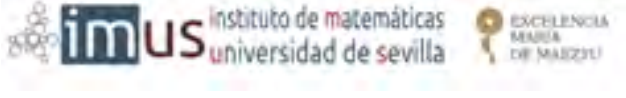
Counterfactual Explanations: Models for Actionable and Explainable Decision-Making

Emilio Carrizosa (IMUS, Universidad de Sevilla)
+ Renato de Leone (Università di Camerino) + Marica Magagnini (Università di Camerino) + Antonio Navas (IMUS, Universidad de Sevilla) + Jasone Ramírez-Ayerbe (Université de Montréal) + Dolores Romero Morales (Copenhagen Business School)

October 23rd, 2025



Keynote: Raiders of the Lost Interpretability

Thursday, August 28, 2025
2:15 PM - 2:55 PM
Penland Suite, JMCC

Overview

Emilio Carrizosa, Professor of Statistics and Operational Research, University of Seville

Details

Chair: Professor Christophe Mues

Speaker

Emilio Carrizosa
Professor of Statistics and Operational Research
University of Seville

Raiders of the Lost Interpretability

Abstract: The adoption of so called Explainable AI, which is typically 'black box' machine learning models accompanied by post-hoc explainability tools, are becoming more common; however, the concern remains for high risk area: can we trust post-hoc explainers?

Biography

Emilio Carrizosa is a data scientist, Professor of Statistics and Operational Research in the University of Seville (Spain) and President of the Spanish Network for Mathematics in Industry math-in

Keynote: Raiders of the Lost Interpretability

Thursday, August 28, 2025
2:15 PM - 2:55 PM
Fenland Suite, JMCC

Overview

Emilio Carrizosa, Professor of Statistics and Operational Research, University of Seville

Details

Chair: Professor Christophe Muller

Speaker

Emilio Carrizosa
Professor of Statistics and Operational Research
University of Seville

Raiders of the Lost Interpretability

Abstract: The adoption of so called Explainable AI, which is typically 'black box' machine learning models accompanied by post-hoc explainability tools, are becoming more common; however, the concern remains for high risk area: can we trust post-hoc explainers?

Biography

Emilio Carrizosa is a data scientist, Professor of Statistics and Operational Research in the University of Seville (Spain) and President of the Spanish Network for Mathematics in Industry math-in

UNIVERSITY OF EDINBURGH
Business School

CRC Credit Research Centre

Home | Conference XIX | Past Conferences | Research | Thought Leadership | News | MSc BMA | About

Home / News / Webinar - Professor Emilio Carrizosa returns to the CRC to finish what he started

Webinar - Professor Emilio Carrizosa returns to the CRC to finish what he started

Time for commercials

EURO Online Seminar Series on Operational Research and Machine Learning

To receive news from us, please register to the mailing list. You will then receive information about our upcoming seminars and the link to participate only. Our seminars are on Mondays, 16.30-17.30 CET, and we typically send the link on Monday morning.

And now, the talk ;)

From the simple to the complex

Jean-Baptiste Lamarck (1744-1829)

"La puissance de la vie tend continuellement à composer l'organisation. Ce pouvoir essentiel, inhérent à la vie, tend sans cesse à compliquer l'organisation".
(The power of life tends continually to build organization. This essential power, inherent in life, constantly strives to make organization more complex)

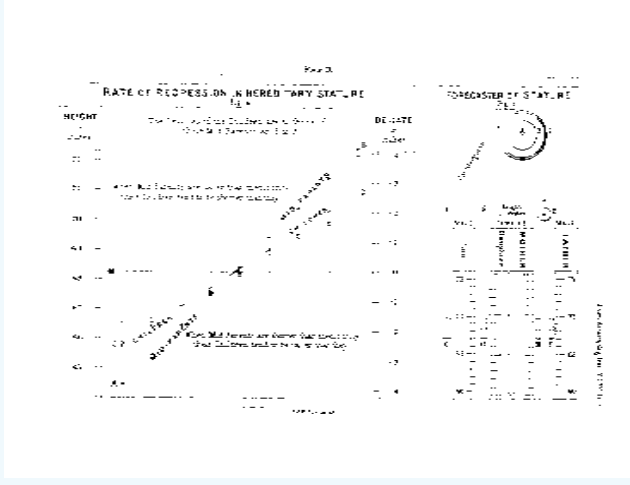
From the simple to the complex

Jean-Baptiste Lamarck (1744-1829)

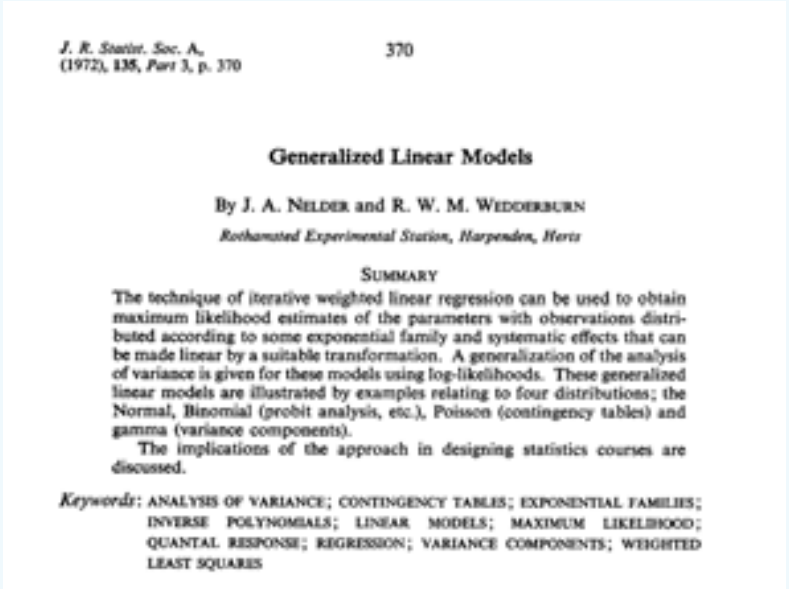
"La puissance de la vie tend continuellement à composer l'organisation. Ce pouvoir essentiel, inhérent à la vie, tend sans cesse à compliquer l'organisation".
(The power of life tends continually to build organization. This essential power, inherent in life, constantly strives to make organization more complex)

$$y(x_1, \ldots, x_p) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p$$

Johann Carl Friedrich Gauß
(1777-1855)



Francis Galton
(1822-1911)



J. R. Statist. Soc. A,
(1972), 135, Part 3, p. 370

370

Generalized Linear Models

By J. A. Nelder and R. W. M. Wedderburn

Rothamsted Experimental Station, Harpenden, Herts

Summary

The technique of iterative weighted linear regression can be used to obtain maximum likelihood estimates of the parameters with observations distributed according to some exponential family and systematic effects that can be made linear by a suitable transformation. A generalization of the analysis of variance is given for these models using log-likelihoods. These generalized linear models are illustrated by examples relating to four distributions; the Normal, Binomial (probit analysis, etc.), Poisson (contingency tables) and gamma (variance components).

The implications of the approach in designing statistics courses are discussed.

Keywords: Analysis of variance; Contingency tables; Exponential families; Inverse polynomials; Linear models; Maximum likelihood; Quantal response; Regression; Variance components; Weighted least squares

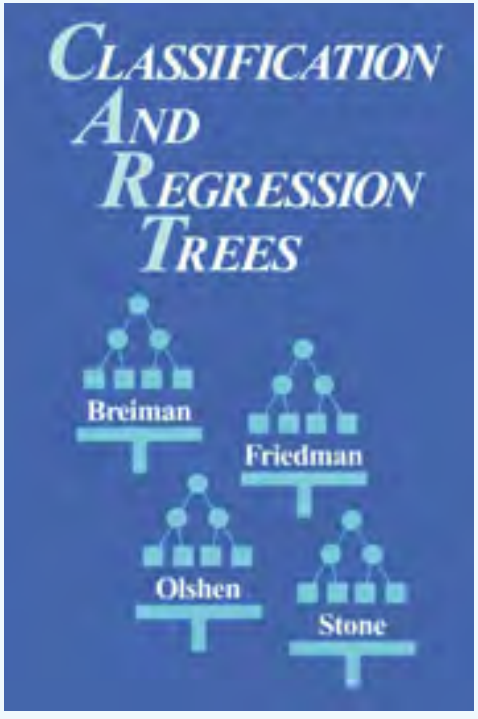
Generalized Linear Model (GLM)

$$g\left(E\left(Y|X=(x_1,\ldots,x_p)\right)\right)=\beta_1 x_1+\beta_2 x_2+\ldots+\beta_p x_p$$

g :↑↑

- Linear regression
- Logistic regression
- Beta regression
- Poisson regression
- ...

Classification and Regression Trees



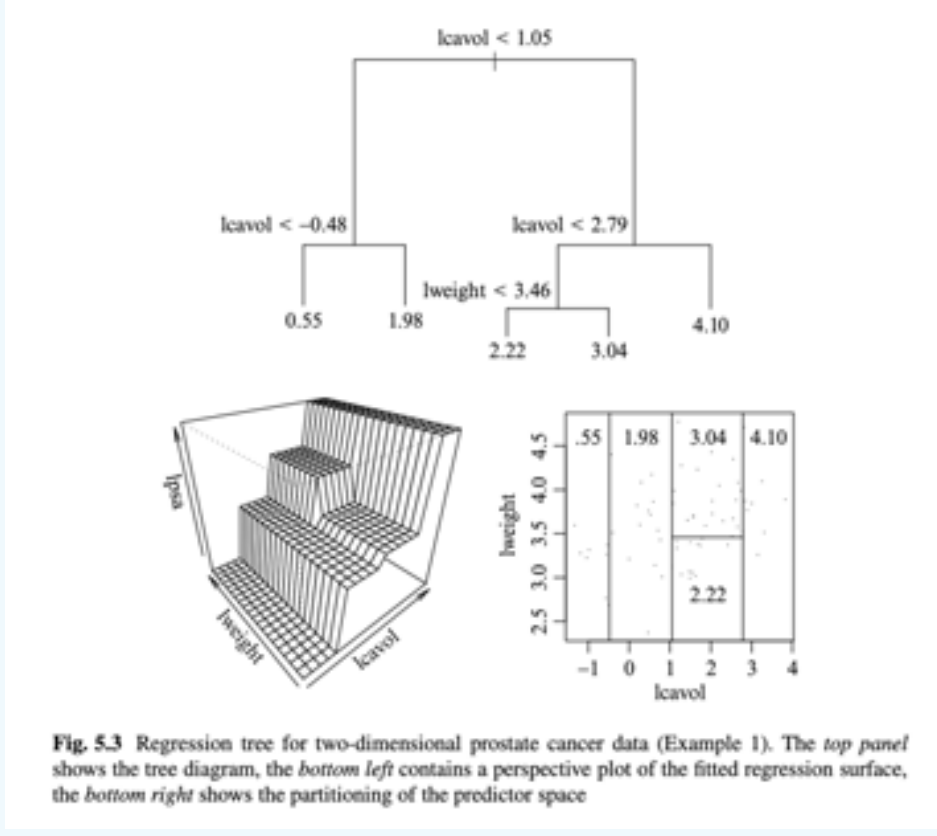
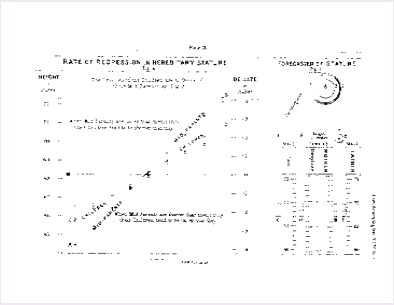


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

From: A. Cutler, D. R. Cutler, J. R. Stevens

19th-20th centuries



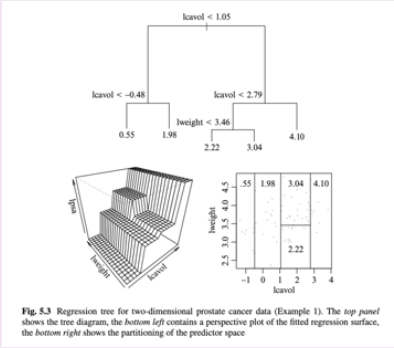
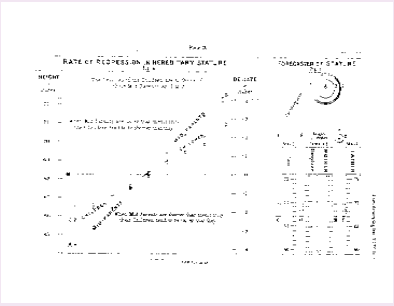


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

19th-20th centuries



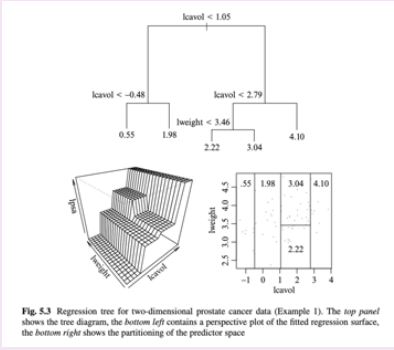
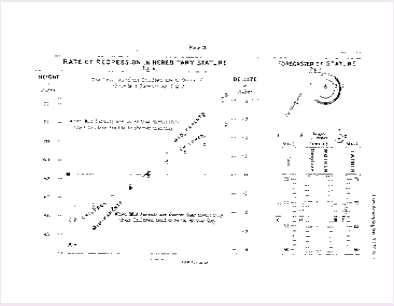


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Late 20th century - 21st century: The blackout

19th-20th centuries



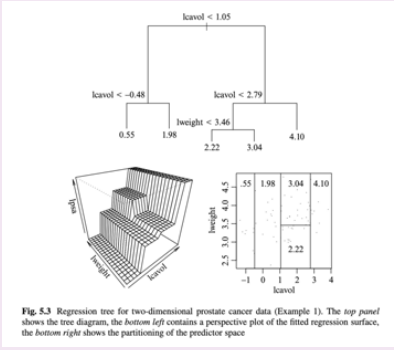
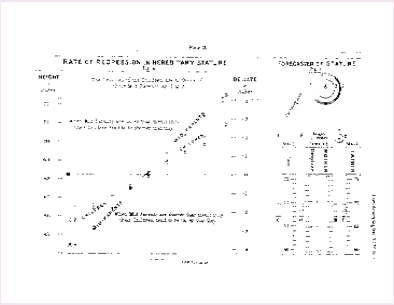


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Late 20th century - 21st century: The blackout

19th-20th centuries



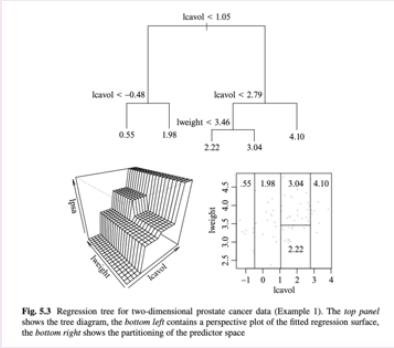


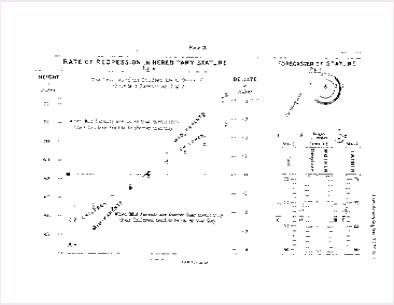
Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Nowadays, 21st century

Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).

19th-20th centuries



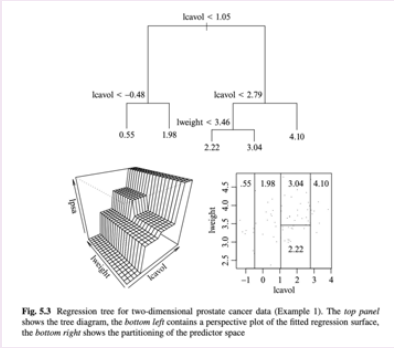


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

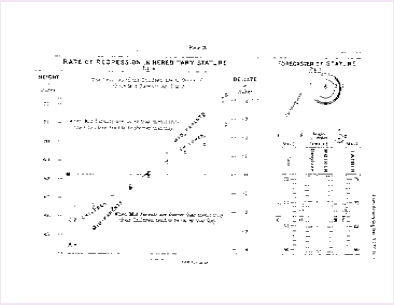
Nowadays, 21st century

Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).
- Different approaches to explainability, Molnar (2020):

- ...

19th-20th centuries



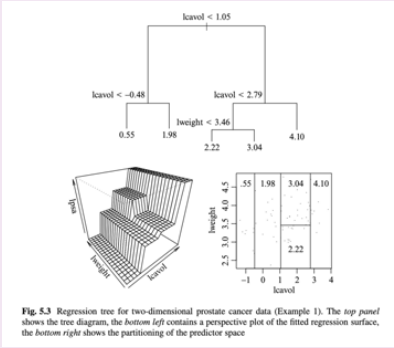


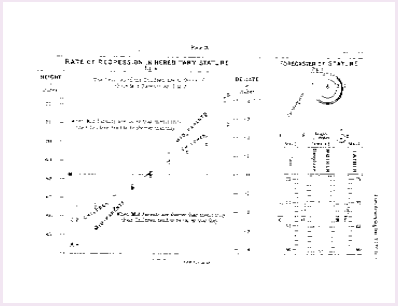
Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Nowadays, 21st century

Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).
- Different approaches to explainability, Molnar (2020):
  - using simple models e.g. Hastie et al. (2015); Friedman et al. (2010); Blanquero et al. (2021)
  - ...

19th-20th centuries



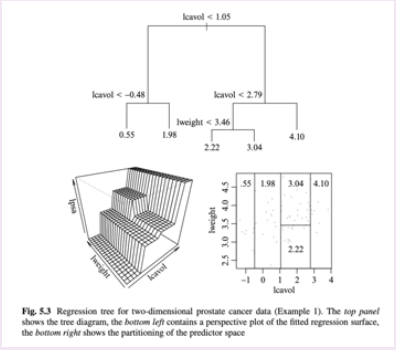


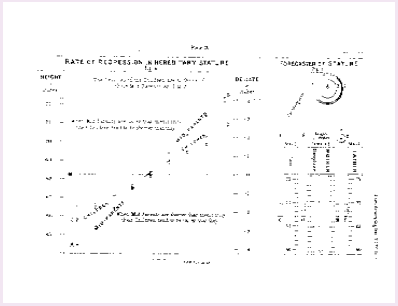
Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Nowadays, 21st century

Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).
- Different approaches to explainability, Molnar (2020):
  - using simple models e.g. Hastie et al. (2015); Friedman et al. (2010); Blanquero et al. (2021)
  - approximating by simple models, e.g. LIME (Local Interpretable Model-Agnostic Explanation), Ribeiro et al. (2016)
  - ...

19th-20th centuries



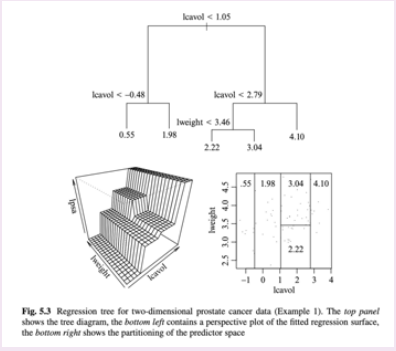


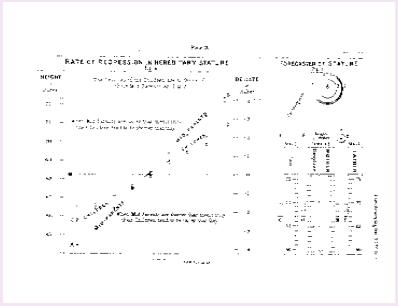
Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Nowadays, 21st century

Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).
- Different approaches to explainability, Molnar (2020):
  - using simple models e.g. Hastie et al. (2015); Friedman et al. (2010); Blanquero et al. (2021)
  - approximating by simple models, e.g. LIME (Local Interpretable Model-Agnostic Explanation), Ribeiro et al. (2016)
  - measuring the importance of attributes: SHapley Additive exPlanations (SHAP), Lundberg and Lee (2017); Permutation Importance, Fisher et al. (2019);
  - ...

19th-20th centuries



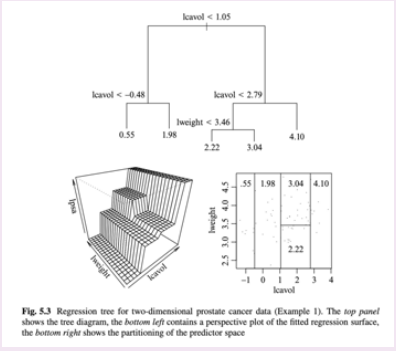


Fig. 5.3 Regression tree for two-dimensional prostate cancer data (Example 1). The *top panel* shows the tree diagram, the *bottom left* contains a perspective plot of the fitted regression surface, the *bottom right* shows the partitioning of the predictor space

Nowadays, 21st century

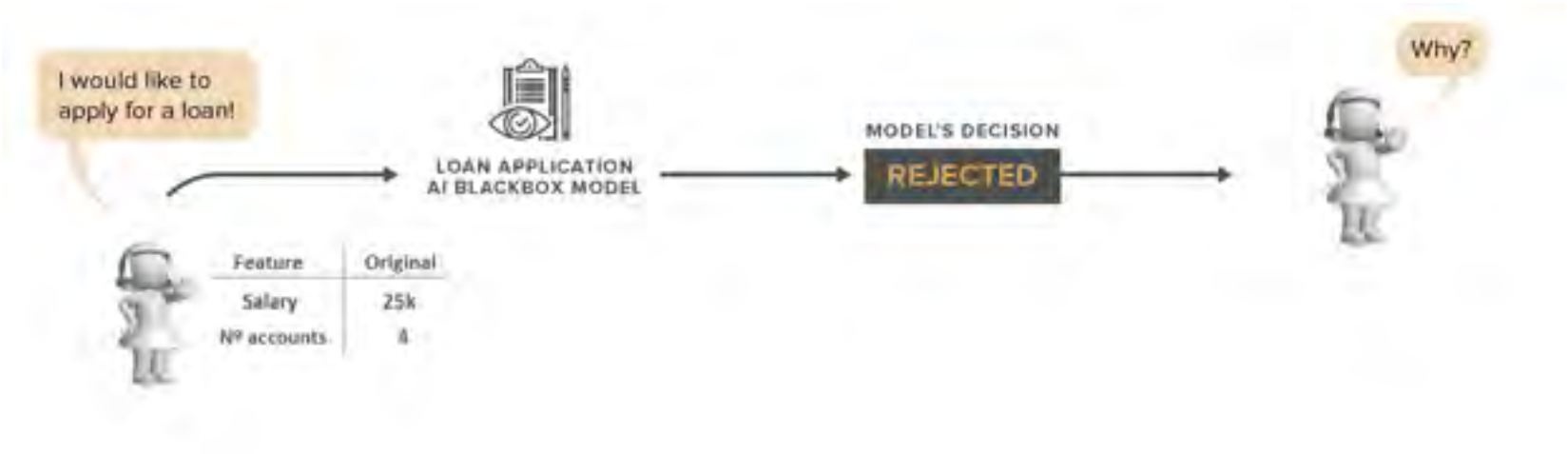
Seeking explainability

- Machine Learning applied more and more in high stakes decision making Zafar et al. (2017); Rudin et al. (2022), even regulated, e.g. EU AI Act Panigutti et al. (2023).
- Different approaches to explainability, Molnar (2020):
  - using simple models e.g. Hastie et al. (2015); Friedman et al. (2010); Blanquero et al. (2021)
  - approximating by simple models, e.g. LIME (Local Interpretable Model-Agnostic Explanation), Ribeiro et al. (2016)
  - measuring the importance of attributes: SHapley Additive exPlanations (SHAP), Lundberg and Lee (2017); Permutation Importance, Fisher et al. (2019);
  - counterfactual explanations, e.g. Wachter et al. (2017); Carrizosa et al. (2024b); Kurtz et al. (2024); Maragno et al. (2024)
  - ...

Optimization

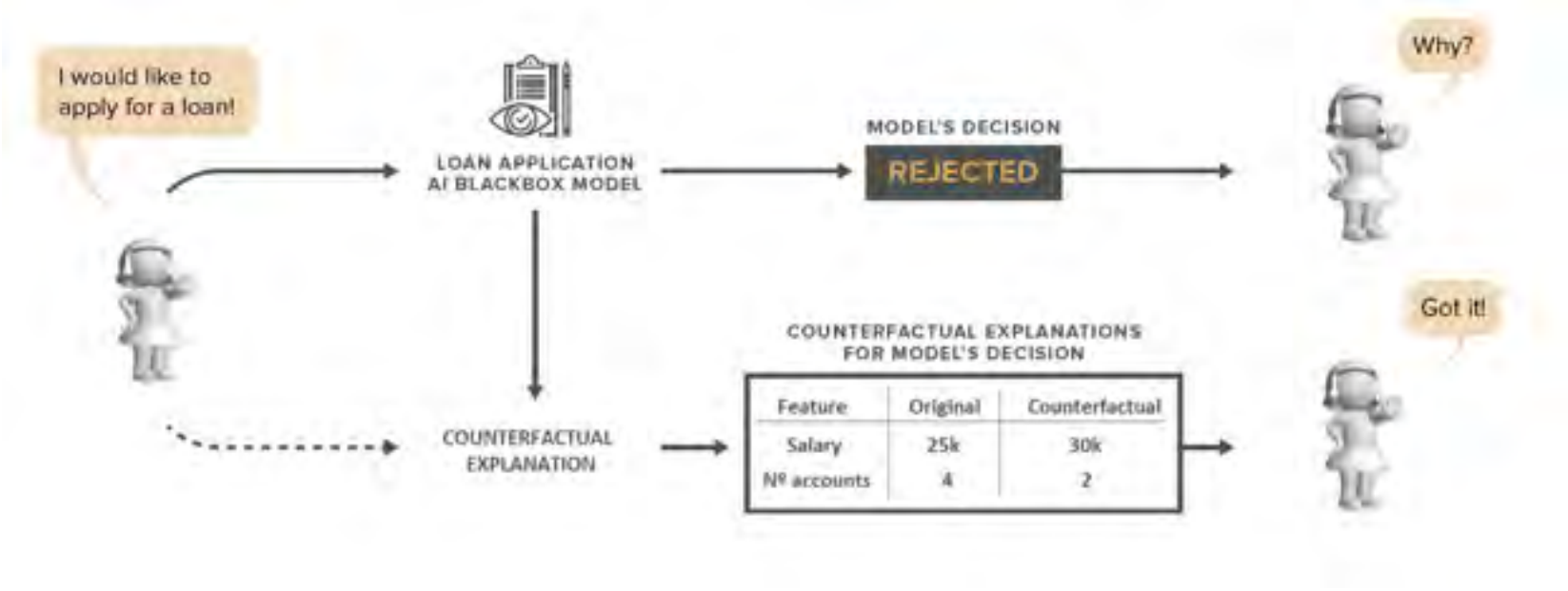
Source: This Week, 1951

Counterfactual Explanations. Motivation



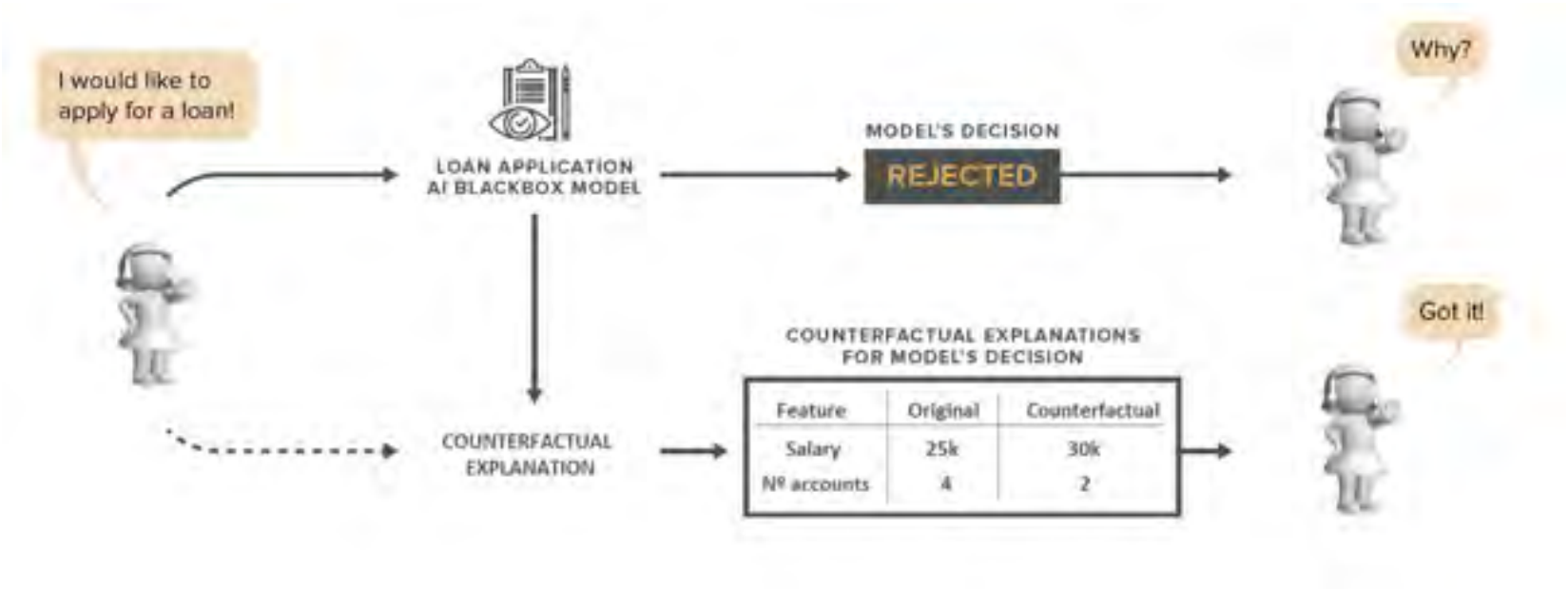
Your loan has been denied

Counterfactual Explanations. Motivation



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

Counterfactual Explanations. Motivation



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: The classifier would have given a probability of $+$ above 0.8, which is the threshold defined by the bank

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models
- $C(\boldsymbol{x}_0, \boldsymbol{x}) = \text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x})$, with $\lambda_c > 0$.

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models
- $C(\boldsymbol{x}_0, \boldsymbol{x}) = \text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x})$, with $\lambda_c > 0$.
  - $\text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x}) = \left|\{j : x_j \neq x_{0j}\}\right|$

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models
- $C(\boldsymbol{x}_0, \boldsymbol{x}) = \text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x})$, with $\lambda_c > 0$.
  - $\text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x}) = \left|\{j : x_j \neq x_{0j}\}\right|$
  - $\text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) = \pi(\sigma(\boldsymbol{x} - \boldsymbol{x}_0))$, where

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models
- $C(\boldsymbol{x}_0, \boldsymbol{x}) = \text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x})$, with $\lambda_c > 0$.
  - $\text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x}) = \left|\{j : x_j \neq x_{0j}\}\right|$
  - $\text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) = \pi(\sigma(\boldsymbol{x} - \boldsymbol{x}_0))$, where
    - π : convex $\uparrow$ in $\mathbb{R}_+$,

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Counterfactual Explanations. Ingredients

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^p$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,

find the changes (to some $\boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \boldsymbol{x})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$

$$\begin{array}{ll} \min_{\boldsymbol{x}} & C(\boldsymbol{x}_0, \boldsymbol{x}) \\ \text{s.t.} & P(\boldsymbol{x}) \geq \tau \\ & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array} \qquad \begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -P(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

- $\mathcal{X}(\boldsymbol{x}_0)$ (see (Smyth and Keane, 2022)):
  - Endogenous: Points from some training set $\longrightarrow$ discrete optimization models
  - Exogenous: Synthetic data $\longrightarrow$ (mixed integer) nonlinear optimization models
- $C(\boldsymbol{x}_0, \boldsymbol{x}) = \text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x})$, with $\lambda_c > 0$.
  - $\text{Complexity}(\boldsymbol{x}_0, \boldsymbol{x}) = \left|\{j : x_j \neq x_{0j}\}\right|$
  - $\text{Dissimilarity}(\boldsymbol{x}_0, \boldsymbol{x}) = \pi(\sigma(\boldsymbol{x} - \boldsymbol{x}_0))$, where
    - π : convex $\uparrow$ in $\mathbb{R}_+$,
    - σ : gauge in $\mathbb{R}^p$ (definite positive, positively homogeneous and subadditive), e.g. norms (not a good idea!!!), quantile gauges, (Carrizosa et al., 2024b) or skewed norms, (Plastria, 1992; Drezner and Drezner, 2021)

(Artelt and Hammer, 2019; Carrizosa et al., 2024b; Guidotti, 2022; Stepin et al., 2021; Verma et al., 2021)

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)
  - Additive trees (e.g. Random Forests, XGBoost, ...)

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)
  - Additive trees (e.g. Random Forests, XGBoost, ...)

From Scores to Probabilities

- Probability of $\boldsymbol{x}$ being classified in the positive class:

$$P(\boldsymbol{x}) = g(f(\boldsymbol{x})),$$

with $g : \uparrow$ (link function)

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)
  - Additive trees (e.g. Random Forests, XGBoost, ...)

From Scores to Probabilities

- Probability of $\boldsymbol{x}$ being classified in the positive class:

$$P(\boldsymbol{x}) = g(f(\boldsymbol{x})),$$

  with $g : \uparrow$ (link function)
- Example: Logistic regression: $g(t) = \frac{1}{1+e^{-t}}$ (applied to a linear f)

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)
  - Additive trees (e.g. Random Forests, XGBoost, ...)

From Scores to Probabilities

- Probability of $\boldsymbol{x}$ being classified in the positive class:

$$P(\boldsymbol{x}) = g(f(\boldsymbol{x})),$$

  with $g : \uparrow$ (link function)
- Example: Logistic regression: $g(t) = \frac{1}{1+e^{-t}}$ (applied to a linear f)
- Example: SVM. g sigmoidal, (Benítez-Peña et al., 2024; Platt, 1999), with parameters estimated via maximum likelihood from a training sample, and f : linear combination of kernels

Score-based classifiers

(Carrizosa et al., 2021; Gambella et al., 2021)

- Score function $f : \mathcal{X} \longrightarrow \mathbb{R}$.
- Example: linear classifiers, $f(\boldsymbol{x}) = \boldsymbol{\beta}^\top \boldsymbol{x} + b$.
  - Logistic regression
  - Support Vector Machines (with linear kernel)
  - Additive trees (e.g. Random Forests, XGBoost, ...)

From Scores to Probabilities

- Probability of $\boldsymbol{x}$ being classified in the positive class:

$$P(\boldsymbol{x}) = g(f(\boldsymbol{x})),$$

  with $g : \uparrow$ (link function)
- Example: Logistic regression: $g(t) = \frac{1}{1+e^{-t}}$ (applied to a linear f)
- Example: SVM. g sigmoidal, (Benítez-Peña et al., 2024; Platt, 1999), with parameters estimated via maximum likelihood from a training sample, and f : linear combination of kernels

$$\begin{array}{ll} \min_{\boldsymbol{x}} & (C(\boldsymbol{x}_0, \boldsymbol{x}), -f(\boldsymbol{x})) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Model valid for classification; also for regression, (Carrizosa and Navas-Orozco, 2025b)

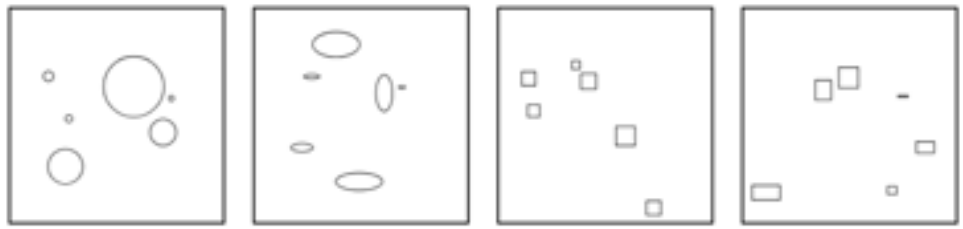
Numerical approaches:

- smooth optimization, e.g., (Joshi et al., 2019; Ramakrishnan et al., 2020; Wachter et al., 2017; Mothilal et al., 2020; Lucic et al., 2019)
- mixed integer optimization, e.g., (Bogetoft et al., 2024; Carrizosa et al., 2024a,c; Contardo et al., 2024; Cui et al., 2015; Fischetti and Jo, 2018; Kanamori et al., 2020, 2021; Magagnini et al., 2025b; Maragno et al., 2022; Parmentier and Vidal, 2021; Russell, 2019)
- multi-objective optimization, e.g., (Dandl et al., 2020; Del Ser et al., 2022; Raimundo et al., 2022),
- robust optimization, e.g., (Maragno et al., 2024; Virgolin and Fracaros, 2023),
- heuristic and metaheuristic approaches, e.g., (Carrizosa and Navas-Orozco, 2025b; Guidotti et al., 2019; Karimi et al., 2021; Magagnini et al., 2025a; Poyiadzi et al., 2020)

Beyond Counterfactual Explanations:

Robust CEs

(Magagnini et al., 2025a; Carrizosa and Navas-Orozco, 2025a,b)



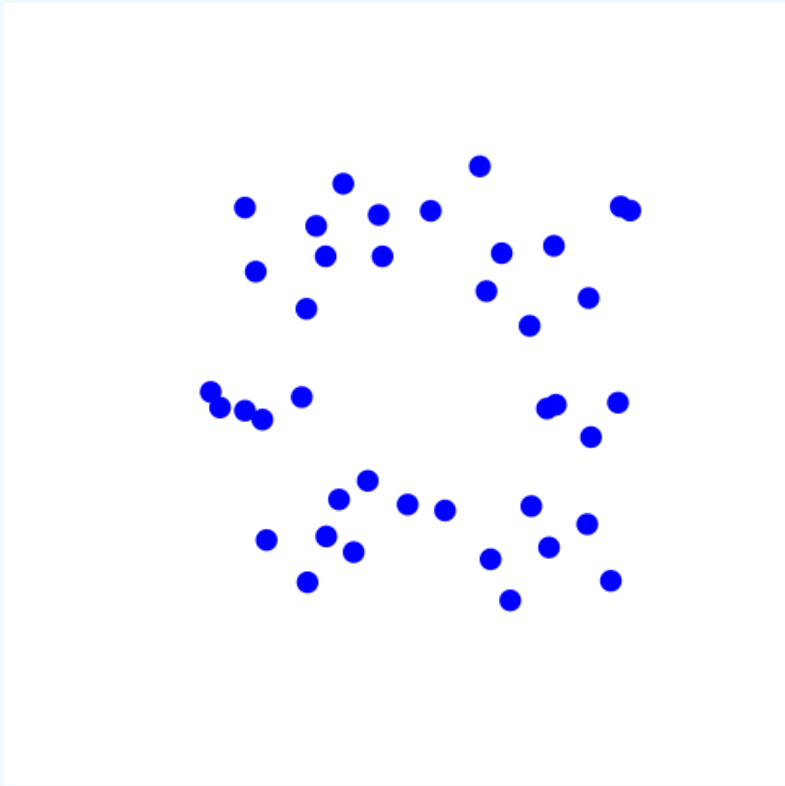
Uncertainty on the data

- Data points replaced by sets (e.g. convex compact sets symmetric w.r.t. the record)
- For k-NN as prediction model, the Integer Programming formulation of (Contardo et al., 2024) is extended in (Magagnini et al., 2025b), and solved with a Gaussian Variable Neighborhood Search as in (Carrizosa et al., 2012).

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -g^{-1}(\beta^\top \boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

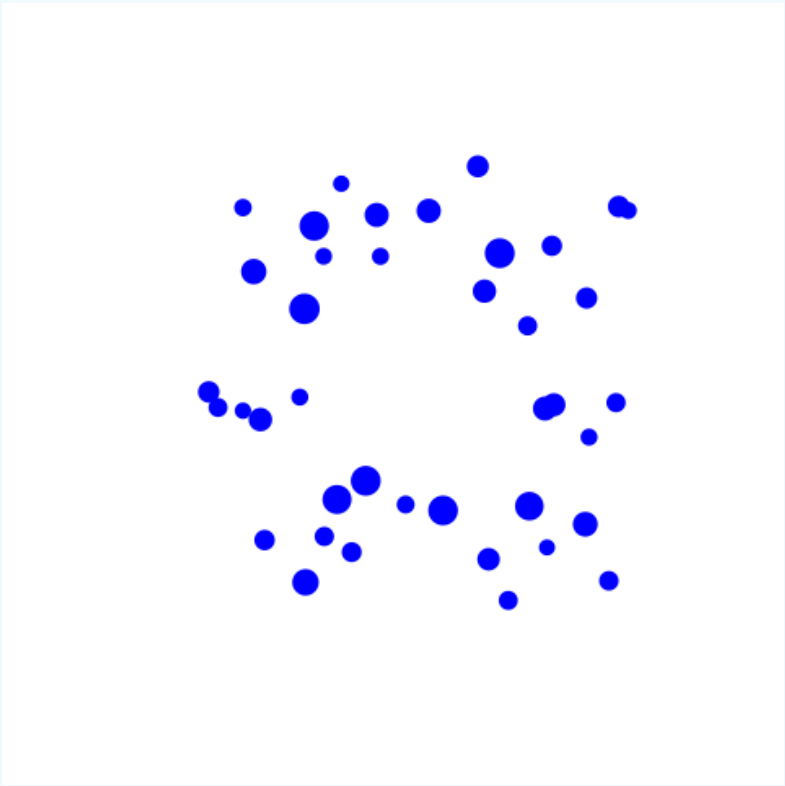


$$\beta = (2.07, -0.50)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -g^{-1}(\beta^\top \boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

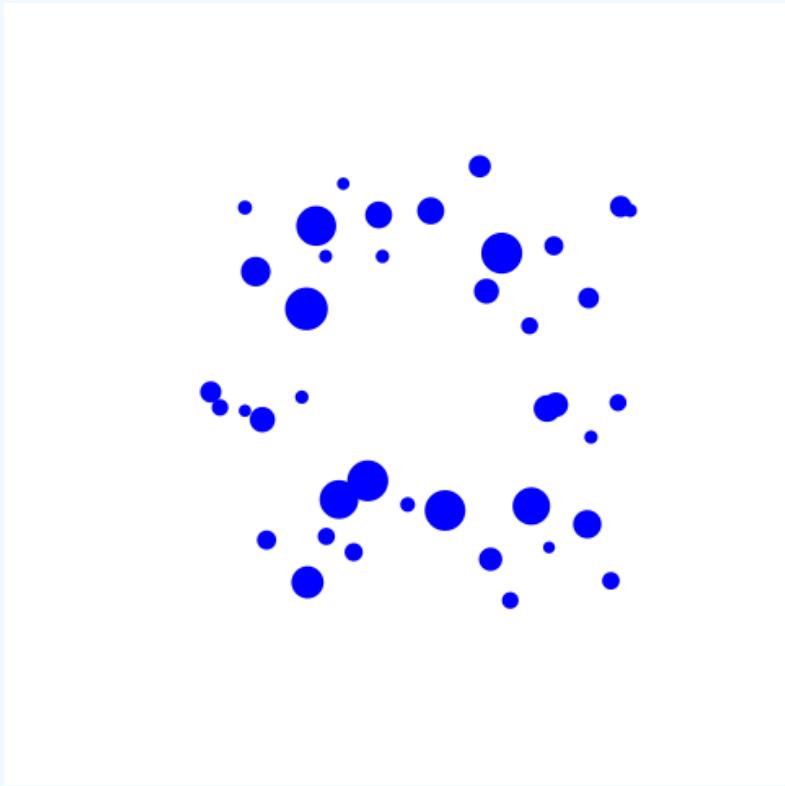


$$\beta = (2.09, -0.44)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -g^{-1}(\beta^\top \boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

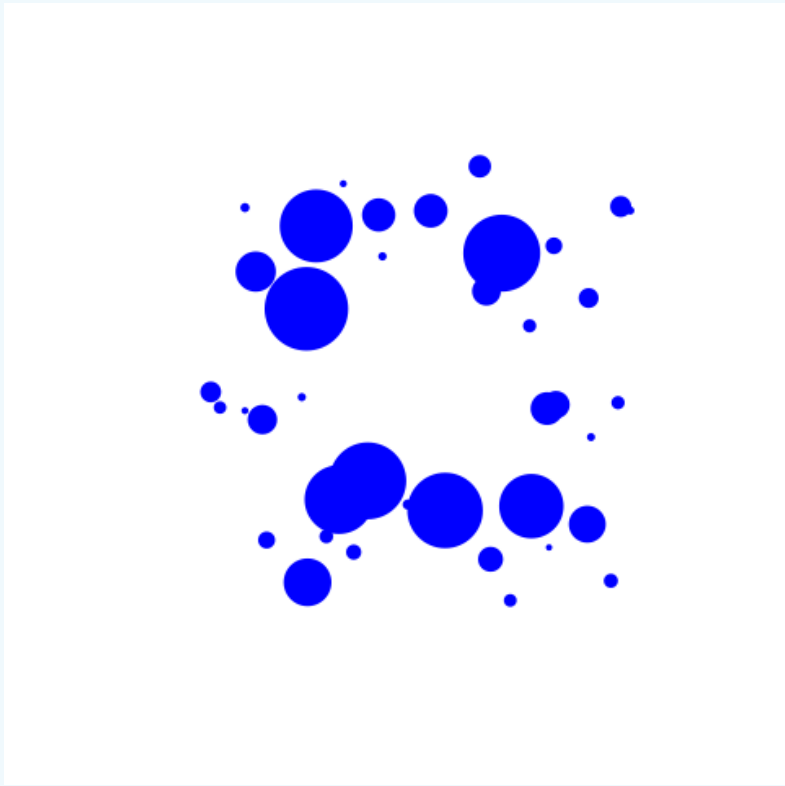


$$\beta = (2.12, -0.36)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -g^{-1}(\beta^\top \boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

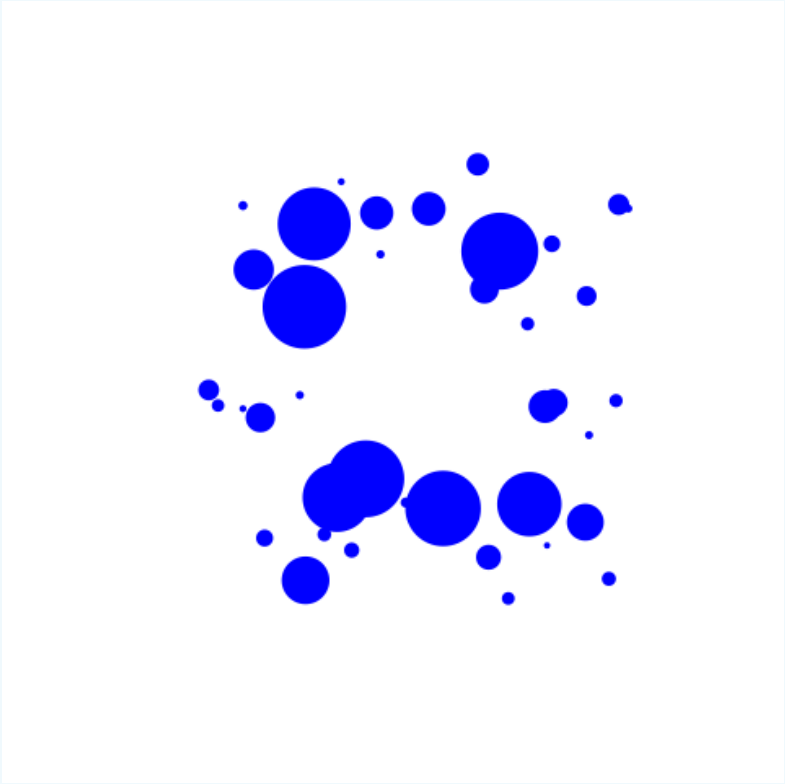


$$\beta = (2.17, -0.18)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{x} & \left(C(x_0, x), \max_{\beta \in \mathcal{B}} -g^{-1}(\beta^\top x)\right) \\ \text{s.t.} & x \in \mathcal{X}(x_0) \end{array}$$

Sample

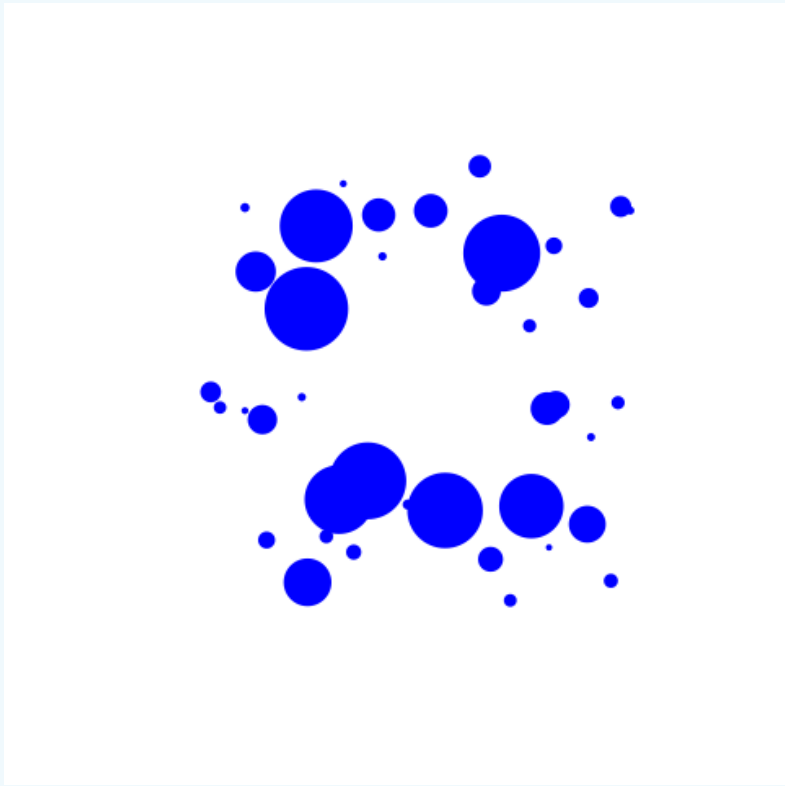


$$\beta = (2.17, -0.18)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -g^{-1}(\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x})\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

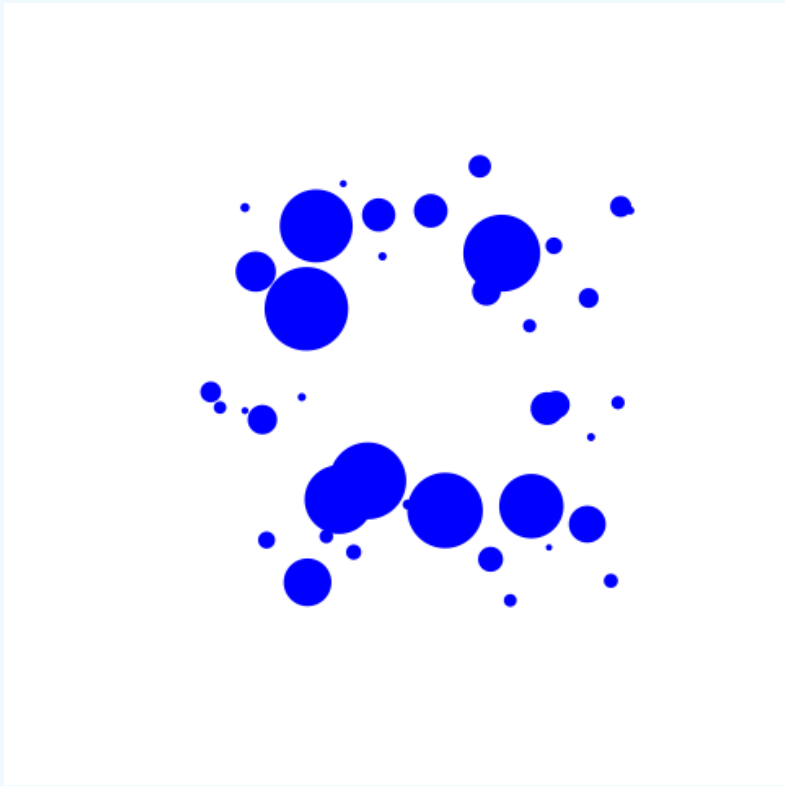


$$\beta = (2.17, -0.18)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \min_{\boldsymbol{x}} & \left(C(\boldsymbol{x}_0, \boldsymbol{x}), -\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}\right) \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \end{array}$$

Sample

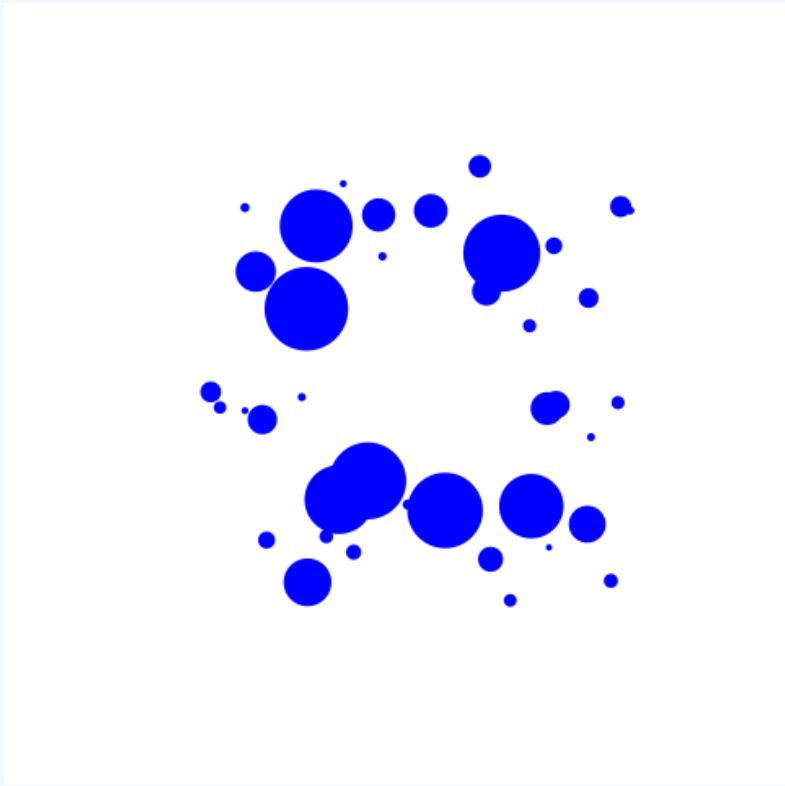


$$\beta = (2.17, -0.18)$$

Uncertainty in the model. Linear scoring function (GLM)

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

Sample



$$\beta = (2.17, -0.18)$$

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

The choice of $\mathcal{B}$

Set of Maximum Likelihood Estimates β when the data distribution P is at "distance" at most κ from the empirical distribution $P_{\boldsymbol{\omega}_0}$.

- Distributions P considered: those with same (finite) support as $P_{\boldsymbol{\omega}_0}$, therefore identified by the probability vector $\boldsymbol{\omega}$

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

The choice of $\mathcal{B}$

Set of Maximum Likelihood Estimates β when the data distribution P is at "distance" at most κ from the empirical distribution $P_{\boldsymbol{\omega}_0}$.

- Distributions P considered: those with same (finite) support as $P_{\boldsymbol{\omega}_0}$, therefore identified by the probability vector $\boldsymbol{\omega}$
- $D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0})$ (Kullback-Leibler divergence)

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

The choice of $\mathcal{B}$

Set of Maximum Likelihood Estimates β when the data distribution P is at "distance" at most κ from the empirical distribution $P_{\boldsymbol{\omega}_0}$.

- Distributions P considered: those with same (finite) support as $P_{\boldsymbol{\omega}_0}$, therefore identified by the probability vector $\boldsymbol{\omega}$
- $D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0})$ (Kullback-Leibler divergence)
- $\mathcal{L}_{\boldsymbol{\omega}}(\beta)$

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

The choice of $\mathcal{B}$

Set of Maximum Likelihood Estimates β when the data distribution P is at "distance" at most κ from the empirical distribution $P_{\boldsymbol{\omega}_0}$.

- Distributions P considered: those with same (finite) support as $P_{\boldsymbol{\omega}_0}$, therefore identified by the probability vector $\boldsymbol{\omega}$
- $D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0})$ (Kullback-Leibler divergence)
- $\mathcal{L}_{\boldsymbol{\omega}}(\beta)$
- $\mathcal{B} = \left\{ \arg\max_{\beta} \mathcal{L}_{\boldsymbol{\omega}}(\beta) \text{ for some } \boldsymbol{\omega}, D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0}) \leq \kappa \right\}$

$$\begin{array}{ll} \max_{\boldsymbol{x}} & \min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x} \\ \text{s.t.} & \boldsymbol{x} \in \mathcal{X}(\boldsymbol{x}_0) \\ & C(\boldsymbol{x}_0, \boldsymbol{x}) \leq \tau \end{array}$$

- Objective function: concave
- Subgradients at $\boldsymbol{x}$ obtained by solving $\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$
- Cutting planes

The choice of $\mathcal{B}$

Set of Maximum Likelihood Estimates β when the data distribution P is at "distance" at most κ from the empirical distribution $P_{\boldsymbol{\omega}_0}$.

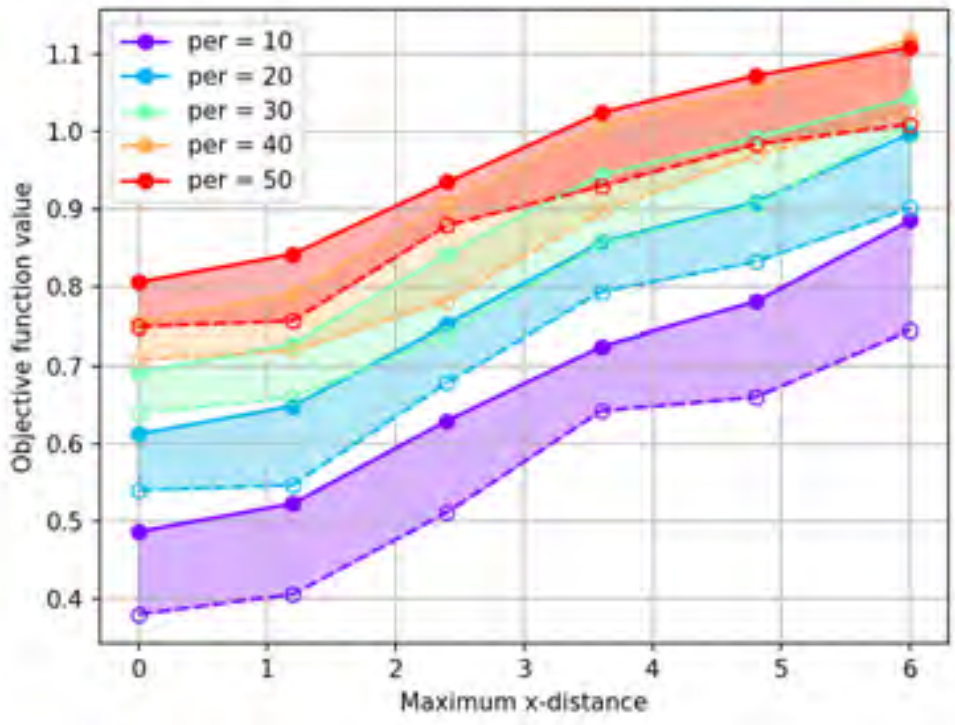
- Distributions P considered: those with same (finite) support as $P_{\boldsymbol{\omega}_0}$, therefore identified by the probability vector $\boldsymbol{\omega}$
- $D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0})$ (Kullback-Leibler divergence)
- $\mathcal{L}_{\boldsymbol{\omega}}(\beta)$
- $\mathcal{B} = \left\{ \arg\max_\beta \mathcal{L}_{\boldsymbol{\omega}}(\beta) \text{ for some } \boldsymbol{\omega}, D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0}) \leq \kappa \right\}$
- Assumptions (strong concavity, coercivity, smoothness) are imposed on $\mathcal{L}_{\boldsymbol{\omega}}(\beta)$ to have $\mathcal{B}$ well defined.

$$\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$$

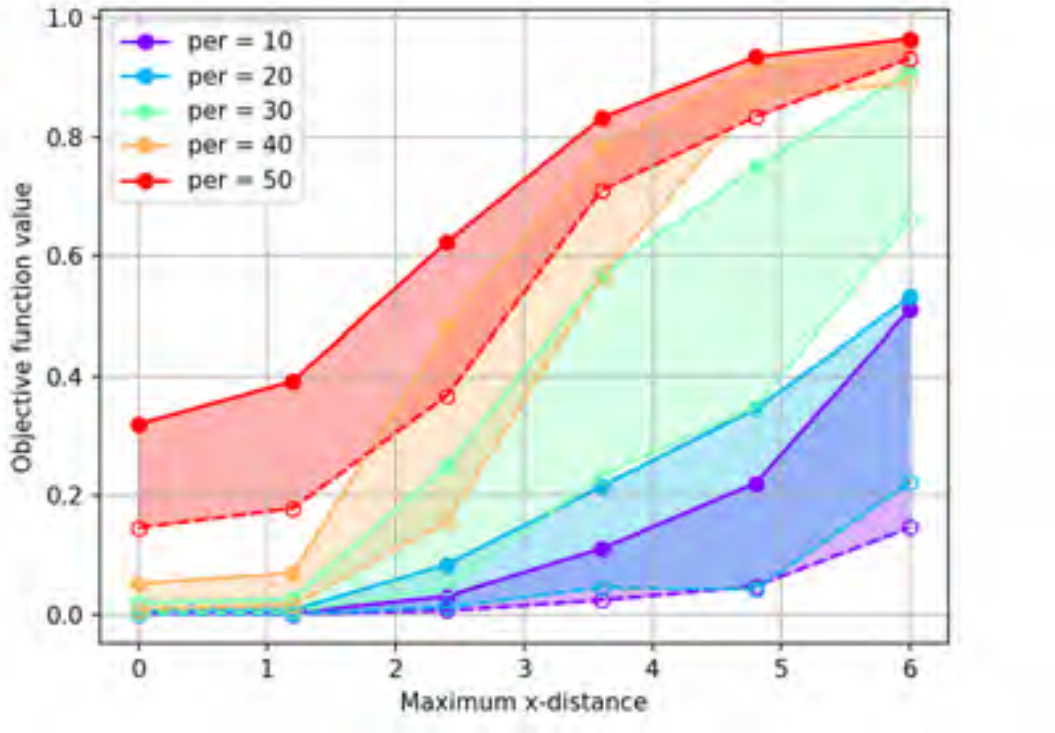
$$\begin{array}{ll} \min_{\boldsymbol{x},\boldsymbol{\omega}} & \beta^\top \boldsymbol{x} \\ \text{s.t.} & \beta \in \arg\max_\beta \mathcal{L}_{\boldsymbol{\omega}}(\beta) \\ & D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0}) \leq \kappa \end{array}$$

$$\min_{\beta \in \mathcal{B}} \beta^\top \boldsymbol{x}$$

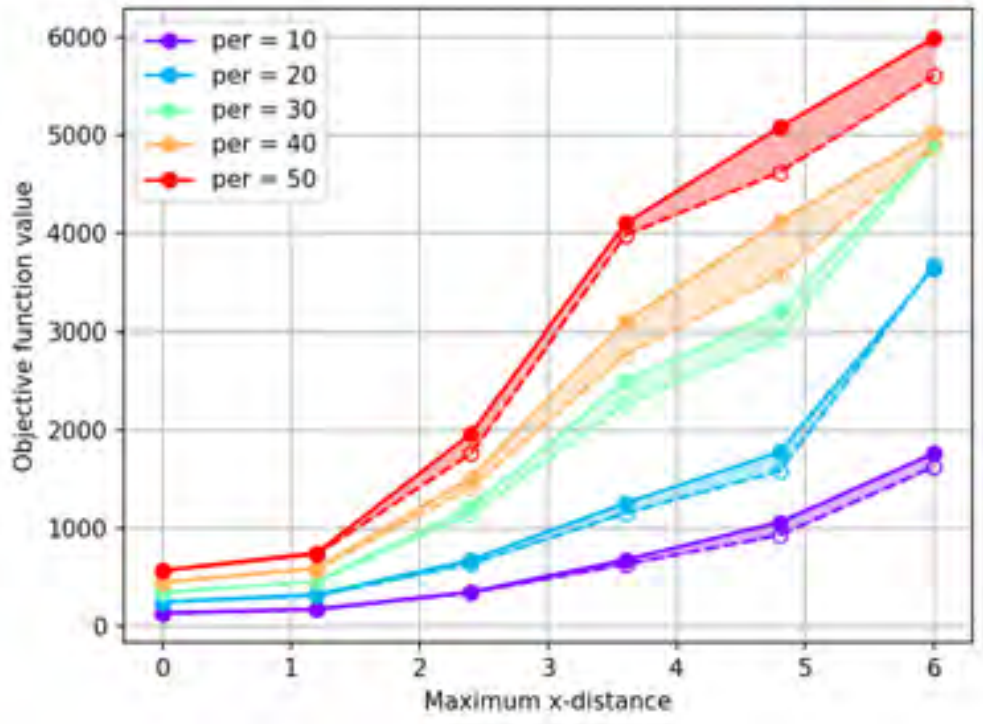
$$\begin{array}{ll} \min_{\boldsymbol{x},\boldsymbol{\omega}} & \beta^\top \boldsymbol{x} \\ \text{s.t.} & \nabla_\beta \mathcal{L}_{\boldsymbol{\omega}}(\beta) = 0 \\ & D_{KL}(P_{\boldsymbol{\omega}}, P_{\boldsymbol{\omega}_0}) \leq \kappa \end{array}$$



Communities and crime (Linear regression) $\kappa = 0, \kappa = 1$.



Breast cancer Wisconsin (Logistic regression) $\kappa = 0, \kappa = 1$.



Seoul bike sharing (Poisson regression) $\kappa = 0, \kappa = 1$.

Beyond Counterfactual Explanations:

Counterfactual Plans

(Carrizosa et al., 2025)

Counterfactual Plans

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^J$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,
- **probability threshold values** $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_R$,

Counterfactual Plans

Given

- a binary classification problem on $\mathcal{X} \subset \mathbb{R}^J$, with classes $\{+, -\}$ (class $+$: the good guys)
- a probabilistic classifier $P : \mathcal{X} \to [0, 1]$, $P(\boldsymbol{x})$: probability of belonging to class $+$
- $\boldsymbol{x}_0 \in \mathcal{X}$,
- **probability threshold values** $\tau_1 < \tau_2 < \ldots < \tau_R$,

find the changes (**sequentially** to some $\underline{\boldsymbol{x}} := (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_R) \in \underline{\mathcal{X}}(\boldsymbol{x}_0)$) with minimum cost $C(\boldsymbol{x}_0, \underline{\boldsymbol{x}})$ that cause $\boldsymbol{x}_0$ to increase their probability from $P(\boldsymbol{x}_0)$ to τ_1, then to τ_2, ..., and finally to τ_R

$$\begin{array}{lll} \min_{\underline{\boldsymbol{x}}} & C(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) & \\ \text{s.t.} & P(\boldsymbol{x}_j) \geq \tau_j & j = 1, 2, \ldots, R \\ & \underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\boldsymbol{x}_0) & \end{array}$$

Cost function C

$$\boldsymbol{C}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}),$$

- $\text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}))$, with $\omega_r > 0$, π : convex $\uparrow$ in $\mathbb{R}_+$, σ : quantile gauge in $\mathbb{R}^J$

Cost function C

$$\boldsymbol{C}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}),$$

- $\text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}))$, with $\omega_r > 0$, π : convex $\uparrow$ in $\mathbb{R}_+$, σ : quantile gauge in $\mathbb{R}^J$
- $\text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \left| \{ j : x_{rj} \neq x_{r-1_j} \text{ for at least one } r \} \right|$

Cost function C

$$\boldsymbol{C}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}),$$

- $\text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}))$, with $\omega_r > 0$, π : convex $\uparrow$ in $\mathbb{R}_+$, σ : quantile gauge in $\mathbb{R}^J$
- $\text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \left| \{ j : x_{rj} \neq x_{r-1_j} \text{ for at least one } r \} \right|$

$$\begin{array}{ll} \min_{\underline{\boldsymbol{x}}} & \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1})) + \lambda \sum_{j=1}^{J} t_j \\ \text{s.t.} & f(\boldsymbol{x}_r) \geq g^{-1}(\tau_r) \qquad r = 1, 2, \ldots, R \\ & (1 - t_j)\left(\boldsymbol{x}_{rj} - \boldsymbol{x}_{r-1_j}\right) = 0 \qquad r = 1, 2, \ldots, R,\ j = 1, 2, \ldots, J \\ & \underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\boldsymbol{x}_0) \\ & t_j \in \{0, 1\} \qquad j = 1, 2, \ldots, J \end{array}$$

Cost function C

$$\boldsymbol{C}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}),$$

- $\text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}))$, with $\omega_r > 0$, π : convex $\uparrow$ in $\mathbb{R}_+$, σ : quantile gauge in $\mathbb{R}^J$
- $\text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \left| \{ j : x_{rj} \neq x_{r-1_j} \text{ for at least one } r \} \right|$

$$\begin{array}{ll} \min_{\underline{\boldsymbol{x}}} & \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1})) + \lambda \sum_{j=1}^{J} t_j \\ \text{s.t.} & f(\boldsymbol{x}_r) \geq g^{-1}(\tau_r) \qquad r = 1, 2, \ldots, R \\ & (1 - t_j)\left(\boldsymbol{x}_{rj} - \boldsymbol{x}_{r-1_j}\right) = 0 \qquad r = 1, 2, \ldots, R,\ j = 1, 2, \ldots, J \\ & \underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\boldsymbol{x}_0) \\ & t_j \in \{0, 1\} \qquad j = 1, 2, \ldots, J \end{array}$$

Linear objective if σ_0 : ℓ_1 and π : affine

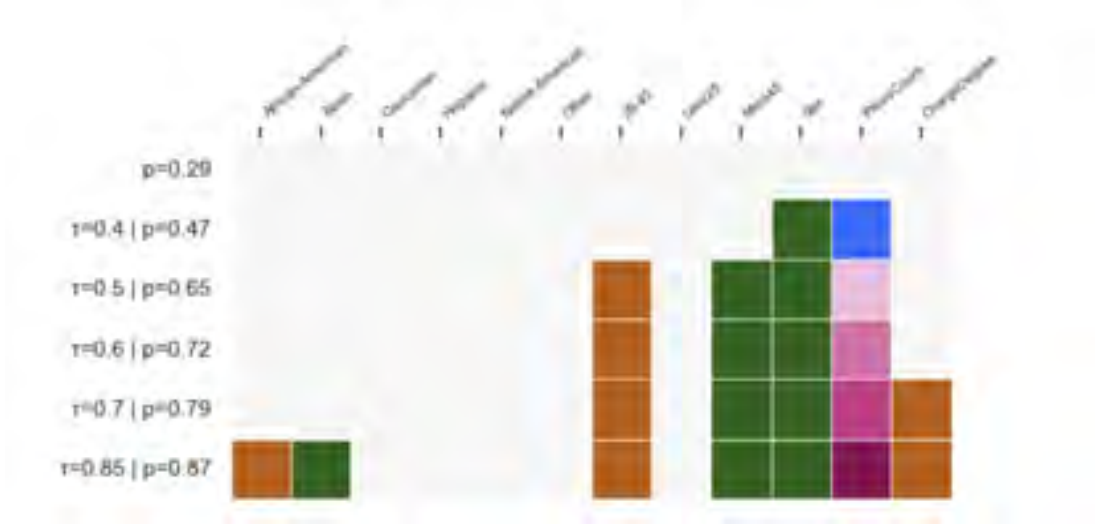
Cost function C

$$\boldsymbol{C}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) + \lambda_c \text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}),$$

- $\text{Dissimilarity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1}))$, with $\omega_r > 0$, π : convex $\uparrow$ in $\mathbb{R}_+$, σ : quantile gauge in $\mathbb{R}^J$
- $\text{Complexity}(\boldsymbol{x}_0, \underline{\boldsymbol{x}}) = \left| \{ j : x_{rj} \neq x_{r-1_j} \text{ for at least one } r \} \right|$

$$\begin{array}{ll} \min_{\underline{\boldsymbol{x}}} & \sum_{r=1}^{R} \omega_r \pi(\sigma(\boldsymbol{x}_r - \boldsymbol{x}_{r-1})) + \lambda \sum_{j=1}^{J} t_j \\ \text{s.t.} & f(\boldsymbol{x}_r) \geq g^{-1}(\tau_r) \qquad r = 1, 2, \ldots, R \\ & (1 - t_j)\left(\boldsymbol{x}_{rj} - \boldsymbol{x}_{r-1_j}\right) = 0 \qquad r = 1, 2, \ldots, R,\ j = 1, 2, \ldots, J \\ & \underline{\boldsymbol{x}} \in \underline{\mathcal{X}}(\boldsymbol{x}_0) \\ & t_j \in \{0, 1\} \qquad j = 1, 2, \ldots, J \end{array}$$

Quadratic convex objective if $\sigma_0 : \ell_2$ and $\pi(t) = t^2$



Logistic regression. $R = 5$ steps. `COMPAS` dataset.

Beyond Counterfactual Explanations:

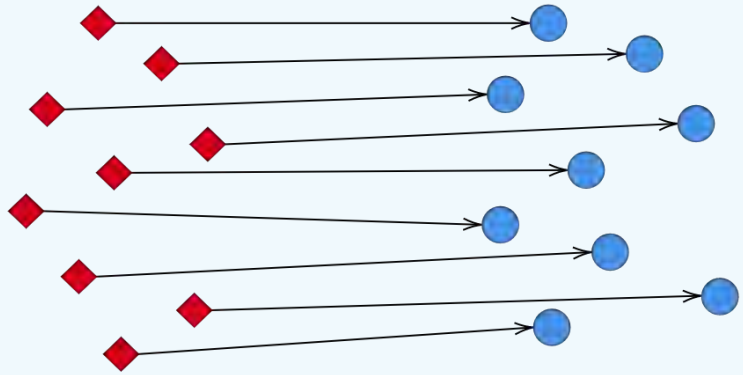
Collective CEs

(Carrizosa et al., 2024a,b)

Collective Counterfactual Explanations

Collective Counterfactual Explanations

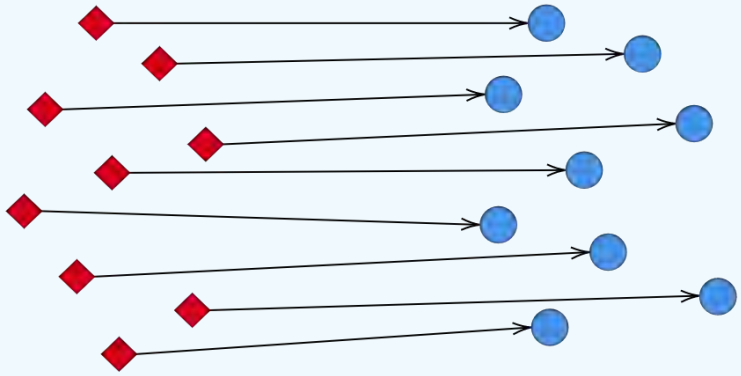
One-for-One Model



(Artelt and Gregoriades, 2024; Carrizosa et al., 2024a,b; Warren et al., 2023)

Collective Counterfactual Explanations

One-for-One Model



(Artelt and Gregoriades, 2024; Carrizosa et al., 2024a,b; Warren et al., 2023)

$$
\begin{array}{lll}
\min_{\underline{x}} & C(\underline{x}_0, \underline{x}) & \\
\text{s.t.} & \beta^\top x_r \geq \tau & r = 1, 2, \ldots, R \\
& \underline{x} \in \underline{\mathcal{X}}(\underline{x}_0) &
\end{array}
$$

- $$\begin{aligned} \boldsymbol{C}(\underline{\boldsymbol{x}}_0, \underline{\boldsymbol{x}}) &= \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \omega_s \pi(\sigma_s(\boldsymbol{x}_r - \boldsymbol{x}_{s0})) \\ &\quad + \lambda_{ind} \sum_{r=1}^{R} \sum_{s \in \mathcal{S}_r} \|\boldsymbol{x}_{s0} - \boldsymbol{x}_r\|_0 + \lambda_{glob} \left\| \left(\max_{r,s \in \mathcal{S}_r} |x_{sj0} - x_{rj}| \right)_{1 \leq j \leq p} \right\|_0 \end{aligned}$$
- $\underline{\boldsymbol{x}}_0 \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}_0) = \prod_s \mathcal{X}(\boldsymbol{x}_{0s})$

-

$$C(\underline{x}_0, \underline{x}) = \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \omega_s \pi(\sigma_s(x_r - x_{s0})) + \lambda_{ind} \sum_{r=1}^{R} \sum_{s \in \mathcal{S}_r} \|x_{s0} - x_r\|_0 + \lambda_{glob} \left\| \left(\max_{r,s \in \mathcal{S}_r} |x_{sj0} - x_{rj}| \right)_{1 \le j \le p} \right\|_0$$

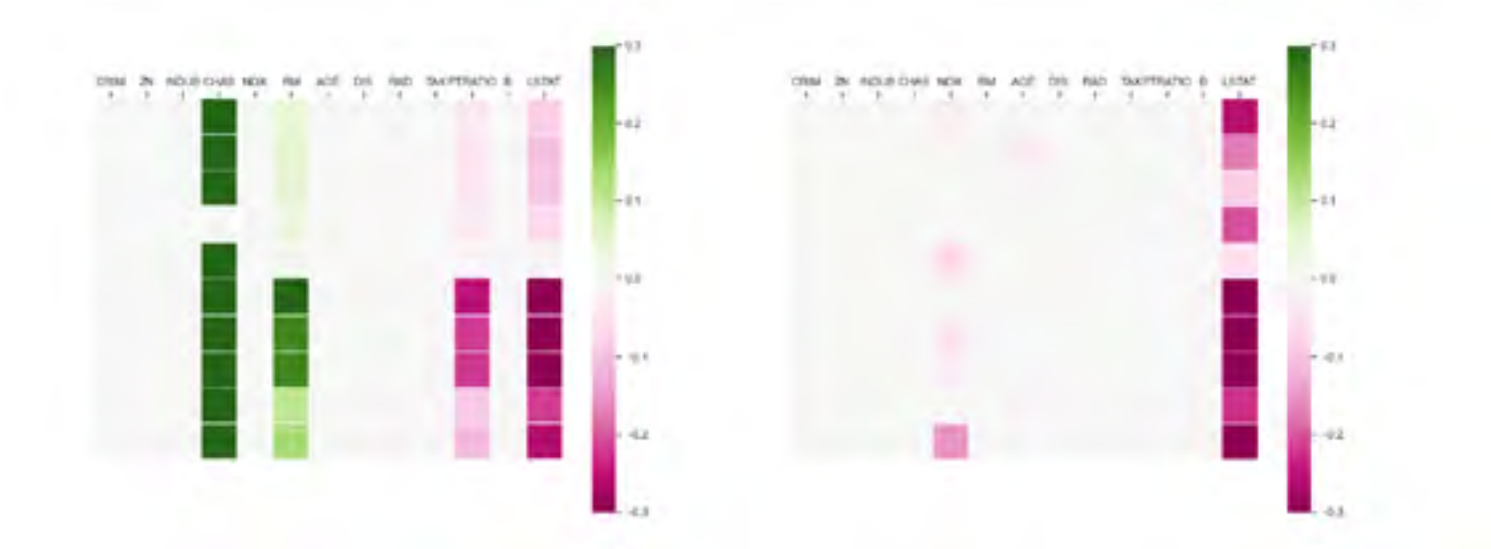
- $\underline{x}_0 \in \underline{\mathcal{X}}(\underline{x}_0) = \prod_s \mathcal{X}(x_{0s})$

$$\begin{array}{ll} \min_{\underline{x}, \xi, \xi^*} & \sum_{s=1}^{S} \omega_s \pi(\sigma_s(x_s - x_{s0})) + \lambda_{ind} \sum_{s=1}^{S} \sum_{j=1}^{p} \xi_{sj} + \lambda_{glob} \sum_{j=1}^{p} \xi_j^* \\ \text{s.t.} & -M_{sj} \xi_{sj} \le x_{sj0} - x_{sj} \le M_{sj} \xi_{sj} \quad j = 1, \ldots, p, s = 1, \ldots, S \\ & \xi_{sj} \in \{0, 1\} \quad s = 1, \ldots, S, j = 1, \ldots p \\ & \xi_j^* \ge \xi_{sj} \quad s = 1, \ldots, S, j = 1, \ldots p \\ & \xi_j^* \in \{0, 1\} \quad j = 1, \ldots, p \\ & \beta^\top x_s \ge \tau \qquad s = 1, \ldots, S \\ & x_s \in \mathcal{X}(x_{s0}) \qquad s = 1, \ldots, S \end{array}$$

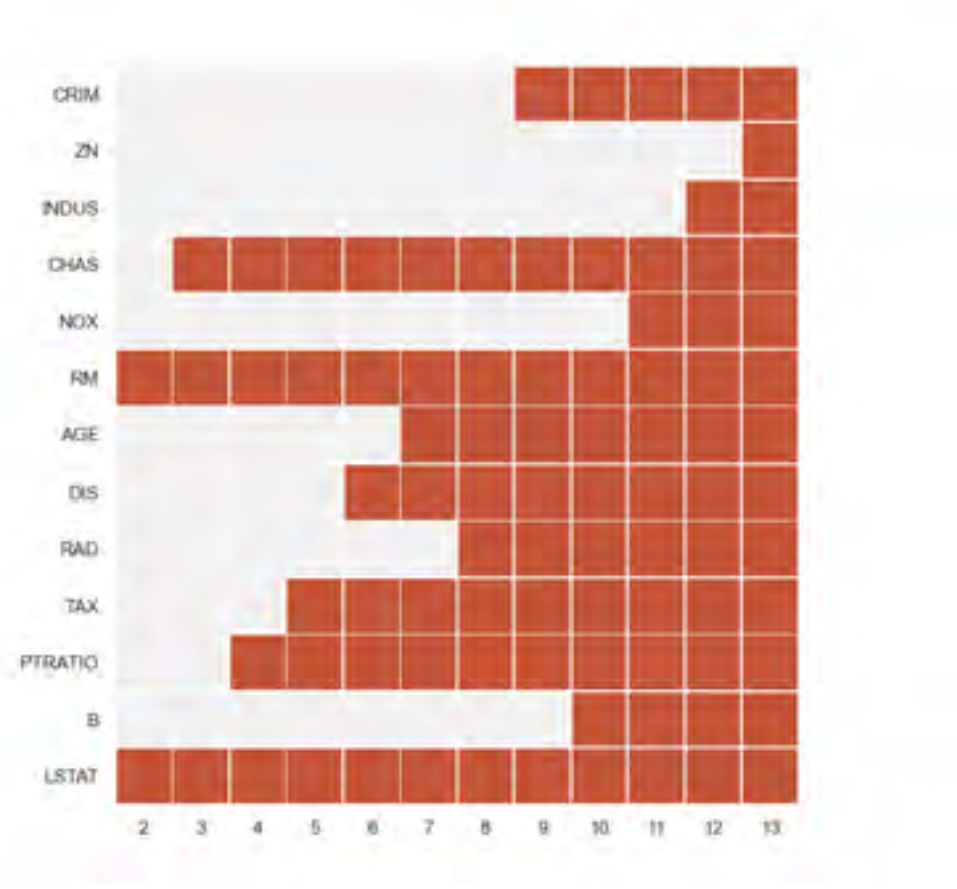
- $$\boldsymbol{C}(\underline{\boldsymbol{x}}_0, \underline{\boldsymbol{x}}) = \sum_{s=1}^{S} \sum_{r \in \mathcal{R}_s} \omega_s \pi(\sigma_s(\boldsymbol{x}_r - \boldsymbol{x}_{s0})) + \lambda_{ind} \sum_{r=1}^{R} \sum_{s \in \mathcal{S}_r} \|\boldsymbol{x}_{s0} - \boldsymbol{x}_r\|_0 + \lambda_{glob} \left\| \left(\max_{r,s \in \mathcal{S}_r} |x_{sj0} - x_{rj}| \right)_{1 \le j \le p} \right\|_0$$
- $\underline{\boldsymbol{x}}_0 \in \underline{\mathcal{X}}(\underline{\boldsymbol{x}}_0) = \prod_s \mathcal{X}(\boldsymbol{x}_{0s})$

$$\begin{array}{ll} \min_{\underline{\boldsymbol{x}}, \boldsymbol{\xi}, \boldsymbol{\xi}^*} & \sum_{s=1}^{S} \omega_s \pi(\sigma_s(\boldsymbol{x}_s - \boldsymbol{x}_{s0})) + \lambda_{ind} \sum_{s=1}^{S} \sum_{j=1}^{p} \xi_{sj} + \lambda_{glob} \sum_{j=1}^{p} \xi_j^* \\ \text{s.t.} & -M_{sj} \xi_{sj} \le x_{sj0} - x_{sj} \le M_{sj} \xi_{sj} \quad j = 1, \ldots, p, s = 1, \ldots, S \\ & \xi_{sj} \in \{0, 1\} \quad s = 1, \ldots, S, j = 1, \ldots p \\ & \xi_j^* \ge \xi_{sj} \quad s = 1, \ldots, S, j = 1, \ldots p \\ & \xi_j^* \in \{0, 1\} \quad j = 1, \ldots, p \\ & \beta^\top \boldsymbol{x}_s \ge \tau \qquad s = 1, \ldots, S \\ & \boldsymbol{x}_s \in \mathcal{X}(\boldsymbol{x}_{s0}) \qquad s = 1, \ldots, S \end{array}$$

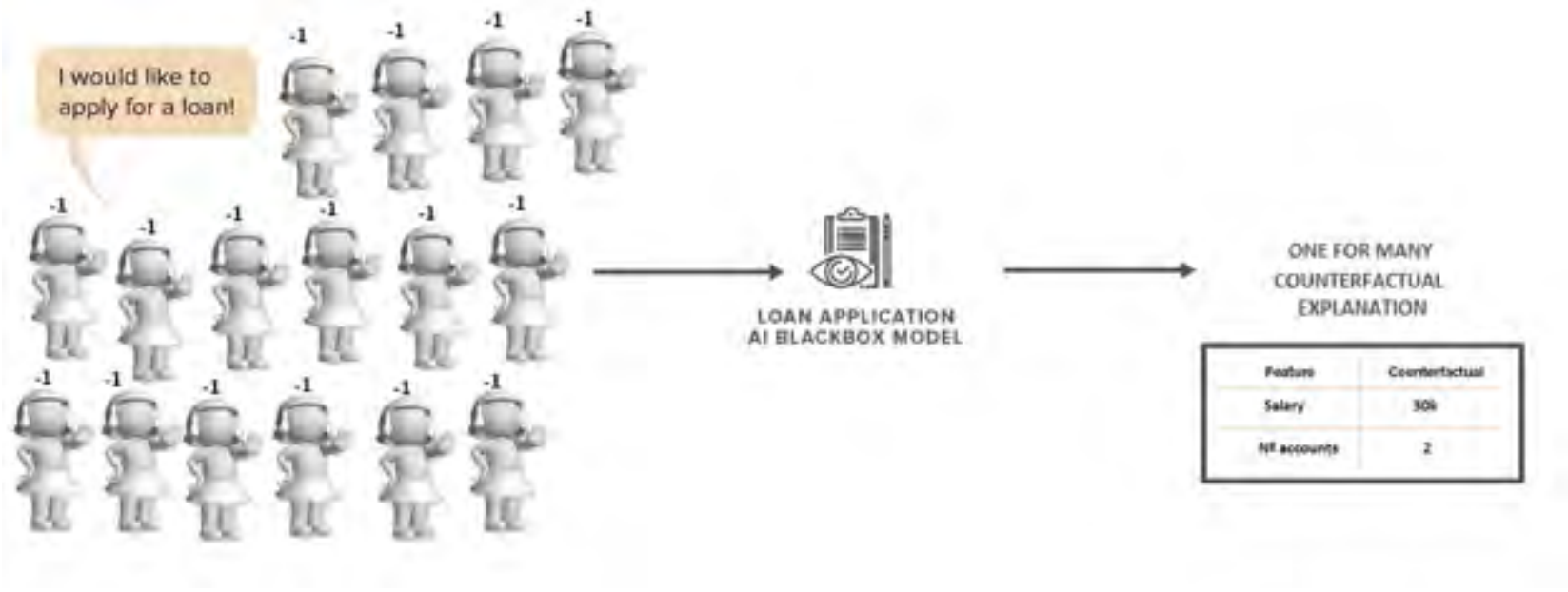
Assuming $\sigma_0 : \ell_2$, $c(t) = t^2$ and $\underline{\mathcal{X}}_0$ to be a bounded polyhedron with some integer coordinates, problem above: Mixed Integer Convex Quadratic Model with linear constraints



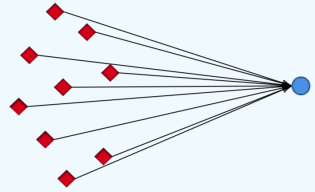
Counterfactual explanations for 10 instances `Boston housing`. Logistic regression and random forest classifiers. Parameters $\lambda_{ind} = 0$, $\lambda_{glob} = 0.2$. The feature perturbations are displayed.

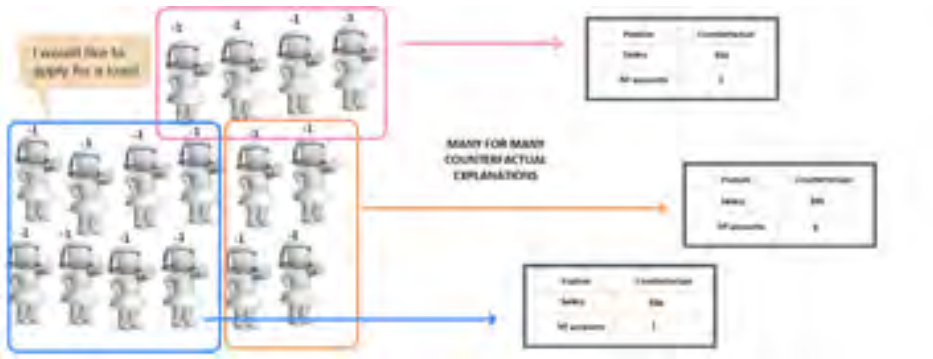


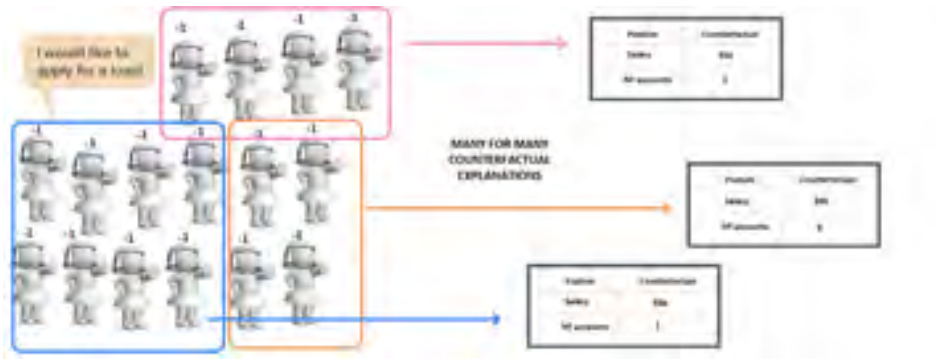
Features that need to be perturbed (in red) for all the instances in the `Boston housing` dataset. The classifier considered is a logistic regression model



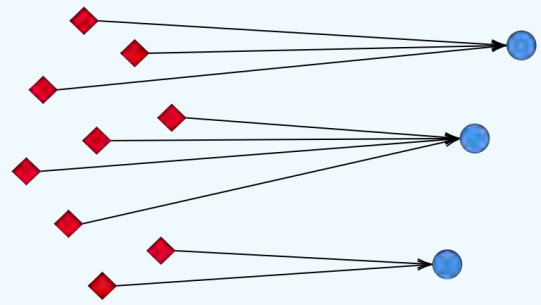
One-for-Many Model

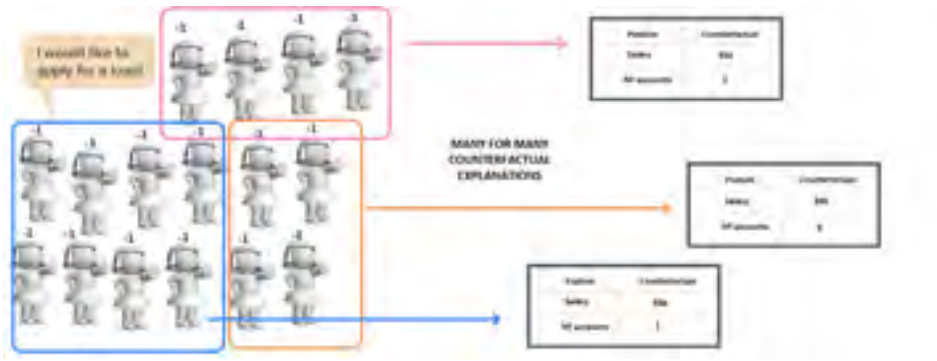



I would like to apply for a loan
MANY FOR MANY COUNTERFACTUAL EXPLANATIONS

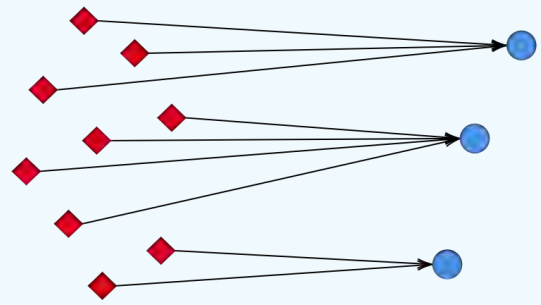


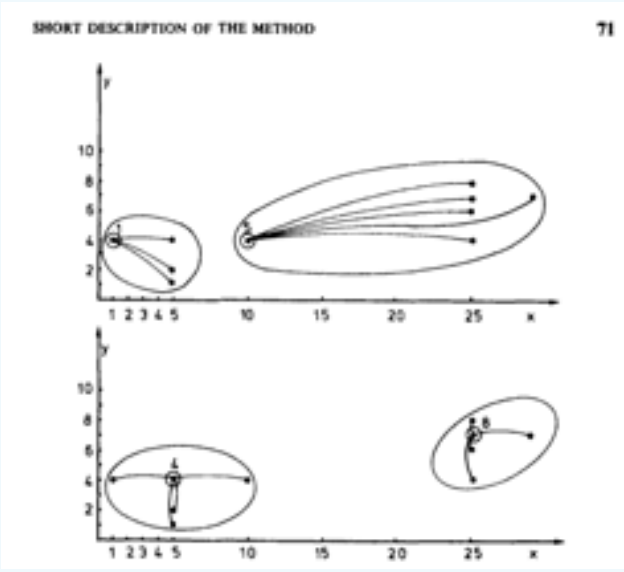
Many-for-Many Model

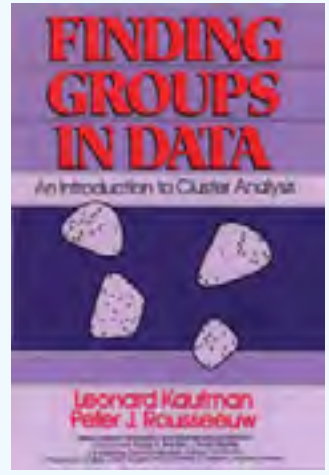




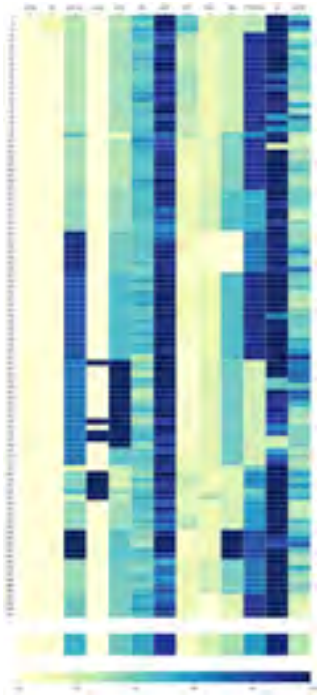
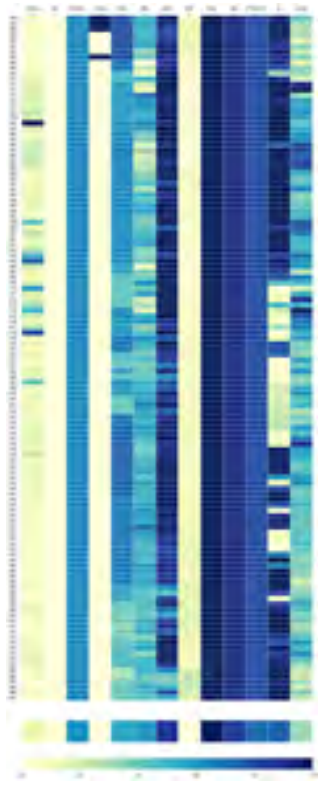
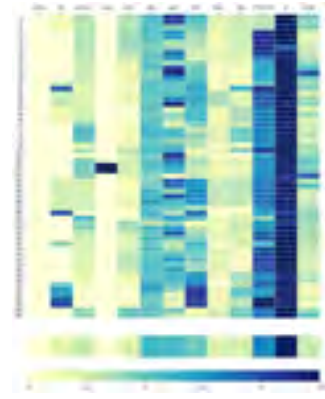
Many-for-Many Model







(Kaufman and Rousseeuw, 1990)

$R = 3$ counterfactual explanations for all the instances in the `Boston housing`. Logistic regression.

We have proposed:

Unified approach to Counterfactual Explanations via Mathematical Optimization

We have proposed:

Unified approach to Counterfactual Explanations via Mathematical Optimization

- Applicable to benchmark classification and regression models such as GLMs, SVM, RF, XGBoost ...

Optimization is crucial to address such problems. The structure of the prediction model/space (not us!!!) will determine the techniques to be used

We have proposed:

Unified approach to Counterfactual Explanations via Mathematical Optimization

- Applicable to benchmark classification and regression models such as GLMs, SVM, RF, XGBoost ...
- Explanations for a single instance or for a collective of instances

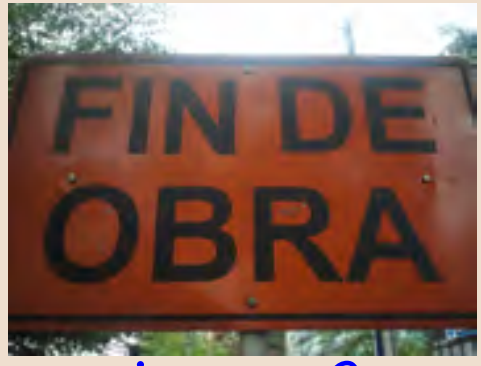
Optimization is crucial to address such problems. The structure of the prediction model/space (not us!!!) will determine the techniques to be used

We have proposed:

Unified approach to Counterfactual Explanations via Mathematical Optimization

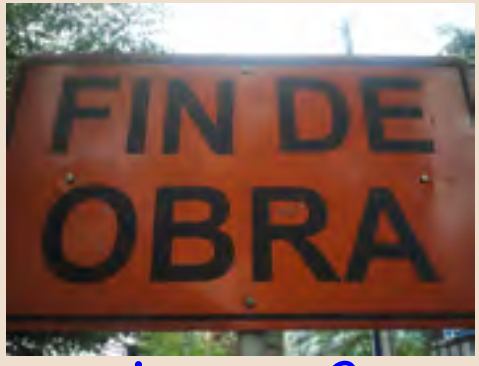
- Applicable to benchmark classification and regression models such as GLMs, SVM, RF, XGBoost ...
- Explanations for a single instance or for a collective of instances
- Different models for costs, searching sparsity
- Different interactions models

Optimization is crucial to address such problems. The structure of the prediction model/space (not us!!!) will determine the techniques to be used



ecarrizosa@us.es

RG



ecarrizosa@us.es

RG

LinkedIn

ecarrizosa@us.es

References I

Artelt, A. and Gregoriades, A. (2024). Supporting organizational decisions on how to improve customer repurchase using multi-instance counterfactual explanations. *Decision Support Systems*, 182:114249.

Artelt, A. and Hammer, B. (2019). On the computation of counterfactual explanations – a survey. *arXiv preprint arXiv:1911.07749*.

Benítez-Peña, S., Blanquero, R., Carrizosa, E., and Ramírez-Cobo, P. (2024). Cost-sensitive probabilistic predictions for support vector machines. *European Journal of Operational Research*, 314(1):268–279.

Blanquero, R., Carrizosa, E., Ramírez-Cobo, P., and Sillero-Denamiel, M. R. (2021). A cost-sensitive constrained lasso. *Advances in Data Analysis and Classification*, 15:121–158.

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

Carrizosa, E., Dražić, M., Dražić, Z., and Mladenović, N. (2012). Gaussian variable neighborhood search for continuous optimization. *Computers & Operations Research*, 39(9):2206–2213.

Carrizosa, E., Molero-Río, C., and Romero Morales, D. (2021). Mathematical optimization in classification and regression trees. *TOP*, 29(1):5–33.

Carrizosa, E. and Navas-Orozco, A. (2025a). Building Robust Counterfactual Explanations with Statistical Guarantees. Technical report, Universidad de Sevilla, `https://www.researchgate.net/publication/396529217_Building_Robust_Counterfactual_Explanations_with_Statistical_Guarantees`.

References II

Carrizosa, E. and Navas-Orozco, A. (2025b). Robust counterfactual explanations in classification and regression. Technical report, Universidad de Sevilla, `https://www.researchgate.net/publication/392923891_Robust_counterfactual_explanations_in_classification_and_regression`.

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

Carrizosa, E., Ramírez-Ayerbe, J., and Romero Morales, D. (2024b). Mathematical optimization modelling for group counterfactual explanations. *European Journal of Operational Research*, 319:399–412.

Carrizosa, E., Ramírez Ayerbe, J., and Romero Morales, D. (2024c). A new model for counterfactual analysis for functional data. *Advances in Data Analysis and Classification*, 18:981–1000.

Carrizosa, E., Ramírez-Ayerbe, J., and Romero Morales, D. (2025). Multistage counterfactual decisions. Technical report, Université de Montréal, In Preparation.

Contardo, C., Fukasawa, R., Rousseau, L.-M., and Vidal, T. (2024). Optimal counterfactual explanations for k-nearest neighbors using mathematical optimization and constraint programming. In *International Symposium on Combinatorial Optimization*, pages 318–331. Springer.

Cui, Z., Chen, W., He, Y., and Chen, Y. (2015). 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.

References III

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

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

Drezner, T. and Drezner, Z. (2021). Asymmetric distance location model. *INFOR: Information Systems and Operational Research*, 59(1):102–110.

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

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*.

Gambella, C., Ghaddar, B., and Naoum-Sawaya, J. (2021). Optimization models for machine learning: A survey. *European Journal of Operational Research*, 290(3):807–828.

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

References IV

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. (2019). Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

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

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

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

Karimi, A.-H., Barthe, G., Schölkopf, B., and Valera, I. (2021). A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*.

Kaufman, L. and Rousseeuw, P. (1990). *Finding Groups in Data. An Introduction to Cluster Analysis*. John Wiley & Sons, New York.

Kurtz, J., Birbil, S. I., and den Hertog, D. (2024). Counterfactual explanations for linear optimization.

References V

Lucic, A., Oosterhuis, H., Haned, H., and de Rijke, M. (2019). FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. *arXiv preprint arXiv:1911.12199*.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Magagnini, M., Carrizosa, E., and de Leone, R. (2025a). Counterfactual explanations with the k-Nearest Neighborhood classifier and uncertain data. Technical report, Universitá di Camerino.

Magagnini, M., Carrizosa, E., and De Leone, R. (2025b). Nearest neighbors counterfactuals. In Nicosia, G., Ojha, V., Giesselbach, S., Pardalos, M. P., and Umeton, R., editors, *Machine Learning, Optimization, and Data Science*, pages 193–208.

Maragno, D., Kurtz, J., Röber, T., Goedhart, R., Birbil, S., and den Hertog, D. (2024). Finding regions of counterfactual explanations via robust optimization. *INFORMS Journal on Computing*.

Maragno, D., Röber, T. E., and Birbil, I. (2022). Counterfactual explanations using optimization with constraint learning. *arXiv preprint arXiv:2209.10997*.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

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

References VI

Panigutti, C., Hamon, R., Hupont, I., Fernandez Llorca, D., Fano Yela, D., Junklewitz, H., Scalzo, S., Mazzini, G., Sanchez, I., Soler Garrido, J., and Gomez, E. (2023). 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. Association for Computing Machinery.

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

Plastria, F. (1992). On destination optimality in asymmetric distance Fermat-Weber problems. *Annals of Operations Research*, 40:355–369.

Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in Large Margin Classifiers*, 10(3):61–74.

Poyiadzi, R., Sokol, K., Santos-Rodriguez, R., De Bie, T., and Flach, P. (2020). FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350.

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

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

References VII

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

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

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

Smyth, B. and Keane, M. T. (2022). A few good counterfactuals: generating interpretable, plausible and diverse counterfactual explanations. In *International Conference on Case-Based Reasoning*, pages 18–32.

Stepin, I., Alonso, J., Catalá, A., and Pereira-Fariña, M. (2021). A survey of contrastive and counterfactual explanation generation methods for explainable artificial intelligence. *IEEE Access*, 9:11974–12001.

Verma, S., Dickerson, J., and Hines, K. (2021). Counterfactual explanations for machine learning: Challenges revisited. *arXiv preprint arXiv:2106.07756*.

Virgolin, M. and Fracaros, S. (2023). On the robustness of sparse counterfactual explanations to adverse perturbations. *Artificial Intelligence*, 316:103840.

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

References VIII

Warren, G., Keane, M., Gueret, C., and E., D. (2023). Explaining groups of instances counterfactually for xai: A use case, algorithm and user study for group-counterfactuals. *arXiv preprint arXiv:2303.09297*.

Zafar, M., Valera, I., Gomez Rodriguez, M., and Gummadi, K. (2017). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pages 962–970. PMLR.