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
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

Professor Paula Carroll
Chair of the EURO WISDOM Forum

Talk about: Introduction to the EURO WISDOM Forum

April 20
Speakers

Online Seminar Series
Machine Learning NeEDS...

NeEDS - Network of European Data Scientists

Public
81 videos 2,219 views Updated yesterday


Leading academics from around the globe talk about their current work in this burgeoning area. The speakers come from

1. Machine Learning NeEDS Mathematical Optimization with Prof. Ilker Birbil
   NeEDS - Network of European Data Scientists • 705 views • 3 years ago

2. Machine Learning NeEDS Mathematical Optimization with Prof Andrea Lodi
   NeEDS - Network of European Data Scientists • 573 views • 2 years ago

3. YOUNG Seminar Series Machine Learning NeEDS Mathematical Optimization on March 13, 2023
   NeEDS - Network of European Data Scientists • 561 views • 1 year ago

4. Machine Learning NeEDS Mathematical Optimization with Prof Nathan Kallus
   NeEDS - Network of European Data Scientists • 537 views • 2 months ago

5. Machine Learning NeEDS Mathematical Optimization with Prof Adam Elmachtoub
   NeEDS - Network of European Data Scientists • 486 views • 5 months ago

6. YOUNG Machine Learning NeEDS Mathematical Optimization on February 1, 2021
   NeEDS - Network of European Data Scientists • 409 views • 3 years ago
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 $\mathbf{x}^0$ classified as ‘-1’, and we want to give insights on how to change the features to be classified as ‘+1’.
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 $\mathbf{x}^0$ 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 $P : \mathcal{X} \to [0, 1]$, $P(x)$: probability of belonging to class +1, and

  - $x_0 \in \mathcal{X}$,

- the goal is to find the changes (to some $x \in \mathcal{X}(x^0) \subset \mathcal{X}$) with minimum cost $C(x, x^0)$ that cause $x^0$ to increase the probability $P(x^0)$ to $P(x)$.

\[
\begin{align*}
\min_x & \quad C(x, x^0) \\
\text{s.t.} & \quad P(x) \geq \nu \\
& \quad x \in \mathcal{X}(x^0) \\
\min_x & \quad (C(x, x^0), -P(x)) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0) \\
\min_x & \quad (1 - \lambda)C(x, x^0) - \lambda P(x) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0)
\end{align*}
\]
Counterfactual explanations

- We are given a **probabilistic** classifier $P : \mathcal{X} \to [0, 1]$, $P(x)$: probability of belonging to class $+1$, and
- $x_0 \in \mathcal{X}$,
- the goal is to find the changes (to some $x \in \mathcal{X}(x^0) \subset \mathcal{X}$) with minimum cost $C(x, x^0)$ that cause $x^0$ to increase the probability $P(x^0)$ to $P(x)$

$$
\begin{align*}
\min_x & \quad C(x, x^0) \\
\text{s.t.} & \quad P(x) \geq \nu \\
& \quad x \in \mathcal{X}(x^0)
\end{align*}$$

$$
\begin{align*}
\min_x & \quad (C(x, x^0), -P(x)) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0)
\end{align*}$$

$$
\begin{align*}
\min_x & \quad (1 - \lambda)C(x, x^0) - \lambda P(x) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0)
\end{align*}$$
Counterfactual explanations

- We are given a **probabilistic** classifier $P : \mathcal{X} \rightarrow [0, 1]$, $P(x) : \text{probability of belonging to class } +1$, and

- $x_0 \in \mathcal{X}$,

- the goal is to find the changes (to some $x \in \mathcal{X}(x^0) \subset \mathcal{X}$) **with minimum cost** $C(x, x^0)$ that cause $x^0$ to increase the probability $P(x^0)$ to $P(x)$

\[
\begin{align*}
\min_x & \quad C(x, x^0) \\
\text{s.t.} & \quad P(x) \geq \nu \\
& \quad x \in \mathcal{X}(x^0)
\end{align*}
\]

\[
\begin{align*}
\min_x & \quad (C(x, x^0), -P(x)) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0)
\end{align*}
\]

\[
\begin{align*}
\min_x & \quad (1 - \lambda)C(x, x^0) - \lambda P(x) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0)
\end{align*}
\]
Counterfactual explanations

- \( \mathcal{X} \)
  - Defined by features (tabular data), or
  - More complex data such as functional one (Carrizosa et al., 2023)

- \( \mathcal{X}(x^0) \)
  - Points from some training set \( \rightarrow \) discrete optimization models
  - Synthetic data \( \rightarrow \) (mixed integer) nonlinear optimization models

If \( P(x) = \varphi(f(x)) \) and \( \varphi \uparrow \), then

\[ P(x) \geq \nu \iff f(x) \geq \varphi^{-1}(\nu) \]

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

- \( C(x, x^0) = \text{Dissimilarity}(x, x^0) + \lambda_c \text{Complexity}(x, x^0) \)
  - Dissimilarity\((x, x^0)\) is usually modeled with \( \ell_p \) 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\((x, x^0)\) 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then
  $$P(x) \geq \nu \iff f(x) \geq \varphi^{-1}(\nu)$$

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

- $C(x, x^0) = \text{Dissimilarity}(x, x^0) + \lambda_c \text{Complexity}(x, x^0)$
  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then

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

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

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

  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then

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

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

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

  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then
  $$P(x) \geq \nu \iff f(x) \geq \varphi^{-1}(\nu)$$

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

- $C(x, x^0) = \text{Dissimilarity}(x, x^0) + \lambda_c \text{Complexity}(x, x^0)$
  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then

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

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

- $C(x, x^0) = \text{Dissimilarity}(x, x^0) + \lambda_c \text{Complexity}(x, x^0)$
  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then
  \[
P(x) \geq \nu \iff f(x) \geq \varphi^{-1}(\nu)
  \]
  These are known as score-based classifiers, e.g., LR, SVM, NN, RF, XGBoost

- $C(x, x^0) = \text{Dissimilarity}(x, x^0) + \lambda_c \text{Complexity}(x, x^0)$
  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then
  $$P(x) \geq \nu \iff f(x) \geq \varphi^{-1}(\nu)$$

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

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

  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then

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

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

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

  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ 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)

- $\mathcal{X}(x^0)$
  - Points from some training set $\rightarrow$ discrete optimization models
  - Synthetic data $\rightarrow$ (mixed integer) nonlinear optimization models

- If $P(x) = \varphi(f(x))$ and $\varphi \uparrow$, then

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

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

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

  - Dissimilarity$(x, x^0)$ is usually modeled with $\ell_p$ 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$(x, x^0)$ can be measured with the zero norm, or more complex sparsity measures (Blanquero et al., 2023)
Counterfactual explanations for Logistic Regression

\[
\min_{x \in X(x^0)} \|x^0 - x\|_2^2 + \lambda_{ind} \|x^0 - x\|_0 \\
\text{s.t. } f^{LR}(x) \geq \varphi^{-1}(\nu)
\]
Counterfactual explanations for Logistic Regression

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

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

\[
\text{s.t. } wx + b \geq -\log \left( \frac{1 - \nu}{\nu} \right)
\]
Counterfactual explanations for Additive Tree Models (RF, XGBoost, etc)

\[
\min_{\mathbf{x} \in \mathcal{X}(\mathbf{x}^0)} \left\| \mathbf{x}^0 - \mathbf{x} \right\|_2^2 + \lambda_{ind} \left\| \mathbf{x}^0 - \mathbf{x} \right\|_0 \\
\text{s.t. } f^{\text{ATM}}(\mathbf{x}) \geq \nu
\]
Counterfactual explanations for Additive Tree Models (RF, XGBoost, etc)

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

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

s.t. $$\sum_{t=1}^{T} \sum_{l \in \mathcal{L}_l^t} w^t z^t_l \geq \nu$$

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

- Introduction
- On Group Counterfactual Analysis
- Counterfactual Analysis Beyond Machine Learning
- Conclusions
Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa, Jasone Ramírez-Ayerbe, Dolores Romero Morales
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

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa, Jasone Ramírez-Ayerbe, Dolores Romero Morales
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**), and sufficiently far 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.

Mathematical optimization modelling for group counterfactual explanations
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

---

Mathematical optimization modelling for group counterfactual explanations

**European Journal of Operational Research**

Available online 5 January 2024

In Press, Corrected Proof  What’s this? 🔄

---

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

---

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa a, b, Jasone Ramírez-Ayerbe a, b, 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

Mathematical optimization modelling for group counterfactual explanations

Emilio Carrizosa, Jasone Ramírez-Ayerbe, Dolores Romero Morales
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 \( R_s \) : set of indices \( r \in \{1, 2, \ldots, R\} \) s.t. counterfactuals \( x_r \) are associated with instance \( x_s^0 \)
- For each \( r \in \{1, 2, \ldots, R\} \), define \( S_r \) : set of indices \( s \in \{1, 2, \ldots, S\} \) s.t. instances \( x_s^0 \) are associated with counterfactual \( x_r \)
- Note: \( r \in R_s \) iff \( s \in S_r \)
- \( R_s, S_r \) : given? decision variables?
Group Counterfactual Explanations. Allocation Rules

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 \( x_r \) are associated with instance \( x_s^0 \)
- For each \( r \in \{1, 2, \ldots, R\} \), define \( \mathcal{S}_r \) : set of indices \( s \in \{1, 2, \ldots, S\} \) s.t. instances \( x_s^0 \) are associated with counterfactual \( x_r \)
- 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

For each $s \in \{1, 2, \ldots, S\}$, define $R_s$ : set of indices $r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $x_r$ are associated with instance $x_0^s$

For each $r \in \{1, 2, \ldots, R\}$, define $S_r$ : set of indices $s \in \{1, 2, \ldots, S\}$ s.t. instances $x_0^s$ are associated with counterfactual $x_r$

Note: $r \in R_s$ iff $s \in S_r$

$R_s, S_r$ : given? decision variables?
Group Counterfactual Explanations. Allocation Rules

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 $R_s : \text{set of indices } r \in \{1, 2, \ldots, R\}$ s.t. counterfactuals $x_r$ are associated with instance $x_0^s$
- For each $r \in \{1, 2, \ldots, R\}$, define $S_r : \text{set of indices } s \in \{1, 2, \ldots, S\}$ s.t. instances $x_0^s$ are associated with counterfactual $x_r$
- Note: $r \in R_s$ iff $s \in S_r$
- $R_s, S_r : \text{given? decision variables?}$
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 \( R_s : \) set of indices \( r \in \{1, 2, \ldots, R\} \) s.t. counterfactuals \( x_r \) are associated with instance \( x_s^0 \)
- For each \( r \in \{1, 2, \ldots, R\} \), define \( S_r : \) set of indices \( s \in \{1, 2, \ldots, S\} \) s.t. instances \( x_s^0 \) are associated with counterfactual \( x_r \)
- Note: \( r \in R_s \) iff \( s \in S_r \)
- \( R_s, S_r : \text{given? decision variables?} \)
Group Counterfactual Explanations. Ingredients

\[
\begin{align*}
\min_x & \quad (C(x^0, x), -P(x)) \\
\text{s.t.} & \quad x \in \mathcal{X}(x^0),
\end{align*}
\]

where

- \( x^0 = (x^0_1, \ldots, x^0_S) \): \( S \) instances for which counterfactuals are sought
- \( x = (x_1, \ldots, x_R) \): \( R \) counterfactuals
- \( x \in \mathcal{X}(x^0) \subset \mathcal{X} := \mathcal{X}^R \)
- \( C(x^0, x) \): cost incurred when \( x^0 \) is perturbed to yield \( x \)
- \( P(x) \): component-wise nondecreasing function of the probabilities \( P(x) \) of the counterfactuals
Cost function $\mathcal{C}$

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

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

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

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

- **Dissimilarity**: A plausible choice would be $\sum_{s=1}^{S} \sum_{r \in R_s} \text{Dissimilarity}(x^0_s, x_r)$

- **Complexity**: At instance level with the zero norm, group level, etc

$$\gamma_0(x^0, x) = \left\| \left( \max_i |x^0_{ij} - x_{ij}| \right)^j \right\|_0$$
Cost function $C$

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

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

$$\gamma_0(x^0, x) = \left\| \left( \max_i |x^0_{ij} - x_{ij}| \right)_{j=1}^J \right\|_0$$
• \( P(x) = \min_{r=1,...,R} P(x_r) \)

• \( P(x) = \frac{1}{R} \sum_{r=1}^{R} |S_r| P(x_r) \)

• \( P(x) = \left( \prod_{r=1}^{R} P(x_r)^{|S_r|} \right)^{1/R} (\log(P(x)) : \text{concave for logistic classifier!}) \)

• ...
Probabilities $P$

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

...
Probabilities $P$

- $P(x) = \min_{r=1,\ldots,R} P(x_r)$
- $P(x) = \frac{1}{R} \sum_{r=1}^{R} |S_r| P(x_r)$
- $P(x) = \left( \prod_{r=1}^{R} P(x_r)^{|S_r|} \right)^{1/R}$

$(\log(P(x)) : \text{concave for logistic classifier!})$
Probabilities $P$

- $P(x) = \min_{r=1, \ldots, R} P(x_r)$
- $P(x) = \frac{1}{R} \sum_{r=1}^{R} |S_r| P(x_r)$
- $P(x) = \left( \prod_{r=1}^{R} P(x_r)^{|S_r|} \right)^{1/R} \left( \log(P(x)) \right)$: 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

\[
\begin{align*}
\min_{x \in \mathcal{X}(x^0)} & \quad \sum_{s=1}^{S} \|x_s^0 - x_s\|_2^2 + \lambda_{ind} \sum_{s=1}^{S} \|x_s^0 - x_s\|_0 + \lambda_{glob} \gamma_0(x^0, x) \\
\text{s.t.} & \quad f(x_s) \geq \varphi^{-1}(\nu) \quad \forall s = 1, 2, \ldots, S
\end{align*}
\]
One-for-One CEs in Carrizosa et al. (2024b)

Housing dataset with Logistic Regression. CEs to be predicted in ‘+1’ class. Heatmap indicates perturbations.
One-for-One with Lipschitz continuity in Carrizosa et al. (2024a)

For some threshold value $\tau$

\[ d(x_i, x_j) \leq \tau d(x_i^0, x_j^0), \quad \forall i, j \]  \hfill (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)

\[
\min_{x_r \in \mathcal{X}^r(x_s^0)} \|x_s^0 - x_r\|^2_2 + \lambda_{ind} \|x_s^0 - x_r\|_0 \\
\text{s.t. } f(x_r) \geq \varphi^{-1}(\nu)
\]
Many-for-one CEs in Carrizosa et al. (2024b)

\[(x_r)_{LSTAT} \leq Q_1 \quad \text{or} \quad Q_1 < (x_r)_{LSTAT} \leq Q_3 \quad \text{or} \quad (x_r)_{LSTAT} > Q_3\]

(a) Perturbations for \(x_1^0\)
(b) Perturbations for \(x_2^0\)

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 \( R < I \)

\[
\begin{align*}
\min_{x \in \mathcal{X}, y} & \quad \sum_{r=1}^{R} \sum_{s=1}^{S} y_{sr} \|x^0_s - x_r\|_2^2 \\
\text{s.t.} & \quad wx_r + b \geq \varphi^{-1}(\nu) \quad \forall r = 1, 2, \ldots, R \\
& \quad \sum_{r=1}^{R} y_{sr} = 1 \quad \forall s = 1, 2, \ldots, S \\
& \quad y_{sr} \in \{0, 1\} \quad \forall s = 1, 2, \ldots, S \quad \forall r = 1, 2, \ldots, R.
\end{align*}
\]
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 → Bilevel Optimization
Given a benchmarking model and an inefficient firm, we find a CE, i.e., a counterfactual firm with a better efficiency → 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 → 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 → **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 → **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

$$\min_{\hat{x}, E} C(x^0, \hat{x})$$

subject to

$$\hat{x} \in \mathbb{R}^l_+$$

$$E \geq E^*$$

$$E \in \arg \min_{\bar{E}, \lambda^0, ..., \lambda^K} \left\{ \bar{E} : \bar{E} \hat{x} \geq \sum_{k=0}^{K} \lambda^k x^k, \ y^0 \leq \sum_{k=0}^{K} \lambda^k y^k, \ \bar{E} \geq 0, \ \lambda \in \mathbb{R}^{K+1}_+ \right\}$$
From bilevel to single level

\[
\begin{align*}
\min_{\hat{x}, F, \beta, \gamma, u, v, w} & \quad C(x^0, \hat{x}) \\
\text{s.t.} & \quad F \leq F^* \\
& \quad \hat{x} \geq \sum_{k=0}^{K} \beta^k x^k \\
& \quad Fy^0 \leq \sum_{k=0}^{K} \beta^k y^k \\
& \quad \gamma I \leq M_I u \\
& \quad \hat{x} - \sum_{k=1}^{K} \beta^k x^k \leq M_I (1 - u) \\
& \quad \gamma O \leq M_O v \\
& \quad -Fy^0 + \sum_{k=1}^{K} \beta^k y^k \leq M_O (1 - v) \\
& \quad \beta \leq M_I w \\
& \quad \gamma^T_i y^k - \gamma^T O y^k \leq M_I (1 - w_k) \forall k \\
& \quad \gamma^T O y^0 = 1 \\
& \quad \gamma^T_i x^k - \gamma^T O y^k \geq 0 \quad k = 0, \ldots, K \\
& \quad \gamma I, \gamma O, F, \beta \geq 0 \\
& \quad u, v, w \in \{0, 1\} \\
& \quad \hat{x} \in \mathbb{R}^I_+ \\
\end{align*}
\]

With the cost function

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

we obtain a Mixed Integer Convex Quadratic with Linear Constraints formulation
From bilevel to single level

\[
\begin{align*}
\min_{\hat{x}, F, \beta, \gamma, u, v, w} \quad & C(x^0, \hat{x}) \\
\text{s.t.} \quad & F \leq F^* \\
& \hat{x} \geq \sum_{k=0}^{K} \beta^k x^k \\
& Fy^0 \leq \sum_{k=0}^{K} \beta^k y^k \\
& \gamma_l \leq M_l u \\
& \hat{x} - \sum_{k=1}^{K} \beta^k x^k \leq M_l (1 - u) \\
& \gamma_o \leq M_o v \\
& -Fy^0 + \sum_{k=1}^{K} \beta^k y^k \leq M_o (1 - v) \\
& \beta \leq M_t w \\
& \gamma_l^T x^k - \gamma_o^T y^k \leq M_t (1 - w_k) \forall k \\
& \gamma_o^T y^0 = 1 \\
& \gamma_l^T x^k - \gamma_o^T y^k \geq 0 \quad k = 0, \ldots, K \\
& \gamma_l, \gamma_o, F, \beta \geq 0 \\
& u, v, w \in \{0, 1\} \\
& \hat{x} \in \mathbb{R}^I_+ \\
\end{align*}
\]

With the cost function

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

we obtain a Mixed Integer Convex Quadratic with Linear Constraints formulation.
Results for banking branches
Results for banking branches

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

![Graphs showing accuracy, compression, number of features, and inconsistency for different discretization methods.](image)
Outline

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


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


References II


