

- Local Interpretable Model-agnostic Explanations (LIME) assume that even if a ML model is very complex, the local prediction can be described with a simpler model
- LIME explains **individual** predictions of **any** black-box model by approximating the model **locally** with an interpretable model
- Called local surrogate models $\rightsquigarrow$ often inherently interpretable models such as linear models or classification/regression trees are chosen

LIME

- Local Interpretable Model-agnostic Explanations (LIME) assume that even if a ML model is very complex, the local prediction can be described with a simpler model
- LIME explains **individual** predictions of **any** black-box model by approximating the model **locally** with an interpretable model
- Called local surrogate models $\rightsquigarrow$ often inherently interpretable models such as linear models or classification/regression trees are chosen
- LIME should answer why a ML model predicted $\hat{y}$ for input $\mathbf{x}$

LIME

- Local Interpretable Model-agnostic Explanations (LIME) assume that even if a ML model is very complex, the local prediction can be described with a simpler model
- LIME explains **individual** predictions of **any** black-box model by approximating the model **locally** with an interpretable model
- Called local surrogate models $\rightsquigarrow$ often inherently interpretable models such as linear models or classification/regression trees are chosen
- LIME should answer why a ML model predicted $\hat{y}$ for input $\mathbf{x}$
- LIME is model-agnostic and can handle tabular, image and text data

LIME: CHARACTERISTICS

Definition:
LIME provides a local explanation for a black-box model $\hat{f}$ in form of a model $\hat{g} \in \mathcal{G}$ with $\mathcal{G}$ as the class of potential (interpretable) models

Model g should have two characteristics:

1. **Interpretable**: relation between the input variables and the response are easy to understand
2. **Locally faithful** / **Fidelity**: similar behavior as $\hat{f}$ in the vicinity of the obs. being predicted

Formally, we want to receive a model $\hat{g}$ with **minimal complexity and maximal local-fidelity**

MODEL COMPLEXITY

We can measure the complexity of a model $\hat{g}$ using a complexity measure $J(\hat{g})$

Example: Linear model

- Let $\mathcal{G} = \left\{g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x})\right\}$ be the class of generalized linear models
- $s(\cdot)$: identity function for linear regression or logistic sigmoid function for logistic regression

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$ could be the L_0 loss, i.e., the number of non-zero coefficients

MODEL COMPLEXITY

We can measure the complexity of a model $\hat{g}$ using a complexity measure $J(\hat{g})$

Example: Linear model

- Let $\mathcal{G} = \left\{g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x})\right\}$ be the class of generalized linear models
- $s(\cdot)$: identity function for linear regression or logistic sigmoid function for logistic regression

$\rightsquigarrow$ $J(g) = \sum_{j=1}^{p} \mathcal{I}_{\{\theta_j \neq 0\}}$ could be the L_0 loss, i.e., the number of non-zero coefficients

Example: Tree

- Let $\mathcal{G} = \left\{g : \mathcal{X} \to \mathbb{R} \mid g(\mathbf{x}) = \sum_{m=1}^{M} c_m \mathcal{I}_{\{\mathbf{x} \in Q_m\}}\right\}$ be the class of trees
  i.e., the class of additive models (e.g., constant c_m) over the leaf-rectangles Q_m

$\rightsquigarrow$ $J(g)$ could measure the number of terminal/leaf nodes

LOCAL MODEL FIDELITY

- g is locally faithful to $\hat{f}$ w.r.t. $\mathbf{x}$ if for $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ close to $\mathbf{x}$, predictions of $\hat{g}(\mathbf{z})$ are close to $\hat{f}(\mathbf{z})$
- In an optimization task: the closer $\mathbf{z}$ is to $\mathbf{x}$, the closer $\hat{g}(\mathbf{z})$ should be to $\hat{f}(\mathbf{z})$

LOCAL MODEL FIDELITY

- g is locally faithful to $\hat{f}$ w.r.t. $\mathbf{x}$ if for $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ close to $\mathbf{x}$, predictions of $\hat{g}(\mathbf{z})$ are close to $\hat{f}(\mathbf{z})$
- In an optimization task: the closer $\mathbf{z}$ is to $\mathbf{x}$, the closer $\hat{g}(\mathbf{z})$ should be to $\hat{f}(\mathbf{z})$
- Two required measures:
  1. A proximity (similarity) measure $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g. the exponential kernel:

$$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$$

  with σ as the kernel width and d as the Euclidean distance (numeric features) or the Gower distance (mixed features)

LOCAL MODEL FIDELITY

- g is locally faithful to $\hat{f}$ w.r.t. $\mathbf{x}$ if for $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ close to $\mathbf{x}$, predictions of $\hat{g}(\mathbf{z})$ are close to $\hat{f}(\mathbf{z})$
- In an optimization task: the closer $\mathbf{z}$ is to $\mathbf{x}$, the closer $\hat{g}(\mathbf{z})$ should be to $\hat{f}(\mathbf{z})$
- Two required measures:
  1. A proximity (similarity) measure $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g. the exponential kernel:

     $$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$$

     with σ as the kernel width and d as the Euclidean distance (numeric features) or the Gower distance (mixed features)
  2. A distance measure or loss function $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error

     $$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = (\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z}))^2$$

LOCAL MODEL FIDELITY

- g is locally faithful to $\hat{f}$ w.r.t. $\mathbf{x}$ if for $\mathbf{z} \in \mathcal{Z} \subseteq \mathbb{R}^p$ close to $\mathbf{x}$, predictions of $\hat{g}(\mathbf{z})$ are close to $\hat{f}(\mathbf{z})$
- In an optimization task: the closer $\mathbf{z}$ is to $\mathbf{x}$, the closer $\hat{g}(\mathbf{z})$ should be to $\hat{f}(\mathbf{z})$
- Two required measures:
  1. A proximity (similarity) measure $\phi_{\mathbf{x}}(\mathbf{z})$ between $\mathbf{z}$ and $\mathbf{x}$, e.g. the exponential kernel:

     $$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$$

     with σ as the kernel width and d as the Euclidean distance (numeric features) or the Gower distance (mixed features)
  2. A distance measure or loss function $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$, e.g. the L_2 loss/squared error

     $$L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z})) = (\hat{g}(\mathbf{z}) - \hat{f}(\mathbf{z}))^2$$

- Given points $\mathbf{z}$, we can measure local fidelity of g with respect to $\hat{f}$ in terms of a weighted loss

$$L(\hat{f}, g, \phi_{\mathbf{x}}) = \sum_{\mathbf{z} \in \mathcal{Z}} \phi_{\mathbf{x}}(\mathbf{z}) L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$$

MINIMIZATION TASK

- Optimization objective of LIME:

$$\arg\min_{g \in \mathcal{G}} L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) + J(g)$$

- In practice:
  - LIME only optimizes $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ (model-fidelity)
  - Users decide threshold on model complexity $J(g)$ beforehand
- Goal: **model-agnostic** explainer
  - $\rightsquigarrow$ optimize $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}})$ without making any assumptions about $\hat{f}$
  - $\rightsquigarrow$ learn $\hat{g}$ only approximately

LIME ALGORITHM: OUTLINE

Input:

- Pre-trained model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Model class $\mathcal{G}$ for local surrogate (to limit the complexity of the explanation)

LIME ALGORITHM: OUTLINE

Input:

- Pre-trained model $\hat{f}$
- Observation $\mathbf{x}$ whose prediction $\hat{f}(\mathbf{x})$ we want to explain
- Model class $\mathcal{G}$ for local surrogate (to limit the complexity of the explanation)

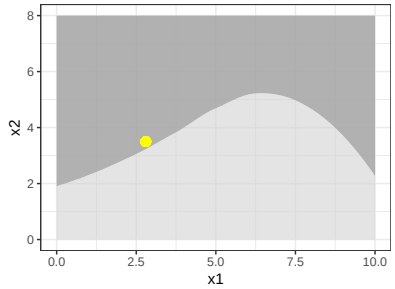
Algorithm:

1. Independently sample new points $\mathbf{z} \in \mathcal{Z}$
2. Retrieve predictions $\hat{f}(\mathbf{z})$ for obtained points $\mathbf{z}$
3. Weight $\mathbf{z} \in \mathcal{Z}$ by their proximity $\phi_{\mathbf{x}}(\mathbf{z})$
4. Train an interpretable surrogate model g on weighted data points $\mathbf{z} \in \mathcal{Z}$
   $\rightsquigarrow$ predictions $\hat{f}(\mathbf{z})$ are the target of this model
5. Return the interpretable model $\hat{g}$ as the explainer

LIME ALGORITHM: EXAMPLE

Illustration of LIME based on a classification task:

- Light/dark gray background: prediction surface of a classifier
- Yellow point: **x** to be explained
- $\mathcal{G}$: class of logistic regression models



LIME ALGORITHM: EXAMPLE (STEP 1+2: SAMPLING) ▸ Ribeiro. 2016

Strategies for sampling:

- Uniformly sample new points from the feasible feature range
- Use the training data set with or without perturbations
- Draw samples from the estimated univariate distribution of each feature
- Create an equidistant grid over the supported feature range

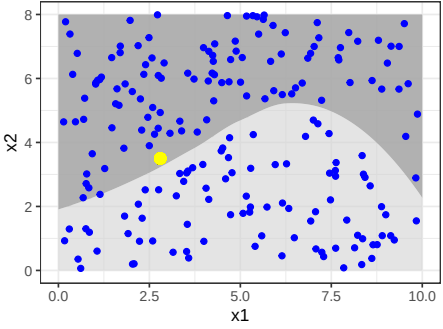


Figure: Uniformly sampled

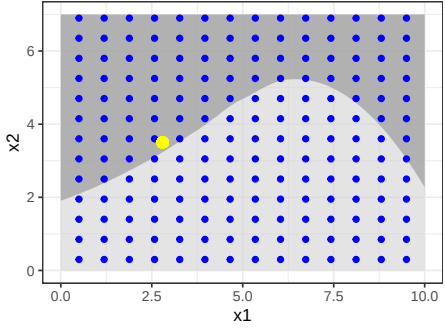
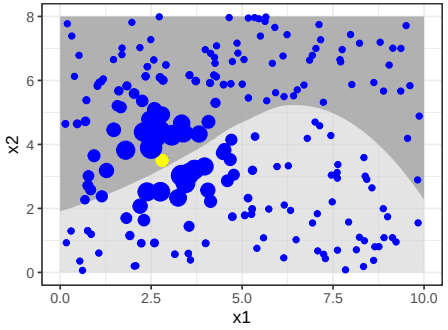


Figure: Equidistant grid

LIME ALGORITHM: EXAMPLE (STEP 3: PROXIMITY) ▸ Ribeiro. 2016

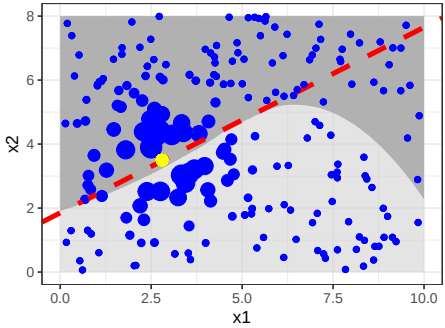
In this example, we use the exponential kernel defined on the Euclidean distance d

$$\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2).$$



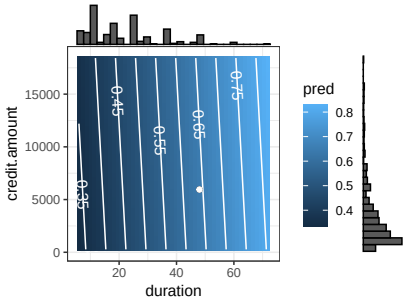
LIME ALGORITHM: EXAMPLE (STEP 4: SURROGATE) ▸ Ribeiro. 2016

In our example, we fit a **logistic regression** model (consequently, $L(\hat{f}(\mathbf{z}), \hat{g}(\mathbf{z}))$ is the Bernoulli loss)



Interpretable Machine Learning

LIME Examples



Learning goals

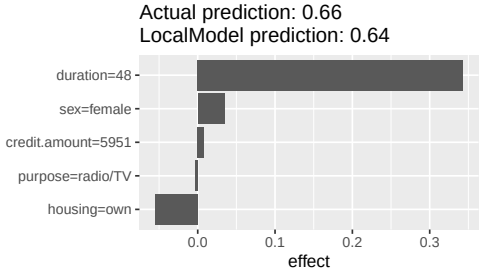
- See real-world data examples
- See application to image and text data

EXAMPLE ON CREDIT DATASET (TABULAR)

- Model: SVM with RBF kernel
- $\mathbf{x}$: first data point of the dataset with $\hat{f}_{bad}(\mathbf{x}) = 0.658$
- $\mathbf{z}$: training data $\rightsquigarrow$ weighted by the Gower proximity
- Surrogate model $\hat{g}$: L_1-regularized linear model with 5 features

age	sex	job	housing	saving	checking	credit.amount	duration	purpose
22	female	2	own	little	moderate	5951	48	radio/TV

EXAMPLE ON CREDIT DATASET (CONT'D)

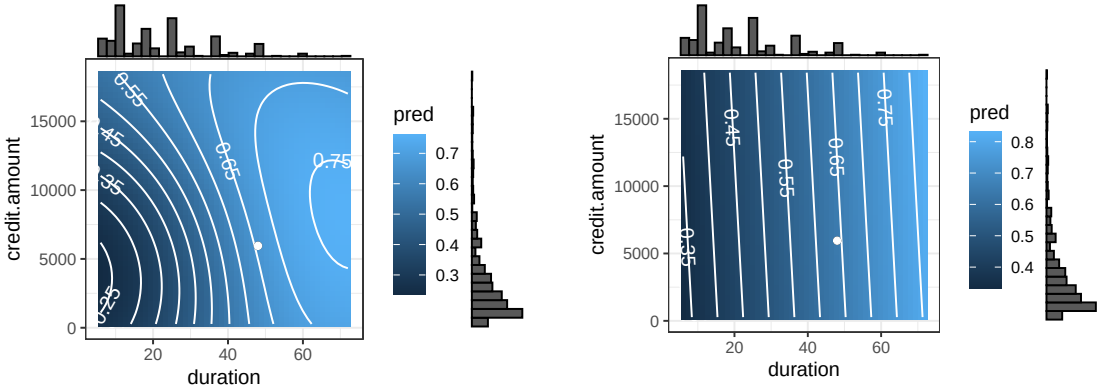


Effects of surrogate model, i.e. $\hat{\theta}^T\mathbf{x}$

- The local model prediction for $\mathbf{x}$ is $\hat{g}(\mathbf{x}) = 0.64$ vs. $\hat{f}(\mathbf{x}) = 0.658$
- $\hat{g}$ has a local fidelity of $L(\hat{f}, \hat{g}, \phi_{\mathbf{x}}) = 4.82$ with $\phi_{\mathbf{x}}(\mathbf{z})$ as the Gower proximity and $L(\hat{f}_{bad}(\mathbf{z}), g(\mathbf{z}))$ as the euclidean distance

EXAMPLE ON CREDIT DATASET (CONT'D)

- 2-dim ICE plots (aka. prediction surface plot) of credit amount and duration show how the surrogate model g linearly approximates the previously nonlinear prediction surface of $\hat{f}_{bad}$



2-dim ICE plot of $\hat{f}_{bad}$ (**left**) and surrogate g (**right**) for features duration and credit amount. The white dot is **x**. The histograms display the marginal distribution of the training data **X**.

LIME FOR TEXT DATA ▸ Shen, Ian, (2019)

LIME can also be applied to text data:

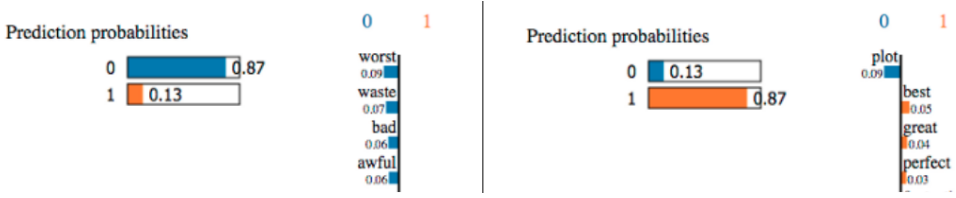
- Raw text representations:
  - Binary vector indicating the presence or absence of a word
  - A vector of word counts
- Examples for *“This text is the first text."* and *“Finally, this is the last one."*:

this	text	is	the	first	finally	last	one
1	2	1	1	1	0	0	0
1	0	1	1	0	1	1	1

- **Sampling**: Randomly set the entry of individual words to 0; equal to removing all occurrences of this word in the text.
- **Proximity**: Exponential kernel with cosine distance.
  - Neglects words that do not occur in both texts
  - Measures the distance irrespective of the text size

LIME FOR TEXT DATA (CONT'D) ▸ Shen, Ian, (2019)

- Random forest classifier labeling movie reviews from IMDB
  - 0: negative
  - 1: positive
- Surrogate model is a sparse linear model

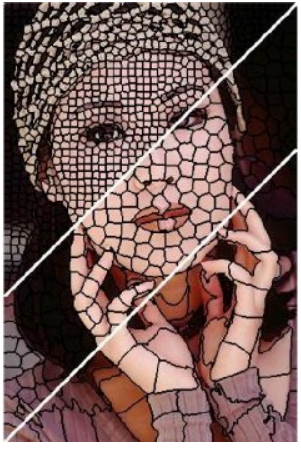


Words like "worst" or "waste" indicate negative review while words like "best" or "great" indicate positive review

LIME FOR IMAGE DATA

LIME also works for image data:

- **Idea**: Each obs. is represented by a binary vector indicating the presence or absence of superpixels ▸ Achanta et al. 2012
- Superpixels are interconnected pixels with similar colors (absence of a single pixel might not have a (strong) effect on the prediction)
- **Warning**: Size of superpixels needs to be determined before the segmentation takes place
- **Sampling**: Randomly switching some of the super pixels “off", i.e., by coloring some superpixels uniformly



Example for superpixels of different sizes

LIME FOR IMAGE DATA (CONT'D) ▸ Ribeiro. 2016

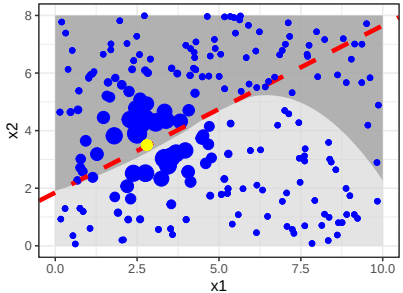
- Explaining prediction of pre-trained inception neural network classifier
- **Sampling**: Graying out all superpixels besides 10 superpixels
- **Surrogate**: Locally weighted sparse linear models
- **Proximity**: Exponential kernel with euclidean distance

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Top 3 classes predicted

Interpretable Machine Learning

LIME Pitfalls



Learning goals

- Learn why LIME should be used with caution
- Possible pitfalls of LIME

LIME PITFALLS

- LIME is one of the best known interpretable ML methods
  $\rightsquigarrow$ But several papers caution to be careful in practice
- Problems can occur on different levels which are described subsequently:
  - Sampling procedure (extrapolation)
  - Definition of locality (sensitivity)
  - Scope of feature effects (local vs. global)
  - Faithfulness (trade-off with sparsity)
  - Surrogate model (hiding biases, robustness)
  - Definition of superpixels in case of image data (sensitivity)

PITFALL: SAMPLING

- **Pitfall**: Common sampling strategies for $\mathbf{z} \in Z$ do not account for correlation between features
- **Implication**: Unlikely data points might be used to learn local explanation models

PITFALL: SAMPLING

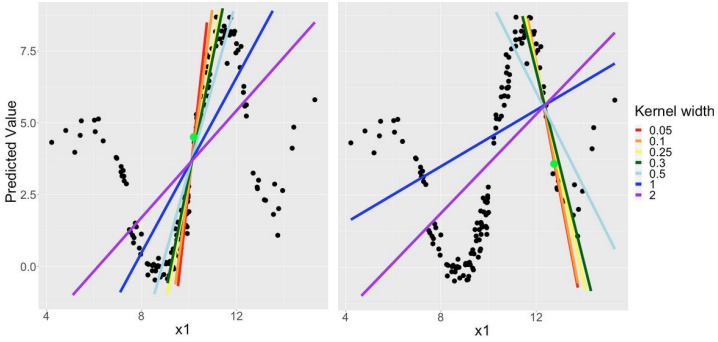
- **Pitfall**: Common sampling strategies for $\mathbf{z} \in Z$ do not account for correlation between features
- **Implication**: Unlikely data points might be used to learn local explanation models
- **Solution I**: Conditional sampling
  $\rightsquigarrow$ particularly difficult for high dimensional or mixed feature spaces
- **Solution II**: Use training data to fit surrogate model
  $\rightsquigarrow$ only works well with enough data near $\mathbf{x}$

LIME PITFALL: LOCALITY

- **Pitfall**: Difficult to define locality (= how samples are weighted locally)
  $\rightsquigarrow$ Strongly affects local model, but there is no automatic procedure for choosing neighborhood
- Originally, an exponential kernel as proximity measure between $\mathbf{x}$ and $\mathbf{z}$ was proposed:
  $\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ where d is a distance measure and σ is the kernel width

LIME PITFALL: LOCALITY

- **Pitfall**: Difficult to define locality (= how samples are weighted locally)
  $\rightsquigarrow$ Strongly affects local model, but there is no automatic procedure for choosing neighborhood
- Originally, an exponential kernel as proximity measure between **x** and **z** was proposed:
  $\phi_{\mathbf{x}}(\mathbf{z}) = exp(-d(\mathbf{x}, \mathbf{z})^2/\sigma^2)$ where d is a distance measure and σ is the kernel width



- Surrogate models for 2 obs. (green points) for same model with one feature x_1
- Each line refers to a linear surrogate model with different kernel width
- Right figure: larger kernel widths influence lines more

LIME PITFALL: LOCALITY ▸ Kopper et al. 2019

- **Solution I**: Kernel width strongly interacts with locality:
  - Large kernel width leads to interaction with points further away (unwanted)
  - Small kernel width leads to small neighborhood
    ⇝ risk of few data points
    ⇝ potentially fitting more noise

LIME PITFALL: LOCALITY ▸ Kopper et al. 2019

- **Solution I**: Kernel width strongly interacts with locality:
  - Large kernel width leads to interaction with points further away (unwanted)
  - Small kernel width leads to small neighborhood
    $\rightsquigarrow$ risk of few data points
    $\rightsquigarrow$ potentially fitting more noise
- **Solution II**: Use Gower distance where no kernel width needs to be specified
  - **Problem**: data points far away receive weight > 0
    $\rightsquigarrow$ resulting explanations are rather global than local surrogates

PITFALL: LOCAL VS. GLOBAL FEATURES ▸ Laugel et al. 2018

- **Problem**:
  By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)

PITFALL: LOCAL VS. GLOBAL FEATURES ▸ Laugel et al. 2018

- **Problem**:
  By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)
- **Implication**:
  - Some features influence the **global** shape of the black-box model
  - Other **local** features impact predictions only in smaller regions of $\mathcal{X}$

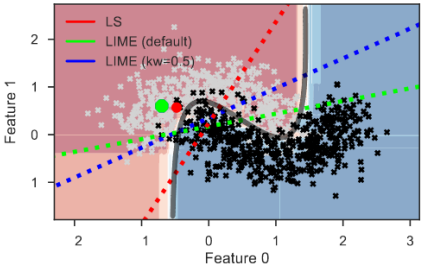
PITFALL: LOCAL VS. GLOBAL FEATURES ▸ Laugel et al. 2018

- **Problem**:
  By sampling obs. for the surrogate model from the whole input space, the influence of local features might be hidden in favor of features with global influence (even for small kernel width)
- **Implication**:
  - Some features influence the **global** shape of the black-box model
  - Other **local** features impact predictions only in smaller regions of $\mathcal{X}$
- **Example**: Decision trees
  $\Rightarrow$ Split features close to root have a more global influence than the ones close to leaves

PITFALL: LOCAL VS. GLOBAL FEATURES – EXAMPLE

▸ Laugel et al. 2018

- Binary classification model
- Right figure:
  - Black and grey crosses: training data
  - Green dot: Obs. to be explained
  - Background color: Classification of random forest
  - Dark grey curve: Classifier's decision boundary
  - Dotted lines: Local decision boundary
- **Observation:** Decision boundaries of LIME with different kernels (blue and green lines) do not match the direction of the local decision boundary (which appears steeper)

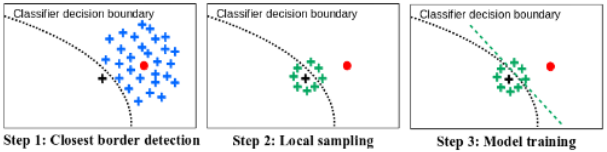


Half-moons dataset

PITFALL: LOCAL VS. GLOBAL FEATURES – SOLUTION

Laugel et al. 2018

- **Solution**: Find closest point to **x** from other class and sample new points **z** around it for higher local accuracy

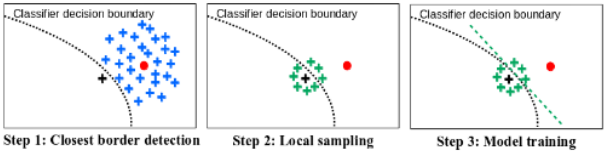


Example: x (red point), closest point from other class (black cross)

PITFALL: LOCAL VS. GLOBAL FEATURES – SOLUTION

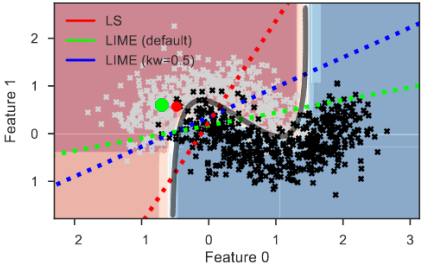
Laugel et al. 2018

- **Solution**: Find closest point to **x** from other class and sample new points **z** around it for higher local accuracy



Example: x (red point), closest point from other class (black cross)

- Red dot (right figure): Closest point from other class
- Red line: Local surrogate (LS) method Laugel et al. 2018
  $\rightsquigarrow$ better approximates the local direction of the decision boundary



Half-moons dataset

PITFALL: FAITHFULNESS

- **Problem**: Trade-off between local fidelity vs. sparsity
- **Observation I**: Low fidelity $\rightsquigarrow$ unreliable explanations
- **Observation II**: High fidelity requires complex models $\rightsquigarrow$ difficult to interpret surrogate model

PITFALL: FAITHFULNESS

- **Problem**: Trade-off between local fidelity vs. sparsity
- **Observation I**: Low fidelity $\rightsquigarrow$ unreliable explanations
- **Observation II**: High fidelity requires complex models $\rightsquigarrow$ difficult to interpret surrogate model
- **Example: Credit data**
  - Original prediction by random forest for one data point $\mathbf{x}$:

$$\hat{f}(\mathbf{x}) = \hat{\mathbb{P}}(y = 1 \mid \mathbf{x}) = 0.143$$

  - Linear model with only three selected features (`age`, `checking.account`, `duration`):

$$g_{lm}(\mathbf{x}) = \hat{\theta}_0 + \hat{\theta}_1 x_{age} + \hat{\theta}_2 x_{checking.account} + \hat{\theta}_3 x_{duration} = 0.283$$

  - Generalized additive model (with all 9 features) is more complex:

$$g_{gam}(\mathbf{x}) = \hat{\theta}_0 + f_{age}(x_{age}) + f_{checking.account}(x_{checking.account}) + f_{duration}(x_{duration}) + \cdots = 0.148$$

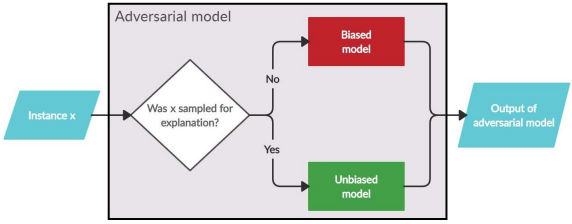
PITFALL: HIDING BIASES ▸ Slack et al. 2020

- **Problem**: Developer could manipulate their model to hide biases
- **Observation**: LIME can sample out-of-distribution points (extrapolation)

PITFALL: HIDING BIASES

▸ Slack et al. 2020

- **Problem**: Developer could manipulate their model to hide biases
- **Observation**: LIME can sample out-of-distribution points (extrapolation)
- **Attack** with adversarial model:
  1. classifier to discriminate between in-distribution and out-of-distribution data points
  2. for in-distribution points, use the original (biased) model
  3. for out-of-distribution points produced for local explanation, use an unbiased model
  - ⇝ LIME samples out-of-distribution points and uses the unbiased model for local explanation
  - ⇝ this hides the bias of the true model

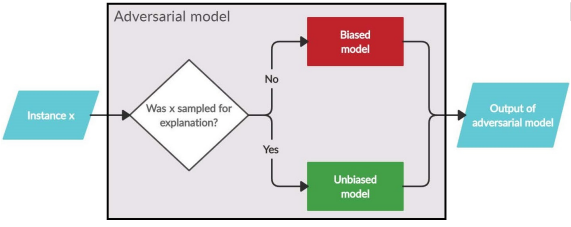


▸ Vres, Sikonja (2021)

PITFALL: HIDING BIASES

▸ Slack et al. 2020

- **Problem**: Developer could manipulate their model to hide biases
- **Observation**: LIME can sample out-of-distribution points (extrapolation)
- **Attack** with adversarial model:
  1. classifier to discriminate between in-distribution and out-of-distribution data points
  2. for in-distribution points, use the original (biased) model
  3. for out-of-distribution points produced for local explanation, use an unbiased model
  - ⇝ LIME samples out-of-distribution points and uses the unbiased model for local explanation
  - ⇝ this hides the bias of the true model



▸ Vres, Sikonja (2021)

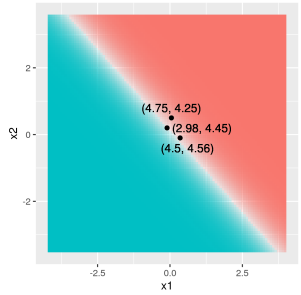
Example: Not using ‘gender‘ to approve a loan

- biased model trained on features correlated with ‘gender‘ (e.g. duration of parental leave)
  ⇝ used to make biased / unfair predictions
- unbiased model trained on features uncorrelated with ‘gender‘
  ⇝ used to produce explanations based on unbiased predictions to hide bias

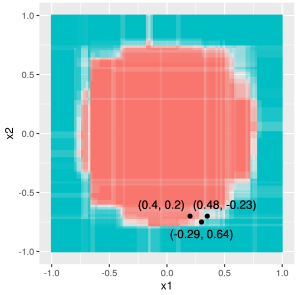
PITFALL: ROBUSTNESS

▸ Alvarez-Melis, D., & Jaakkola, T. 2018

- **Problem**: Instability of explanations
- **Observation**: Explanations of two very close points could vary greatly
  - ⇝ can happen because of other sampled data points **z**



Linear prediction task (logistic regression). Linear surrogate returns similar coefficients for similar points.



Circular prediction task (random forest). Linear surrogate returns different coefficients for similar points.

PITFALL: DEFINITION OF SUPERPIXELS

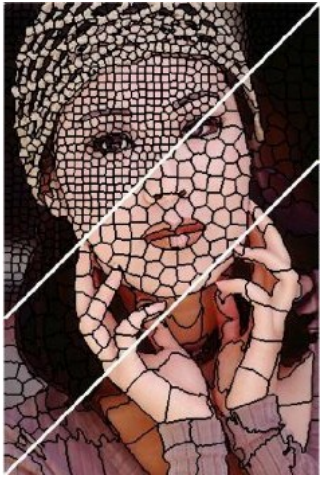
▸ Achanta et al. 2012

- **Problem**: Instability because of specification of superpixels for image data
- **Observation**: Multiple specification of superpixels exist, influencing both the shape and size

PITFALL: DEFINITION OF SUPERPIXELS

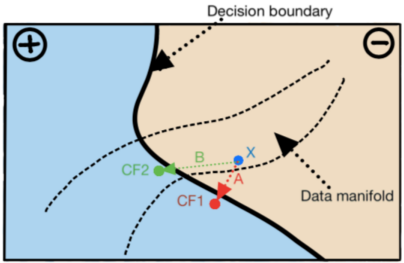
▸ Achanta et al. 2012

- **Problem**: Instability because of specification of superpixels for image data
- **Observation**: Multiple specification of superpixels exist, influencing both the shape and size
- **Implication**: The specification of superpixel has a large influence on the explanations
- **Attack**: Change superpixels as part of an adversarial attack $\rightsquigarrow$ changed explanation



Interpretable Machine Learning

Counterfactual Explanations

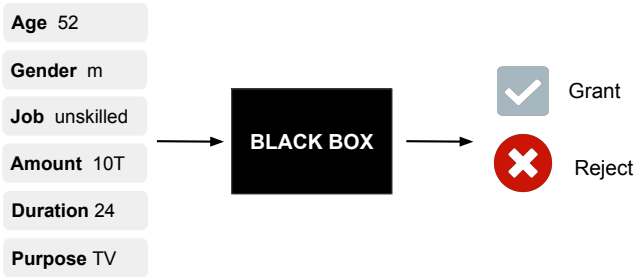


Learning goals

- Understand the motivation behind CEs
- See the mathematical foundation of CEs

EXAMPLE: CREDIT RISK APPLICATION

- **x**: customer and credit information
- *y*: grant or reject credit

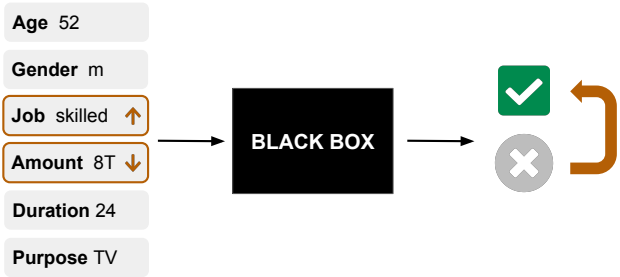


Questions:

- Why was the credit rejected?
- Is it a fair decision?
- **How should x be changed so that the credit is accepted?**

EXAMPLE: CREDIT RISK APPLICATION

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



“If the person was more skilled and the credit amount had been reduced to $8.000, the credit would have been granted."

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\rightsquigarrow$ Useful if there is a chance to change the input features (e.g., by changing behaviour)

COUNTERFACTUAL EXPLANATIONS: MAIN IDEA

- Counterfactual explanations == counterfactuals == CEs
- Explain particular predictions of an ML model by presenting an alternative input whose prediction equals a desired outcome
- Represent **close neighbors** of a data point we are interested in, but belonging to the **desired outcome**
- Reveal which minimal changes to the input are sufficient to receive a different outcome
  $\rightsquigarrow$ Useful if there is a chance to change the input features (e.g., by changing behaviour)
- The targeted audience of CEs are often end-users

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

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

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“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.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“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.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“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.

AIMS & ROLES

CEs can serve various purposes, the user can decide what to learn from them. For example:

“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.

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If *S* was the case, *Q* would have been the case."

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where S had been the case, Q would have been the case

PHILOSOPHICAL BASIS

Counterfactuals have a long-standing tradition in analytic philosophy
$\rightsquigarrow$ Accoding to ▸ Lewis (1973), a **counterfactual conditional** is a statement of the form:

"If S was the case, Q would have been the case."

- S is an event that must relate to a past event that didn't occur
  $\rightsquigarrow$ counterfactuals run **contrary** to the **facts**
- Above statement is true, if in all possible worlds most similar to the actual world where S had been the case, Q would have been the case
- A world is similar to another if laws are maximally preserved between the worlds and only a few facts are changed

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by \$20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant

PHILOSOPHICAL BASIS

- Counterfactuals have largely been studied to explain causal dependence
- Causal dependence underlies the explanatory power
  $\rightsquigarrow$ good CEs point to critical causal factors that drove the algorithmic decision
- If maximal closeness is relaxed, causally irrelevant factors can become part of the explanation
  $\rightsquigarrow$ e.g., decreasing loan amount by $20.000 and being one year older is recommended by the explainer although only loan amount might be causally relevant
- CEs are often contrastive, i.e., they explain a decision by referring to an alternative outcome
  $\rightsquigarrow$ e.g., if the loan applicant was 30 instead of 60 years old, the approved loan would have been over $100.000 instead of $40.000

MATHEMATICAL PERSPECTIVE

Terminology:

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

MATHEMATICAL PERSPECTIVE

Terminology:

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

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint $\mathbf{x}$

MATHEMATICAL PERSPECTIVE

Terminology:

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

A **valid** counterfactual $\mathbf{x}'$ is a datapoint:

1. whose prediction $\hat{f}(\mathbf{x}')$ is equal to the desired prediction y'
2. that is maximally close to the original datapoint **x**

Reformulate these two objectives (denoted by o_1 and o_2) as optimization problem:

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x})$$

- λ_1 and λ_2 balance the two objectives
- Choice of o_p (distance on prediction space) and of o_f (distance on feature space) is crucial

MATHEMATICAL PERSPECTIVE

▸ Dandl et al. (2020)

- Regression: o_p could be the L_1-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L_1-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$

MATHEMATICAL PERSPECTIVE

▸ Dandl et al. (2020)

- Regression: o_p could be the L_1-distance $o_p(\hat{f}(\mathbf{x}'), y') = |\hat{f}(\mathbf{x}') - y'|$
- Classification: L_1-distance for scores and 0-1 Loss for labels, e.g., $o_p(\hat{f}(\mathbf{x}'), y') = \mathcal{I}_{\{\hat{f}(\mathbf{x}') \neq y'\}}$
- o_f could be the Gower distance (suitable for mixed feature space):

$$o_f(\mathbf{x}', \mathbf{x}) = d_G(\mathbf{x}', \mathbf{x}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x_j', x_j) \in [0, 1]$$

The value of δ_G depends on the feature type (numerical or categorical):

$$\delta_G(x_j', x_j) = \begin{cases} \frac{1}{\widehat{R}_j} |x_j' - x_j| & \text{if } x_j \text{ is numerical} \\ \mathcal{I}_{\{x_j' \neq x_j\}} & \text{if } x_j \text{ is categorical} \end{cases}$$

with $\widehat{R}_j$ as the value range of feature j in the training dataset (to ensure that $\delta_G(x_j', x_j) \in [0, 1]$)

FURTHER OBJECTIVES

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

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**

FURTHER OBJECTIVES

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

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this

FURTHER OBJECTIVES

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

Sparsity:

- End-users often prefer short over long explanations
  $\rightsquigarrow$ counterfactuals should be **sparse**
- Objective o_f can take the number of changed features into account (but does not have to)
  $\rightsquigarrow$ e.g., the L_0- and the L_1-norm (similar to LASSO) can do this
- Independently from o_f, sparsity in the changes can be additionally considered by another objective that counts the number of changed features via the L0-norm:

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

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values

FURTHER OBJECTIVES

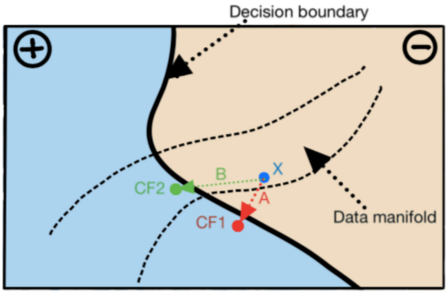
Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$

FURTHER OBJECTIVES

Plausibility:

- CEs should suggest plausible alternatives
  $\rightsquigarrow$ e.g., not plausible to suggest to raise your income and get unemployed at the same time
- CEs should be realistic and adhere to data manifold or originate from distribution of $\mathcal{X}$
  $\rightsquigarrow$ avoid unrealistic combinations of feature values
- Estimating joint distribution of training data is complex, especially for mixed feature spaces
  $\rightsquigarrow$ Proxy: ensure that $\mathbf{x}'$ is close to training data $\mathbf{X}$



Example from ▸ Verma et al. (2020)

- Two possible paths for **x**, originally classified to ⊖
- Two valid CEs in class ⊕: CF1 and CF2
- Path **A** for CF1 is shorter
- Path **B** for CF2 is longer but adheres to data manifold

FURTHER OBJECTIVES

To ensure plausibility, o_4 could, e.g., be the Gower distance of $\mathbf{x}'$ to its nearest data point of the training dataset which we denote $\mathbf{x}^{[1]}$:

$$o_4(\mathbf{x}', \mathbf{X}) = d_G(\mathbf{x}', \mathbf{x}^{[1]}) = \frac{1}{p} \sum_{j=1}^{p} \delta_G(x'_j, x_j^{[1]})$$

We can extend the previous optimization problem by adding o_s (for sparsity) and o_4 (for plausibility):

$$\arg\min_{\mathbf{x}'} \lambda_1 o_p(\hat{f}(\mathbf{x}'), y') + \lambda_2 o_f(\mathbf{x}', \mathbf{x}) + \lambda_3 o_s(\mathbf{x}', \mathbf{x}) + \lambda_4 o_4(\mathbf{x}', \mathbf{X})$$

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

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 a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

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 a given $\mathbf{x}$ (but: time and human processing capacity is limited)
- Focus on one or few CEs (but: by which criterion should they be selected?)

Note:

- As the model is generally non-linear, inconsistent and diverse CEs can arise
  e.g. suggesting either an increase or decrease in credit duration (confuses the explainee)
- How to deal with the Rashomon effect is considered an open problem in IML

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan

REMARKS: MODEL OR REAL-WORLD

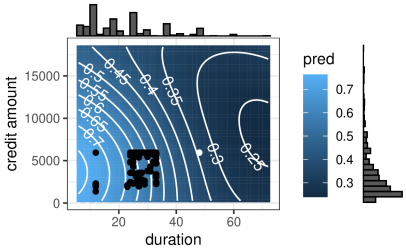
- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ ▸ Karimi et al. (2020) avoid this by considering causal dependencies between features

REMARKS: MODEL OR REAL-WORLD

- Most CEs provide explanations of model predictions, but CEs might appear to explain the real-world for end-users
  $\rightsquigarrow$ Transfer of model explanations to explain real-world is generally not permitted
- Consider a CE that proposes to increase the feature age by 5 to obtain the loan
  $\rightsquigarrow$ a loan applicant takes this information and applies 5 years later for the loan
- However, by then, many other feature values might have changed
  $\rightsquigarrow$ not only age, also other causally dependent features e.g. job status might have changed
  $\rightsquigarrow$ ▸ Karimi et al. (2020) avoid this by considering causal dependencies between features
- Also, the bank's algorithm might change and previous CEs are not applicable anymore

Interpretable Machine Learning

Methods & Discussion of Counterfactual Explanations



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
  $\rightsquigarrow$ so far, all methods remain in the supervised learning paradigm

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

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

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

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 $\Rightarrow$ Model-agnostic and model-specific methods exist

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 $\Rightarrow$ 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

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 $\Rightarrow$ 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

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

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

EXAMPLE: CREDIT DATA

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

EXAMPLE: CREDIT DATA

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

Original observation:

age	sex	job	housing	saving	checking	credit.amount	duration	purpose	$\hat{y}$
22	female	2	own	little	moderate	5951	48	radio/TV	0.34

EXAMPLE: CREDIT DATA

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

Original observation:

age	sex	job	housing	saving	checking	credit.amount	duration	purpose	$\hat{y}$
22	female	2	own	little	moderate	5951	48	radio/TV	0.34

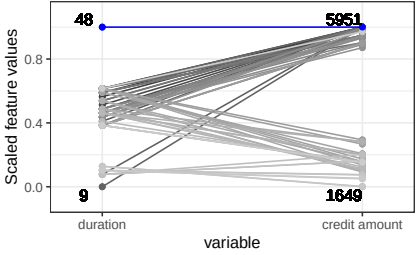
Selected counterfactuals:

age	sex	job	housing	saving	checking	credit.amount	duration	purpose	$\hat{y}$
22	female	2	own	little	moderate	5951	30	radio/TV	0.54
22	female	2	own	little	moderate	2656	24	radio/TV	0.63
22	female	2	own	little	moderate	1454	12	radio/TV	0.72

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 **x**

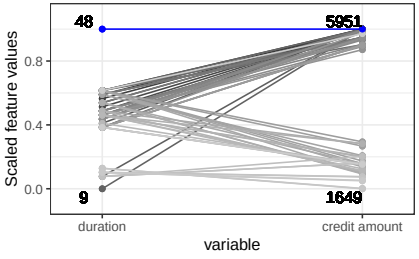


Parallel plot: Grey lines show feature values of CEs $\mathbf{x}'$, blue line are values of **x**. Features without proposed changes are omitted. Bold numbers refer to range of numeric features.

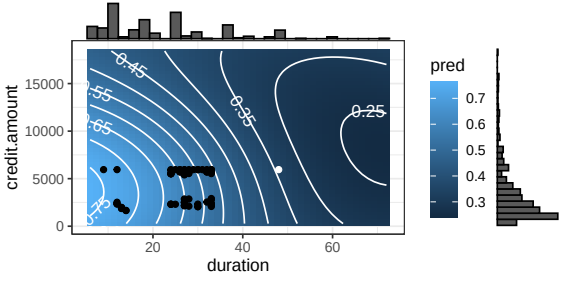
EXAMPLE: CREDIT DATA

▸ 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 **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
  $\rightsquigarrow$ 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 which reduces complexity, but is limited in 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 can be desirable for end-users but not for data scientists searching for model bias

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
  $\rightsquigarrow$ 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 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
  $\rightsquigarrow$ End-users need to be aware that CE provide insights into a model not the real world

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
  $\rightsquigarrow$ 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 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
  $\rightsquigarrow$ 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

PROBLEMS, PITFALLS, & LIMITATIONS

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

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ 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

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ in reality this assumption is often violated and CEs are not reliable anymore

PROBLEMS, PITFALLS, & LIMITATIONS

- **Rashomon effect:** One, few, all? Which CEs should be shown to the end-user?
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ 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
  $\rightsquigarrow$ how faithful are CEs to the models underlying mechanism?