A tour of Mathematical Optimization Models for Group Counterfactual Explanations

MIP2024 Workshop, University of Kentucky, Lexington
June 4, 2024

Dolores Romero Morales
Copenhagen Business School
E: drm.eco@cbs.dk **H**: doloresromero.com **T**: @DoloresRomeroM

This project has received funding from the European Union's Horizon 2020 research and Innovation programme under the Marie Skłodowska-Curie grant agreement No. 822214

Outline

- Introduction
- On Group Counterfactual Analysis
- Counterfactual Analysis Beyond Machine Learning
- Conclusions

Outline

- Introduction
- On Group Counterfactual Analysis
- Counterfactual Analysis Beyond Machine Learning
- Conclusions

When training a machine learning model,
accuracy of its predictions matters, as does its **interpretability/explainability**
(Rudin et al., 2022; European Commission, 2020; Panigutti et al., 2023)

Interpretability in Machine Learning

E.g., optimal trees, see our recent review

Home > TOP > Article

Mathematical optimization in classification and regression trees

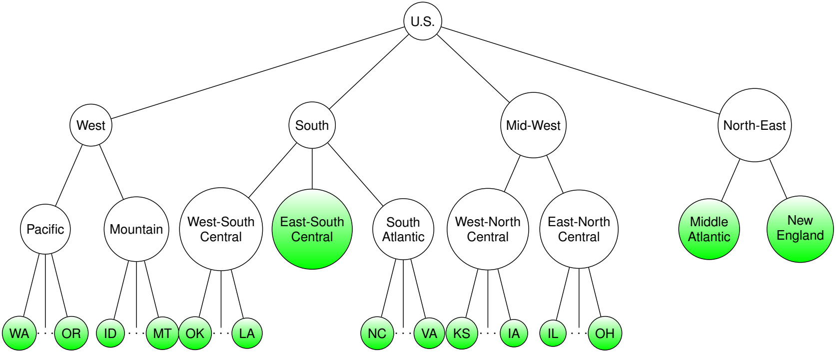
Original Paper | Open access | Published: 17 March 2021

Volume 29, pages 5–33, (2021) Cite this article

Emilio Carrizosa, Cristina Molero-Río & Dolores Romero Morales

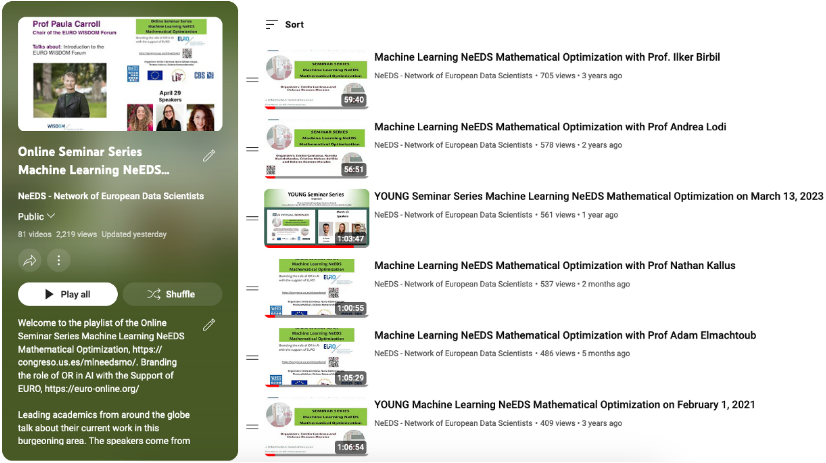
Interpretability in Machine Learning

Sparse models, e.g., Carrizosa et al. (2022) for categorical variables



Interpretability in Machine Learning

and many more at the playlist of the Online Seminar Series ML NeEDS MO



Explainability in Binary Classification

Explainability in Binary Classification

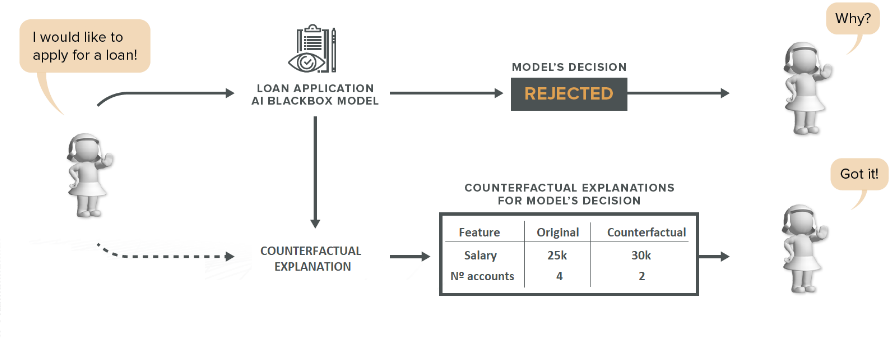
Wlog, we assume that we have a **binary classification problem** on $\mathcal{X} \subset \mathbb{R}^J$ with classes, '+1' and '-1'. The positive class, '+1', implies something good for the individual, e.g., getting a loan, social benefits or parole.

Suppose we have a classifier and an individual $\boldsymbol{x}^0$ classified as '-1', and we want to give insights on how to change the features to be classified as '+1'.

Explainability in Binary Classification

Wlog, we assume that we have a **binary classification problem** on $\mathcal{X} \subset \mathbb{R}^J$ with classes, '+1' and '-1'. The positive class, '+1', implies something good for the individual, e.g., getting a loan, social benefits or parole.

Suppose we have a classifier and an individual $\boldsymbol{x}^0$ classified as '-1', and we want to give insights on how to change the features to be classified as '+1'.



Your loan has been denied. Had your salary been 30k instead of 25k and had you had 2 accounts open instead of 4, your loan would have been accepted (Martens and Provost, 2014; Wachter et al., 2017)

Counterfactual explanations

- We are given a **probabilistic** classifier $P: \mathcal{X} \to [0,1]$, $P(\boldsymbol{x})$: probability of belonging to class +1, and
- $\boldsymbol{x}_0 \in \mathcal{X}$,
- the goal is to find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \subset \mathcal{X}$) **with minimum cost** $C(\boldsymbol{x}, \boldsymbol{x}^0)$ that cause $\boldsymbol{x}^0$ to increase the probability $P(\boldsymbol{x}^0)$ to $P(\boldsymbol{x})$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \nu \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}, \boldsymbol{x}^0), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (1-\lambda)C(\boldsymbol{x}, \boldsymbol{x}^0) - \lambda P(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array}$$

Counterfactual explanations

- We are given a **probabilistic** classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class +1, and
- $\boldsymbol{x}_0 \in \mathcal{X}$,
- the goal is to find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \subset \mathcal{X}$) **with minimum cost** $C(\boldsymbol{x}, \boldsymbol{x}^0)$ that cause $\boldsymbol{x}^0$ to increase the probability $P(\boldsymbol{x}^0)$ to $P(\boldsymbol{x})$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \nu \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}, \boldsymbol{x}^0), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (1-\lambda)C(\boldsymbol{x}, \boldsymbol{x}^0) - \lambda P(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array}$$

Counterfactual explanations

- We are given a **probabilistic** classifier $P: \mathcal{X} \rightarrow [0,1]$, $P(\boldsymbol{x})$: probability of belonging to class +1, and
- $\boldsymbol{x}_0 \in \mathcal{X}$,
- the goal is to find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \subset \mathcal{X}$) **with minimum cost** $C(\boldsymbol{x}, \boldsymbol{x}^0)$ that cause $\boldsymbol{x}^0$ to increase the probability $P(\boldsymbol{x}^0)$ to $P(\boldsymbol{x})$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \nu \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}, \boldsymbol{x}^0), -P(\boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (1-\lambda)C(\boldsymbol{x}, \boldsymbol{x}^0) - \lambda P(\boldsymbol{x}) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0) \end{array}$$

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost

- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost

- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - $\text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - $\text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - $\text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - $\text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost

- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - $\text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - $\text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - $\text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - $\text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations

- $\mathcal{X}$
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)
- $\mathcal{X}(\boldsymbol{x}^0)$
  - Points from some training set $\longrightarrow$ **discrete optimization models**
  - Synthetic data $\longrightarrow$ **(mixed integer) nonlinear optimization models**
- If $P(\boldsymbol{x}) = \varphi(f(\boldsymbol{x}))$ and $\varphi \uparrow$, then

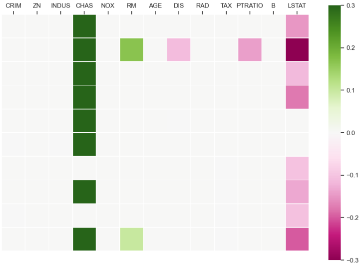
$$P(\boldsymbol{x}) \geq \nu \iff f(\boldsymbol{x}) \geq \varphi^{-1}(\nu)$$

  These are known as **score-based classifiers**, e.g., LR, SVM, NN, RF, XGBoost
- $C(\boldsymbol{x}, \boldsymbol{x}^0) = \text{Dissimilarity}(\boldsymbol{x}, \boldsymbol{x}^0) + \lambda_c \text{Complexity}(\boldsymbol{x}, \boldsymbol{x}^0)$
  - Dissimilarity$(\boldsymbol{x}, \boldsymbol{x}^0)$ is usually modeled with ℓ_p norms, but need to extend, e.g., to asymmetric gauges as in Carrizosa et al. (2024a) for asymmetric costs (Karimi et al., 2021). Also, embeddings may be needed for more complex data
  - Complexity$(\boldsymbol{x}, \boldsymbol{x}^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)

Counterfactual explanations for Logistic Regression

$$
\begin{aligned}
\min_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0)} \quad & \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \lambda_{ind} \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0 \\
\text{s.t.} \quad & f^{\mathrm{LR}}(\boldsymbol{x}) \geq \varphi^{-1}(\nu)
\end{aligned}
$$

Counterfactual explanations for Logistic Regression

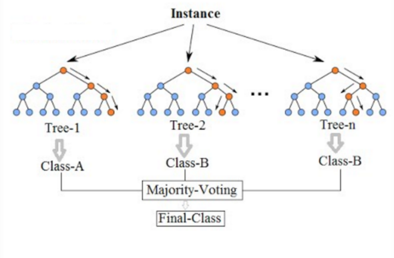


$$\min_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0)} \quad \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \lambda_{ind} \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0$$

$$\text{s.t.} \quad \boldsymbol{wx} + b \geq -\log\left(\frac{1-\nu}{\nu}\right)$$

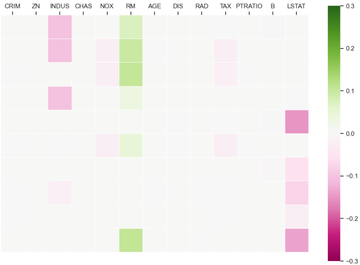
`Housing` dataset with Logistic Regression. CEs to be predicted in '+1' class. Heatmap indicates perturbations

Counterfactual explanations for Additive Tree Models (RF, XGBoost, etc)

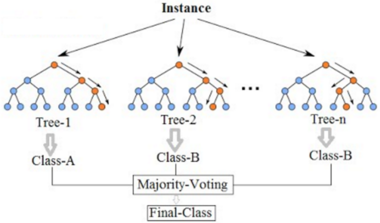


$$
\begin{aligned}
\min_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0)} \quad & \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \lambda_{ind} \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0 \\
\text{s.t.} \quad & f^{\mathrm{ATM}}(\boldsymbol{x}) \geq \nu
\end{aligned}
$$

Counterfactual explanations for Additive Tree Models (RF, XGBoost, etc)



`Housing` dataset with Random Forest. CEs to be predicted in '+1' class. Heatmap indicates perturbations



$$\min_{\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}^0)} \quad \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2 + \lambda_{ind} \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0$$

$$\text{s.t.} \quad \sum_{t=1}^{T} \sum_{l \in \mathcal{L}_+^t} w^t z_l^t \geq \nu$$

$\boldsymbol{z}$ routing of CE in trees of ATM

Counterfactual explanations

Different types of optimization problems:

- **smooth opt**, e.g., Joshi et al. (2019); Ramakrishnan et al. (2020); Wachter et al. (2017); Mothilal et al. (2020); Lucic et al. (2022)
- **MIP**, e.g., Cui et al. (2015); Fischetti and Jo (2018); Kanamori et al. (2020, 2021); Maragno et al. (2022); Parmentier and Vidal (2021); Russell (2019)
- **multi-objective opt**, e.g., Dandl et al. (2020); Del Ser et al. (2022); Raimundo et al. (2022),
- **robust opt**, e.g., Maragno et al. (2024)

Most of the literature focuses on the **single-instance single-counterfactual** setting

(Guidotti, 2022; Karimi et al., 2022; Verma et al., 2022)

Counterfactual explanations

Different types of optimization problems:

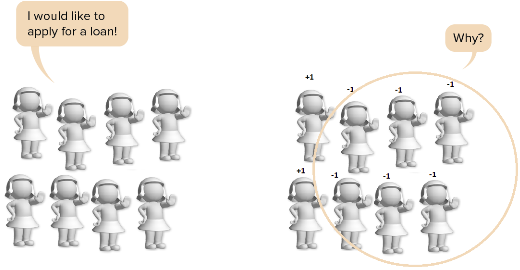
- **smooth opt**, e.g., Joshi et al. (2019); Ramakrishnan et al. (2020); Wachter et al. (2017); Mothilal et al. (2020); Lucic et al. (2022)
- **MIP**, e.g., Cui et al. (2015); Fischetti and Jo (2018); Kanamori et al. (2020, 2021); Maragno et al. (2022); Parmentier and Vidal (2021); Russell (2019)
- **multi-objective opt**, e.g., Dandl et al. (2020); Del Ser et al. (2022); Raimundo et al. (2022),
- **robust opt**, e.g., Maragno et al. (2024)

Most of the literature focuses on the **single-instance single-counterfactual** setting

(Guidotti, 2022; Karimi et al., 2022; Verma et al., 2022)

Outline

Group Counterfactual Analysis in Machine Learning



European Journal of Operational Research

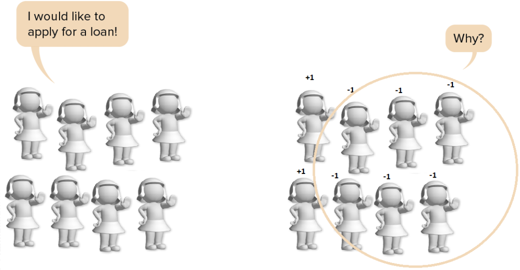
Available online 5 January 2024

In Press, Corrected Proof What's this?

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Group Counterfactual Analysis in Machine Learning



European Journal of Operational Research

Available online 5 January 2024

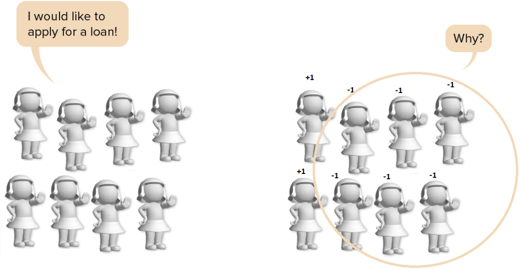
In Press, Corrected Proof What's this?

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Motivation

- **linking constraints** may exist between CEs, e.g., CEs for close individuals should also be close
- several CEs may be sought, sufficiently far (**diverse**) from each other (Wachter et al., 2017)
- a set of **critical features** is sought for CEs (Eckstein et al., 2021; Sharma et al., 2020)
- **benchmarks** for records are sought, i.e., same CE for a group of instances



European Journal of Operational Research

Available online 5 January 2024

In Press, Corrected Proof   What's this?

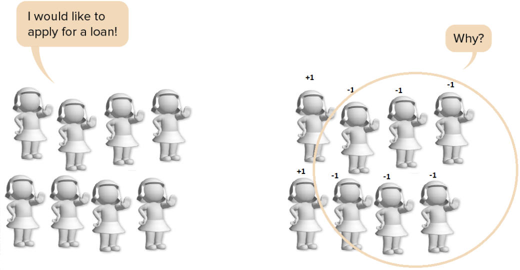
Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Motivation

- **linking constraints** may exist between CEs, e.g., CEs for close individuals should also be close
- several CEs may be sought, sufficiently far (**diverse**) from each other (Wachter et al., 2017)
- a set of **critical features** is sought for CEs (Eckstein et al., 2021; Sharma et al., 2020)
- **benchmarks** for records are sought, i.e., same CE for a group of instances

Group Counterfactual Analysis in Machine Learning



European Journal of Operational Research

Available online 5 January 2024

In Press, Corrected Proof What's this?

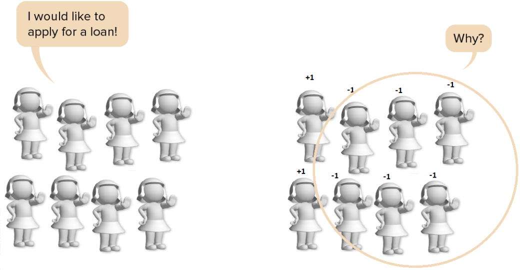
Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Motivation

- **linking constraints** may exist between CEs, e.g., CEs for close individuals should also be close
- several CEs may be sought, sufficiently far (**diverse**) from each other (Wachter et al., 2017)
- a set of **critical features** is sought for CEs (Eckstein et al., 2021; Sharma et al., 2020)
- **benchmarks** for records are sought, i.e., same CE for a group of instances

Group Counterfactual Analysis in Machine Learning



European Journal of Operational Research

Available online 5 January 2024

In Press, Corrected Proof What's this?

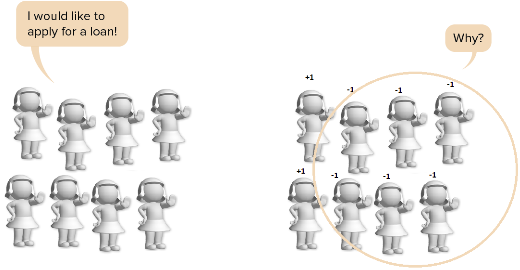
Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Motivation

- **linking constraints** may exist between CEs, e.g., CEs for close individuals should also be close
- several CEs may be sought, sufficiently far (**diverse**) from each other (Wachter et al., 2017)
- a set of **critical features** is sought for CEs (Eckstein et al., 2021; Sharma et al., 2020)
- **benchmarks** for records are sought, i.e., same CE for a group of instances

Group Counterfactual Analysis in Machine Learning



European Journal of Operational Research

Available online 5 January 2024

In Press, Corrected Proof What's this?

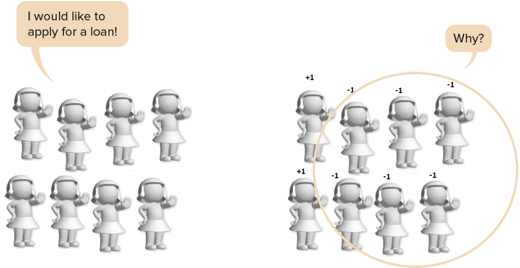
Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Motivation

- **linking constraints** may exist between CEs, e.g., CEs for close individuals should also be close
- several CEs may be sought, sufficiently far (**diverse**) from each other (Wachter et al., 2017)
- a set of **critical features** is sought for CEs (Eckstein et al., 2021; Sharma et al., 2020)
- **benchmarks** for records are sought, i.e., same CE for a group of instances

Group Counterfactual Analysis in Machine Learning



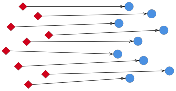
European Journal of Operational Research
Available online 5 January 2024
In Press, Corrected Proof What's this?

Mathematical optimization modelling for group counterfactual explanations

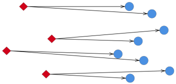
Emilio Carrizosa [a], Jasone Ramírez-Ayerbe [a], Dolores Romero Morales [b]

Jasone Ramírez Ayerbe

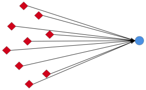
Group Counterfactual Explanations. Allocation Rules



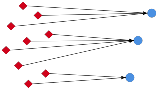
(a) One-for-one



(b) Many-for-one



(c) One-for-all

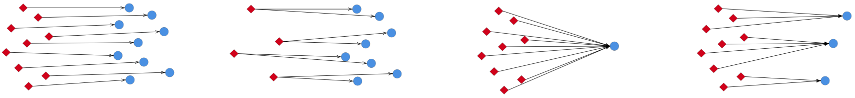


(d) One-for-many

Figure: Allocation rules between instances (in red) and their counterfactual explanations (in blue) in group counterfactual analysis

- For each $s \in \{1, 2, \dots, S\}$, define $\mathcal{R}_s$: set of indices $r \in \{1, 2, \dots, R\}$ s.t. counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$
- For each $r \in \{1, 2, \dots, R\}$, define $\mathcal{S}_r$: set of indices $s \in \{1, 2, \dots, S\}$ s.t. instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$
- Note: $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$
- $\mathcal{R}_s, \mathcal{S}_r$: **given? decision variables?**

Group Counterfactual Explanations. Allocation Rules

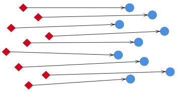


(a) One-for-one (b) Many-for-one (c) One-for-all (d) One-for-many

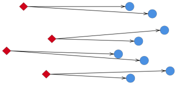
Figure: Allocation rules between instances (in red) and their counterfactual explanations (in blue) in group counterfactual analysis

- For each $s \in \{1, 2, \ldots, S\}$, define $\mathcal{R}_s$: set of indices $r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$
- For each $r \in \{1, 2, \ldots, R\}$, define $\mathcal{S}_r$: set of indices $s \in \{1, 2, \ldots, S\}$ s.t. instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$
- Note: $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$
- $\mathcal{R}_s, \mathcal{S}_r$: **given? decision variables?**

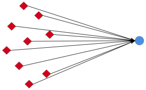
Group Counterfactual Explanations. Allocation Rules



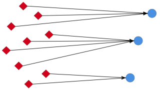
(a) One-for-one



(b) Many-for-one



(c) One-for-all

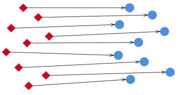


(d) One-for-many

Figure: Allocation rules between instances (in red) and their counterfactual explanations (in blue) in group counterfactual analysis

- For each $s \in \{1, 2, \ldots, S\}$, define $\mathcal{R}_s$: set of indices $r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$
- For each $r \in \{1, 2, \ldots, R\}$, define $\mathcal{S}_r$: set of indices $s \in \{1, 2, \ldots, S\}$ s.t. instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$
- Note: $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$
- $\mathcal{R}_s, \mathcal{S}_r$: **given? decision variables?**

Group Counterfactual Explanations. Allocation Rules

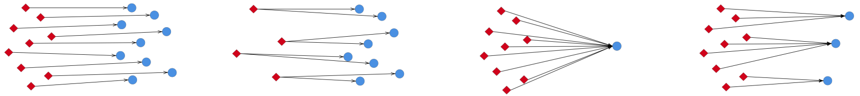


(a) One-for-one (b) Many-for-one (c) One-for-all (d) One-for-many

Figure: Allocation rules between instances (in red) and their counterfactual explanations (in blue) in group counterfactual analysis

- For each $s \in \{1, 2, \ldots, S\}$, define $\mathcal{R}_s$: set of indices $r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$
- For each $r \in \{1, 2, \ldots, R\}$, define $\mathcal{S}_r$: set of indices $s \in \{1, 2, \ldots, S\}$ s.t. instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$
- Note: $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$
- $\mathcal{R}_s, \mathcal{S}_r$: **given? decision variables?**

Group Counterfactual Explanations. Allocation Rules



(a) One-for-one (b) Many-for-one (c) One-for-all (d) One-for-many

Figure: Allocation rules between instances (in red) and their counterfactual explanations (in blue) in group counterfactual analysis

- For each $s \in \{1, 2, \ldots, S\}$, define $\mathcal{R}_s$: set of indices $r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $\boldsymbol{x}_r$ are associated with instance $\boldsymbol{x}_s^0$
- For each $r \in \{1, 2, \ldots, R\}$, define $\mathcal{S}_r$: set of indices $s \in \{1, 2, \ldots, S\}$ s.t. instances $\boldsymbol{x}_s^0$ are associated with counterfactual $\boldsymbol{x}_r$
- Note: $r \in \mathcal{R}_s$ iff $s \in \mathcal{S}_r$
- $\mathcal{R}_s, \mathcal{S}_r$: **given? decision variables?**

Group Counterfactual Explanations. Ingredients

$$\begin{array}{ll} \min_{\underline{\boldsymbol{x}}} & \left(\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}), -\boldsymbol{P}(\underline{\boldsymbol{x}})\right) \\ \text{s.t.} & \underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0), \end{array}$$

where

- $\underline{\boldsymbol{x}}^0 = (\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_S^0)$: S instances for which counterfactuals are sought
- $\underline{\boldsymbol{x}} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_R)$: R counterfactuals
- $\underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}^0) \subset \underline{\mathcal{X}} := \mathcal{X}^R$
- $\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$: cost incurred when $\underline{\boldsymbol{x}}^0$ is perturbed to yield $\underline{\boldsymbol{x}}$
- $\boldsymbol{P}(\underline{\boldsymbol{x}})$: component-wise nondecreasing function of the probabilities $P(\boldsymbol{x})$ of the counterfactuals

Cost function $\boldsymbol{C}$

$$\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \underline{\text{Dissimilarity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) + \lambda_c \underline{\text{Complexity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$$

- Dissimilarity: A plausible choice would be $\sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \text{Dissimilarity}(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$
- Complexity : At instance level with the zero norm, group level, etc

$$\gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \left\| \left(\max_i |x_{ij}^0 - x_{ij}| \right)_{j=1}^{J} \right\|_0$$

Cost function $\boldsymbol{C}$

$$\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \underline{\text{Dissimilarity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) + \lambda_c \underline{\text{Complexity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$$

- Dissimilarity: A plausible choice would be $\sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \text{Dissimilarity}(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$
- Complexity : At instance level with the zero norm, group level, etc

$$\gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \left\| \left(\max_i |x_{ij}^0 - x_{ij}| \right)_{j=1}^{J} \right\|_0$$

Cost function $\boldsymbol{C}$

$$\boldsymbol{C}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \underline{\text{Dissimilarity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) + \lambda_c \underline{\text{Complexity}}(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$$

- Dissimilarity: A plausible choice would be $\sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \text{Dissimilarity}(\boldsymbol{x}_s^0, \boldsymbol{x}_r)$
- Complexity : At instance level with the zero norm, group level, etc

$$\gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}}) = \left\| \left(\max_i |x_{ij}^0 - x_{ij}| \right)_{j=1}^{J} \right\|_0$$

Probabilities $\boldsymbol{P}$

- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \min_{r=1,\ldots,R} P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \frac{1}{R}\sum_{r=1}^{R} |\mathcal{S}_r| P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \left(\prod_{r=1}^{R} P(\boldsymbol{x}_r)^{|\mathcal{S}_r|}\right)^{1/R}$ ($\log(\boldsymbol{P}(\underline{\boldsymbol{x}}))$: concave for logistic classifier!)
- ...

Probabilities $\boldsymbol{P}$

- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \min_{r=1,\dots,R} P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \frac{1}{R}\sum_{r=1}^{R} |\mathcal{S}_r| P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \left(\prod_{r=1}^{R} P(\boldsymbol{x}_r)^{|\mathcal{S}_r|}\right)^{1/R}$ ($\log(\boldsymbol{P}(\underline{\boldsymbol{x}}))$: concave for logistic classifier!)
- ...

Probabilities $\boldsymbol{P}$

- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \min_{r=1,\dots,R} P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \frac{1}{R}\sum_{r=1}^{R} |\mathcal{S}_r| P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \left(\prod_{r=1}^{R} P(\boldsymbol{x}_r)^{|\mathcal{S}_r|}\right)^{1/R}$ (log($\boldsymbol{P}(\underline{\boldsymbol{x}})$) : concave for logistic classifier!)
- ...

Probabilities $\boldsymbol{P}$

- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \min_{r=1,\dots,R} P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \frac{1}{R}\sum_{r=1}^{R} |\mathcal{S}_r| P(\boldsymbol{x}_r)$
- $\boldsymbol{P}(\underline{\boldsymbol{x}}) = \left(\prod_{r=1}^{R} P(\boldsymbol{x}_r)^{|\mathcal{S}_r|}\right)^{1/R}$ ($\log(\boldsymbol{P}(\underline{\boldsymbol{x}}))$: concave for logistic classifier!)
- ...

One-for-one CEs in Carrizosa et al. (2024b)

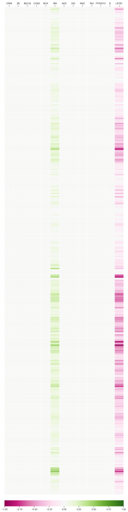
One-for-one CEs

- local and global sparsity are sought for CEs
- linking constraints, such as Lipschitz continuity

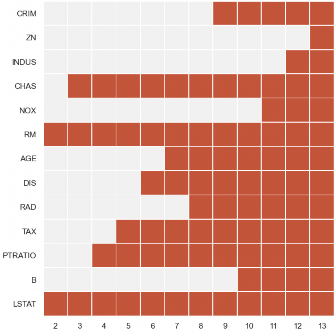
$$\min_{\underline{\boldsymbol{x}} \in \mathcal{X}(\underline{\boldsymbol{x}}^0)} \quad \sum_{s=1}^{S} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_2^2 + \lambda_{ind} \sum_{s=1}^{S} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_s\|_0 + \lambda_{glob} \gamma_0(\underline{\boldsymbol{x}}^0, \underline{\boldsymbol{x}})$$

$$\text{s.t.} \quad f(\boldsymbol{x}_s) \geq \varphi^{-1}(\nu) \quad \forall s = 1, 2, \ldots, S$$

One-for-One CEs in Carrizosa et al. (2024b)



Housing dataset with Logistic Regression. CEs to be predicted in ‘+1’ class. Heatmap indicates perturbations

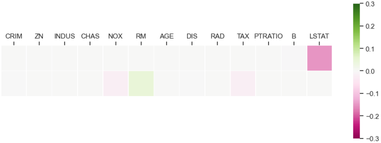


Housing dataset with Logistic Regression. Features that need to be perturbed (in red) for instances to be predicted in ‘+1’ class

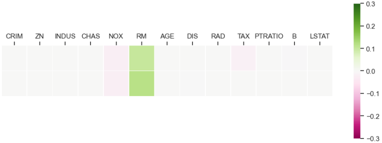
One-for-One with Lipschitz continuity in Carrizosa et al. (2024a)

For some threshold value τ

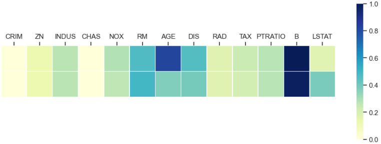
$$d(\boldsymbol{x}_i, \boldsymbol{x}_j) \leq \tau d(\boldsymbol{x}_i^0, \boldsymbol{x}_j^0), \quad \forall i, j \tag{1}$$



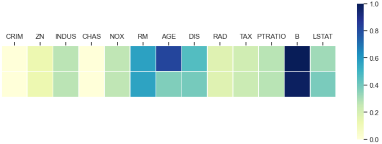
(a) Perturbations without (1)



(b) Perturbations with (1)



(c) Feature values without (1)



(d) Feature values with (1)

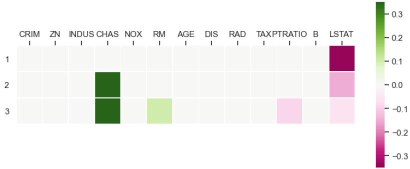
`Housing` dataset with Random Forest. CEs to be predicted in '+1' class. Features perturbations are displayed on the two pictures on the top, with the Lipschitz continuity constraint for $\tau = 10$ and without this constraint, respectively, whereas in the two bottom pictures the corresponding features values are displayed

Many-for-one CEs in Carrizosa et al. (2024b)

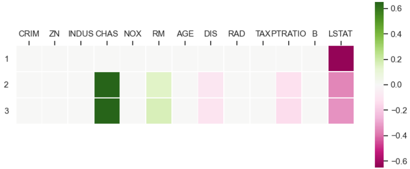
$$\begin{aligned}\min_{\boldsymbol{x}_r \in \mathcal{X}^r(\boldsymbol{x}_s^0)} \quad & \|\boldsymbol{x}_s^0 - \boldsymbol{x}_r\|_2^2 + \lambda_{ind}\|\boldsymbol{x}_s^0 - \boldsymbol{x}_r\|_0 \\ \text{s.t.} \quad & f(\boldsymbol{x}_r) \geq \varphi^{-1}(\nu)\end{aligned}$$

Many-for-one CEs in Carrizosa et al. (2024b)

$$(\boldsymbol{x}_r)_{\text{LSTAT}} \leq Q_1 \quad \text{or} \quad Q_1 < (\boldsymbol{x}_r)_{\text{LSTAT}} \leq Q_3 \quad \text{or} \quad (\boldsymbol{x}_r)_{\text{LSTAT}} > Q_3$$



(a) Perturbations for $\boldsymbol{x}_1^0$



(b) Perturbations for $\boldsymbol{x}_2^0$

`Housing` dataset with Logistic Regression. Many-for-one counterfactual explanations with $R = 3$ for instances two instances.

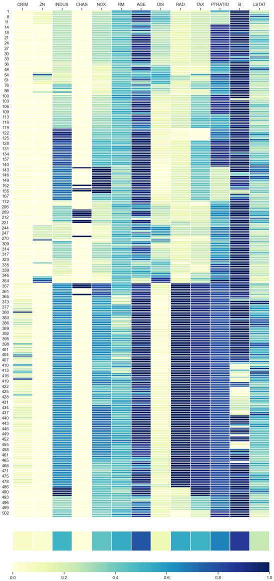
One-for-all and one-for-many CEs in Carrizosa et al. (2024a)

One-for-all and one-for-many CEs

- Identify R CEs for I instances, with $R < I$

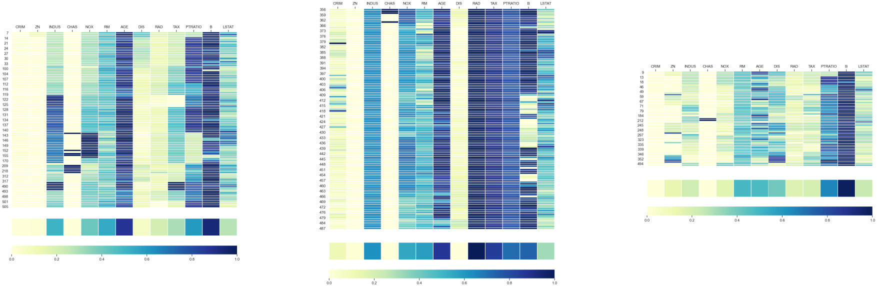
$$
\begin{aligned}
\min_{\underline{\boldsymbol{x}} \in \underline{\mathcal{X}}, \boldsymbol{y}} \quad & \sum_{r=1}^{R} \sum_{s=1}^{S} y_{sr} \|\boldsymbol{x}_s^0 - \boldsymbol{x}_r\|_2^2 & \\
\text{s.t.} \quad & \boldsymbol{w}\boldsymbol{x}_r + b \geq \varphi^{-1}(\nu) & \forall r = 1, 2, \ldots, R \\
& \sum_{r=1}^{R} y_{sr} = 1 & \forall s = 1, 2, \ldots, S \\
& y_{sr} \in \{0, 1\} & \forall s = 1, 2, \ldots, S \quad \forall r = 1, 2, \ldots, R.
\end{aligned}
$$

One-for-all CE in Carrizosa et al. (2024a)



Housing dataset with Logistic Regression. $R = 1$ cluster and corresponding CEs predicted in '+1' class. Heatmaps indicate feature values

One-for-many CEs in Carrizosa et al. (2024a)



`Housing` dataset with Logistic Regression. $R = 3$ clusters and corresponding CEs predicted in '+1' class. Heatmaps indicate feature values

Outline

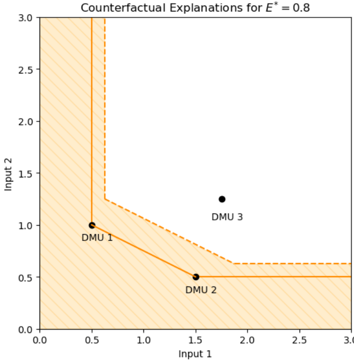
- Introduction
- On Group Counterfactual Analysis
- Counterfactual Analysis Beyond Machine Learning
- Conclusions

Counterfactual Analysis in Benchmarking in Bogetoft et al. (2024)

Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency $\rightarrow$ **Bilevel Optimization**

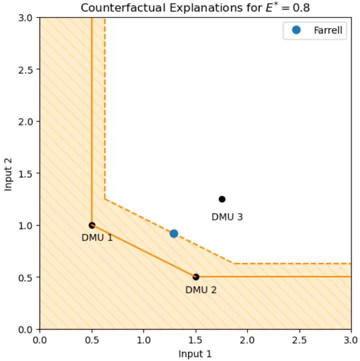
Counterfactual Analysis in Benchmarking in Bogetoft et al. (2024)

Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency $\rightarrow$ **Bilevel Optimization**



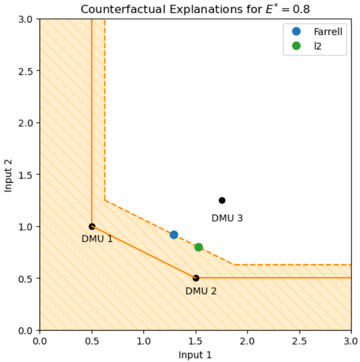
Counterfactual Analysis in Benchmarking in Bogetoft et al. (2024)

Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency $\rightarrow$ **Bilevel Optimization**



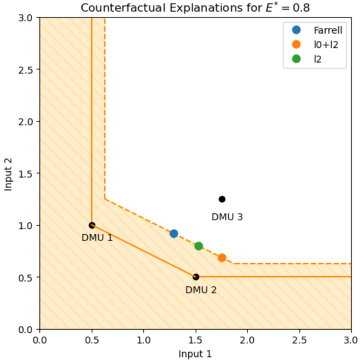
Counterfactual Analysis in Benchmarking in Bogetoft et al. (2024)

Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency $\rightarrow$ **Bilevel Optimization**



Counterfactual Analysis in Benchmarking in Bogetoft et al. (2024)

Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency $\rightarrow$ **Bilevel Optimization**



Counterfactual Explanations for DEA models

It is about minimizing the distance to a complement of a convex set (Thach, 1988)

Counterfactual explanation to be at least E^* efficient

$$
\begin{aligned}
\min_{\hat{\boldsymbol{x}}, E} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) \\
\text{s.t.} \quad & \hat{\boldsymbol{x}} \in \mathbb{R}^I_+ \\
& E \geq E^* \\
& E \in \underset{\bar{E}, \lambda^0, \ldots, \lambda^K}{\arg\min} \left\{ \bar{E} \quad : \quad \bar{E}\hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \lambda^k \boldsymbol{x}^k, \quad \boldsymbol{y}^0 \leq \sum_{k=0}^{K} \lambda^k \boldsymbol{y}^k, \right. \\
& \qquad \left. \bar{E} \geq 0, \quad \boldsymbol{\lambda} \in \mathbb{R}^{K+1}_+ \right\}
\end{aligned}
$$

From bilevel to single level

$$
\begin{aligned}
\min_{\hat{\boldsymbol{x}}, F, \boldsymbol{\beta}, \gamma, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) & & \\
\text{s.t.} \quad & F \leq F^* & & \boldsymbol{\beta} \leq M_f \boldsymbol{w} \\
& \hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \beta^k \boldsymbol{x}^k & & \gamma_I^T \boldsymbol{x}^k - \gamma_O^T \boldsymbol{y}^k \leq M_f(1 - w_k) \forall k \\
& F \boldsymbol{y}^0 \leq \sum_{k=0}^{K} \beta^k \boldsymbol{y}^k & & \gamma_O^T \boldsymbol{y}^0 = 1 \\
& \gamma_I \leq M_I \boldsymbol{u} & & \gamma_I^T \boldsymbol{x}^k - \gamma_O^T \boldsymbol{y}^k \geq 0 \quad k = 0, \ldots, K \\
& \hat{\boldsymbol{x}} - \sum_{k=1}^{K} \beta^k \boldsymbol{x}^k \leq M_I(1 - \boldsymbol{u}) & & \gamma_I, \gamma_O, F, \boldsymbol{\beta} \geq 0 \\
& \gamma_O \leq M_O \boldsymbol{v} & & \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \{0, 1\} \\
& - F \boldsymbol{y}^0 + \sum_{k=1}^{K} \beta^k \boldsymbol{y}^k \leq M_O(1 - \boldsymbol{v}) & & \hat{\boldsymbol{x}} \in \mathbb{R}_+^I
\end{aligned}
$$

With the cost function

$$C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) = \nu_0 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_0 + \nu_1 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_1 + \nu_2 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_2^2,$$

we obtain a Mixed Integer Convex Quadratic with Linear Constraints formulation

From bilevel to single level

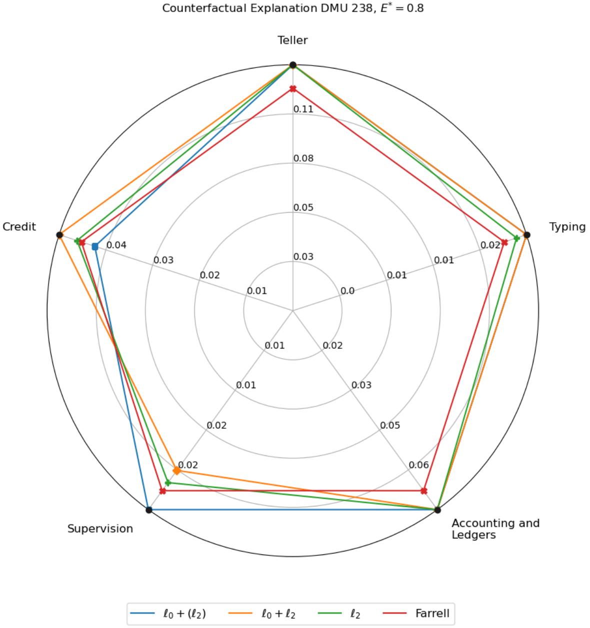
$$
\begin{aligned}
\min_{\hat{\boldsymbol{x}}, F, \boldsymbol{\beta}, \gamma, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}} \quad & C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) && \\
\text{s.t.} \quad & F \leq F^* && \boldsymbol{\beta} \leq M_f \boldsymbol{w} \\
& \hat{\boldsymbol{x}} \geq \sum_{k=0}^{K} \beta^k \boldsymbol{x}^k && \gamma_I^T \boldsymbol{x}^k - \gamma_O^T \boldsymbol{y}^k \leq M_f (1 - w_k) \forall k \\
& F \boldsymbol{y}^0 \leq \sum_{k=0}^{K} \beta^k \boldsymbol{y}^k && \gamma_O^T \boldsymbol{y}^0 = 1 \\
& \gamma_I \leq M_I \boldsymbol{u} && \gamma_I^T \boldsymbol{x}^k - \gamma_O^T \boldsymbol{y}^k \geq 0 \quad k = 0, \ldots, K \\
& \hat{\boldsymbol{x}} - \sum_{k=1}^{K} \beta^k \boldsymbol{x}^k \leq M_I (1 - \boldsymbol{u}) && \gamma_I, \gamma_O, F, \boldsymbol{\beta} \geq 0 \\
& \gamma_O \leq M_O \boldsymbol{v} && \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w} \in \{0, 1\} \\
& - F \boldsymbol{y}^0 + \sum_{k=1}^{K} \beta^k \boldsymbol{y}^k \leq M_O (1 - \boldsymbol{v}) && \hat{\boldsymbol{x}} \in \mathbb{R}_+^I
\end{aligned}
$$

With the cost function

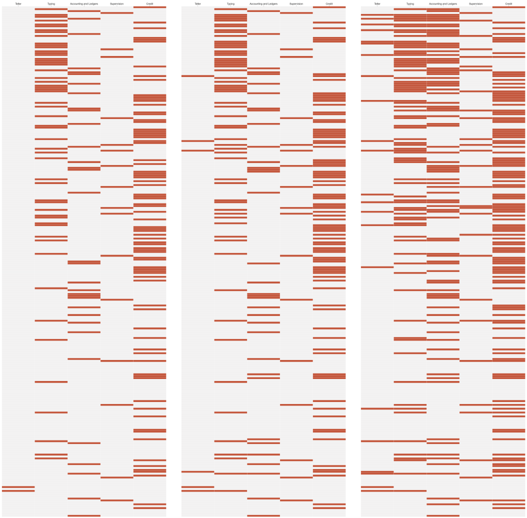
$$C(\boldsymbol{x}^0, \hat{\boldsymbol{x}}) = \nu_0 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_0 + \nu_1 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_1 + \nu_2 \|\boldsymbol{x}^0 - \hat{\boldsymbol{x}}\|_2^2,$$

we obtain a Mixed Integer Convex Quadratic with Linear Constraints formulation

Results for banking branches



Results for banking branches



(a) $C = \ell_0 + (\ell_2)$ (b) $C = \ell_0 + \ell_2$ (c) $C = \ell_2$

Figure: The inputs that change when we impose a desired efficiency of $E^* = 0.8$

Counterfactual Analysis for Supervised Discretization

We study in Piccialli et al. (2024) ...

... how to detect with Counterfactual Analysis **critical thresholds of features** for a given black-box classifier to derive

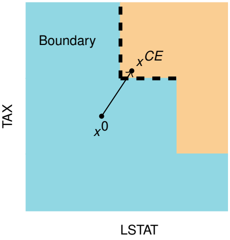
- Feature discretization
- Surrogate white/gray box classifier

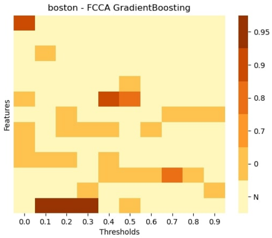
Counterfactual Analysis for Supervised Discretization

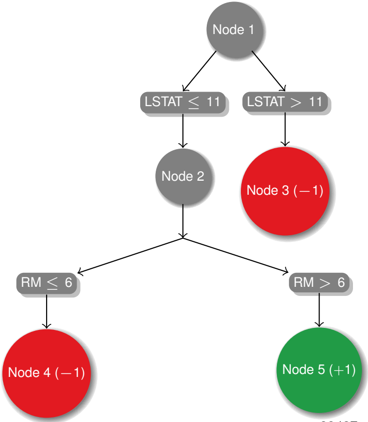
We study in Piccialli et al. (2024) ...

... how to detect with Counterfactual Analysis **critical thresholds of features** for a given black-box classifier to derive

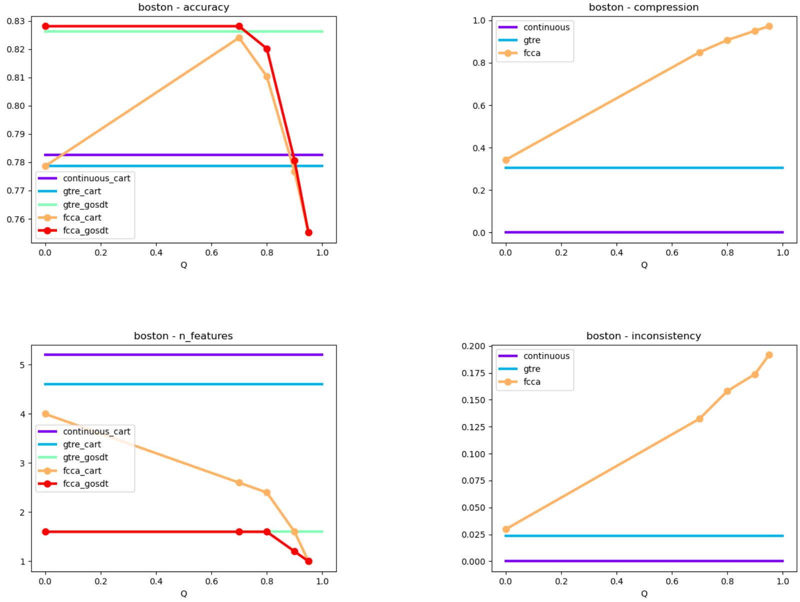
- Feature discretization
- Surrogate white/gray box classifier







Counterfactual Analysis for Supervised Discretization



Outline

Conclusions

- MIP (and more) for Group Counterfactual Analysis
- Connections with Locational Analysis
- Ability to handle decision-making settings beyond ML, such as those arising in Benchmarking
- New opportunities for the community to develop bespoke algorithms

References I

R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. On optimal regression trees to detect critical intervals for multivariate functional data. *Computers and Operations Research*, 152:106152, 2023.

P. Bogetoft, J. Ramírez Ayerbe, and D. Romero Morales. Counterfactual analysis and target setting in benchmarking. *European Journal of Operational Research*, 315(3):1083–1095, 2024.

E. Carrizosa, L.H. Mortensen, D. Romero Morales, and M.R. Sillero-Denamiel. The tree based linear regression model for hierarchical categorical variables. *Expert Systems With Applications*, 203(7):117423, 2022.

E. Carrizosa, J. Ramírez Ayerbe, and D. Romero Morales. A new model for counterfactual analysis for functional data. *Forthcoming in Advances in Data Analysis and Classification* https://www.researchgate.net/publication/363539291_A_New_Model_for_Counterfactual_Analysis_for_Functional_Data, 2023.

E. Carrizosa, J. Ramírez Ayerbe, and D. Romero Morales. Mathematical optimization modelling for group counterfactual explanations. *Forthcoming in European Journal of Operational Research* https://www.researchgate.net/publication/368958766_Mathematical_Optimization_Modelling_for_Group_Counterfactual_Explanations, 2024a.

E. Carrizosa, J. Ramírez Ayerbe, and D. Romero Morales. Generating collective counterfactual explanations in score-based classification via mathematical optimization. *Expert Systems With Applications*, 238:121954, 2024b.

Z. Cui, W. Chen, Y. He, and Y. Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 179–188, 2015.

S. Dandl, C. Molnar, M. Binder, and B. Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer, 2020.

J. Del Ser, A. Barredo-Arrieta, N. Díaz-Rodríguez, F. Herrera, and A. Holzinger. Exploring the trade-off between plausibility, change intensity and adversarial power in counterfactual explanations using multi-objective optimization. *arXiv preprint arXiv:2205.10232*, 2022.

N. Eckstein, A.S. Bates, G.S.X.E. Jefferis, and J. Funke. Discriminative attribution from counterfactuals. *arXiv preprint arXiv:2109.13412*, 2021.

European Commission. *White Paper on Artificial Intelligence : a European approach to excellence and trust*. https://ec.europa.eu/info/publications/white-paper-artificial-intelligence-european-approach-excellence-and-trust_en, 2020.

M. Fischetti and J. Jo. Deep neural networks and mixed integer linear optimization. *Constraints*, 23(3):296–309, 2018.

R. Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Forthcoming in Data Mining and Knowledge Discovery*, 2022.

S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh. Towards realistic individual recourse and actionable explanations in black-box decision making systems. *arXiv preprint arXiv:1907.09615*, 2019.

K. Kanamori, T. Takagi, K. Kobayashi, and H. Arimura. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In C. Bessiere, editor, *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2855–2862, 2020.

References II

K. Kanamori, T. Takagi, K. Kobayashi, Y. Ike, K. Uemura, and H. Arimura. Ordered counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11564–11574, 2021.

A.-H. Karimi, B. Schölkopf, and I. Valera. Algorithmic recourse: from counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 353–362, 2021.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022.

A. Lucic, H. Oosterhuis, H. Haned, and M. de Rijke. FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5313–5322, 2022.

D. Maragno, T. E Röber, and I. Birbil. Counterfactual explanations using optimization with constraint learning. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022.

D. Maragno, J. Kurtz, T.E. Röber, R. Goedhart, Ş.I. Birbil, and D. den Hertog. Finding regions of counterfactual explanations via robust optimization. *Forthcoming in INFORMS Journal on Computing*, 2024.

D. Martens and F. Provost. Explaining data-driven document classifications. *MIS Quarterly*, 38(1):73–99, 2014.

R.K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 607–617, 2020.

C. Panigutti, R. Hamon, I. Hupont, D. Fernandez Llorca, D. Fano Yela, H. Junklewitz, S. Scalzo, G. Mazzini, I. Sanchez, J. Soler Garrido, and E. Gomez. The Role of Explainable AI in the Context of the AI Act. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, pages 1139–1150, New York, NY, USA, 2023.

A. Parmentier and T. Vidal. Optimal counterfactual explanations in tree ensembles. In *International Conference on Machine Learning*, pages 8422–8431. PMLR, 2021.

V. Piccialli, D. Romero Morales, and C. Salvatore. Supervised feature compression based on counterfactual analysis. *European Journal of Operational Research*, 317:273–285, 2024.

M.M. Raimundo, L.G. Nonato, and J. Poco. Mining pareto-optimal counterfactual antecedents with a branch-and-bound model-agnostic algorithm. *Forthcoming in Data Mining and Knowledge Discovery*, pages 1–33, 2022.

G. Ramakrishnan, Y.C. Lee, and A. Albarghouthi. Synthesizing action sequences for modifying model decisions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5462–5469, 2020.

C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys*, 16:1–85, 2022.

C. Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20–28, 2019.

S. Sharma, J. Henderson, and J. Ghosh. Certifai: A common framework to provide explanations and analyse the fairness and robustness of black-box models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 166–172, 2020.

P.T. Thach. The design centering problem as a DC programming problem. *Mathematical Programming*, 41(1):229–248, 1988.

S. Verma, V. Boonsanong, M. Hoang, K.E. Hines, J.P. Dickerson, and C. Shah. Counterfactual explanations and algorithmic recourses for machine learning: A review. *arXiv preprint arXiv:2010.10596*, 2022.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law & Technology*, 31:841–887, 2017.