Counterfactual Explanations via Mathematical Optimization

Emilio Carrizosa[1], **Jasone Ramírez-Ayerbe**[1] and Dolores Romero Morales[2]

[1] Instituto de Matemáticas de la Universidad de Sevilla, Seville, Spain
[2] Copenhagen Business School, Frederiksberg, Denmark

10th November 2021

New Bridges between Mathematics and Data Science, Valladolid, Nov 8 - 11

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

Outline

Interpretability

- The EU imposes the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- Accuracy matters but also Interpretability and Transparency
  - Global and Local Explainability: LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]
  - Fairness [Zafar et al., 2017]
  - **Counterfactual Explanations** [Wachter et al., 2017]

Interpretability in Data Science

Interpretability

- The EU imposes the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- Accuracy matters but also Interpretability and Transparency
  - Global and Local Explainability: LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]
  - Fairness [Zafar et al., 2017]
  - **Counterfactual Explanations** [Wachter et al., 2017]

Interpretability in Data Science

Interpretability

- The EU imposes the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- Accuracy matters but also Interpretability and Transparency
  - Global and Local Explainability: LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]
  - Fairness [Zafar et al., 2017]
  - **Counterfactual Explanations** [Wachter et al., 2017]

Interpretability in Data Science

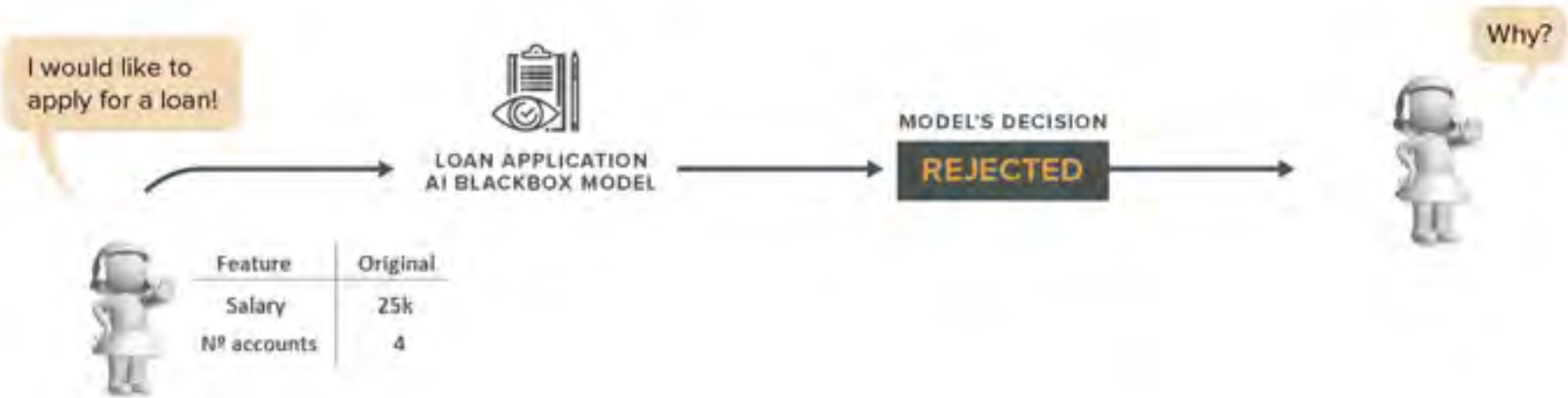
Interpretability

- The EU imposes the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- Accuracy matters but also Interpretability and Transparency
  - Global and Local Explainability: LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]
  - Fairness [Zafar et al., 2017]
  - **Counterfactual Explanations** [Wachter et al., 2017]

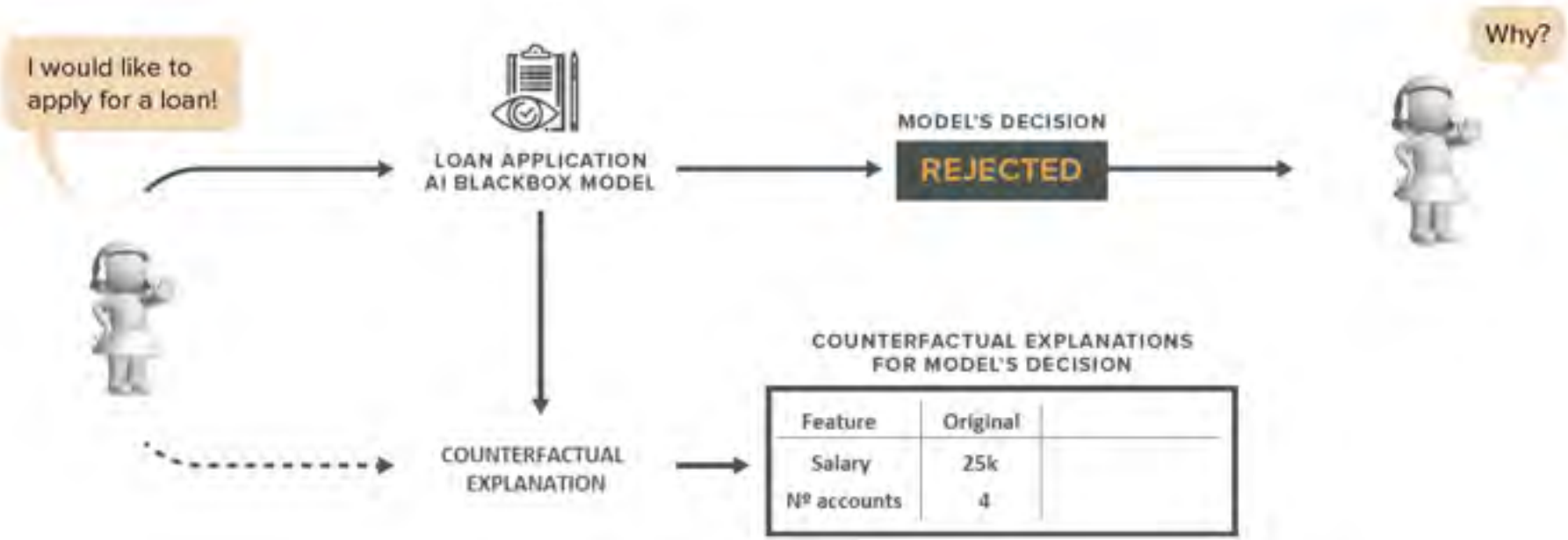
Interpretability

- The EU imposes the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- Accuracy matters but also Interpretability and Transparency
  - Global and Local Explainability: LIME [Ribeiro et al., 2016], SHAP [Lundberg and Lee, 2017]
  - Fairness [Zafar et al., 2017]
  - **Counterfactual Explanations** [Wachter et al., 2017]

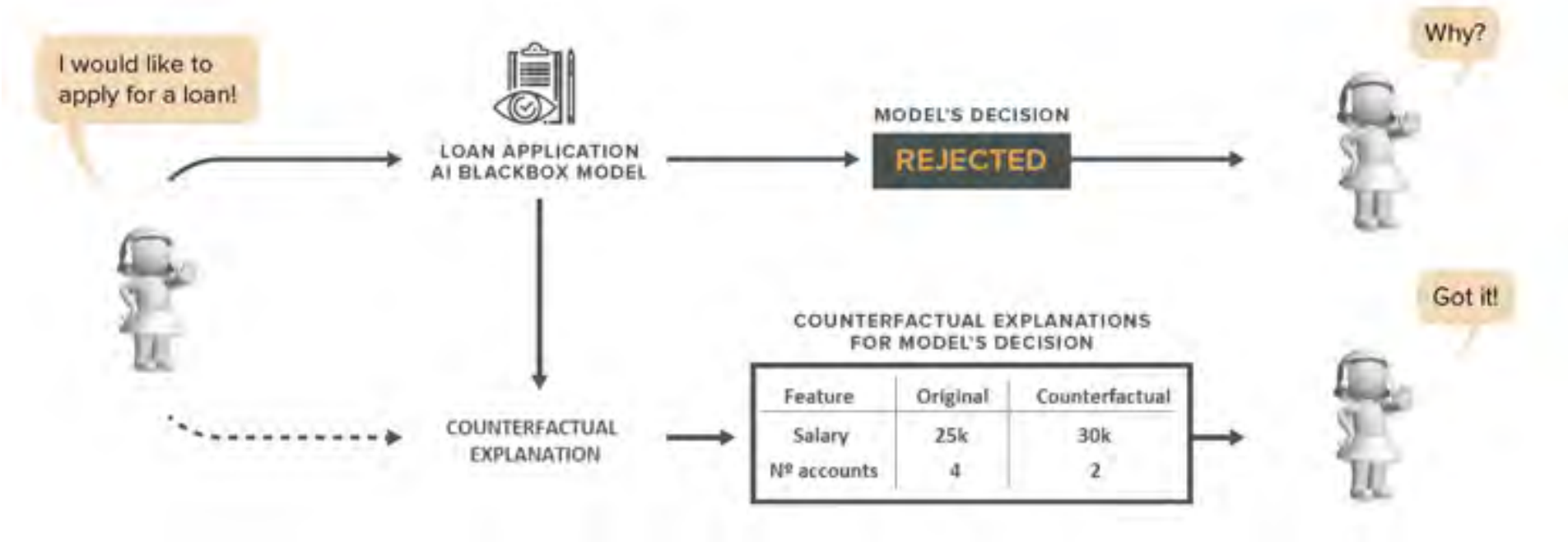
Interpretability in Data Science: Counterfactuals



Interpretability in Data Science: Counterfactuals



Interpretability in Data Science: Counterfactuals



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.

Interpretability in Data Science: Counterfactuals

Suppose we have a multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

State-of-the-Art

- Only valid for specific methodologies:
  - differentiable [Wachter et al., 2017, Mothilal et al., 2020, Looveren et al., 2020],
  - tree-based [Lucic et al., 2019, Lindgren et al., 2016, Parmentier et al., 2021],
- Metaheuristics, e.g., genetic algorithm [Dandl et al., 2020, Guidotti et al., 2019], graph-based [Poyiadzi et al., 2020]

Interpretability in Data Science: Counterfactuals

Suppose we have a multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

State-of-the-Art

- Only valid for specific methodologies:
  - differentiable [Wachter et al., 2017, Mothilal et al., 2020, Lucic et al., 2019]
  - tree-based [Cui et al., 2015, Lindgren et al., 2019, Parmentier and Vidal, 2021]
- Metaheuristics, e.g., genetic algorithm [Dandl et al., 2020, Guidotti et al., 2019], graph-based [Poyiadzi et al., 2020]

Interpretability in Data Science: Counterfactuals

Suppose we have a multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

State-of-the-Art

- Only valid for specific methodologies:
  - differentiable [Wachter et al., 2017, Mothilal et al., 2020, Lucic et al., 2019]
  - tree-based [Cui et al., 2015, Lindgren et al., 2019, Parmentier and Vidal, 2021]
- Metaheuristics, e.g., genetic algorithm [Dandl et al., 2020, Guidotti et al., 2019], graph-based [Poyiadzi et al., 2020]

Interpretability in Data Science: Counterfactuals

Suppose we have a multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

State-of-the-Art

- Only valid for specific methodologies:
  - differentiable [Wachter et al., 2017, Mothilal et al., 2020, Lucic et al., 2019]
  - tree-based [Cui et al., 2015, Lindgren et al., 2019, Parmentier and Vidal, 2021]
- Metaheuristics, e.g., genetic algorithm [Dandl et al., 2020, Guidotti et al., 2019], graph-based [Poyiadzi et al., 2020]

Interpretability in Data Science: Counterfactuals

Suppose we have multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

Our proposal [Carrizosa et al., 2021]:

Unified approach to Counterfactual Explanations for score-based classifiers:

- By means of Mathematical Optimization
- Applicable to multi-class classification models like: LR, SVM, RF, XGBoost...
- Individual and Group level Explanations
- Tabular and functional data

Interpretability in Data Science: Counterfactuals

Suppose we have multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

Our proposal [Carrizosa et al., 2021]:

Unified approach to Counterfactual Explanations for score-based classifiers:

- By means of Mathematical Optimization
- Applicable to multi-class classification models like: LR, SVM, RF, XGBoost...
- Individual and Group level Explanations
- Tabular and functional data

Interpretability in Data Science: Counterfactuals

Suppose we have multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

Our proposal [Carrizosa et al., 2021]:

Unified approach to Counterfactual Explanations for score-based classifiers:

- By means of Mathematical Optimization
- Applicable to multi-class classification models like: LR, SVM, RF, XGBoost...
- Individual and Group level Explanations
- Tabular and functional data

Interpretability in Data Science: Counterfactuals

Suppose we have multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

Our proposal [Carrizosa et al., 2021]:

Unified approach to Counterfactual Explanations for score-based classifiers:

- By means of Mathematical Optimization
- Applicable to multi-class classification models like: LR, SVM, RF, XGBoost...
- Individual and Group level Explanations
- Tabular and functional data

Interpretability in Data Science: Counterfactuals

Suppose we have multi-class classification model with K classes.

Counterfactual Explanations

Given an individual $\boldsymbol{x}$ who has been classified with a **given model** in class k, what are the changes with minimum cost that cause $\boldsymbol{x}$ to be classified in $k^* \neq k$?

Our proposal [Carrizosa et al., 2021]:

Unified approach to Counterfactual Explanations for score-based classifiers:

- By means of Mathematical Optimization
- Applicable to multi-class classification models like: LR, SVM, RF, XGBoost...
- Individual and Group level Explanations
- Tabular and functional data

Outline

Counterfactual Explanations

Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Counterfactual Explanations

Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Counterfactual Explanations

Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Counterfactual Explanations

Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Counterfactual Explanations

Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Counterfactual Explanations

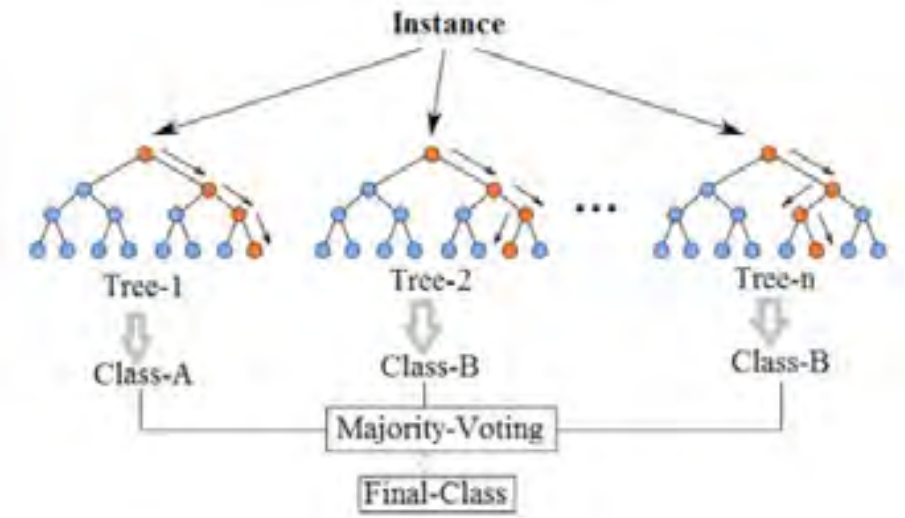
Given:

- a score based classifier with score function $f_k : \mathbb{R}^J \to \mathbb{R}$ for class $k = 1, \ldots, K$.
- an instance with predictor vector $\boldsymbol{x}^0$
- $\boldsymbol{x}^0$ classified in $k^0 \in \arg\max_k f_k(\boldsymbol{x}^0)$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$

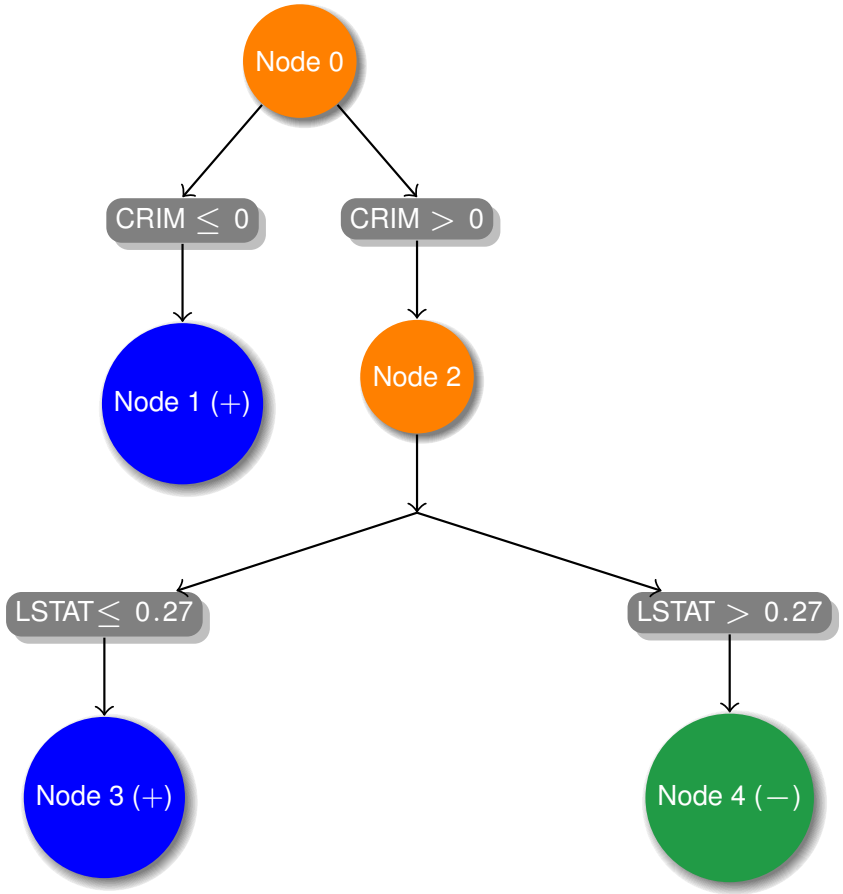
Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Additive Tree Models

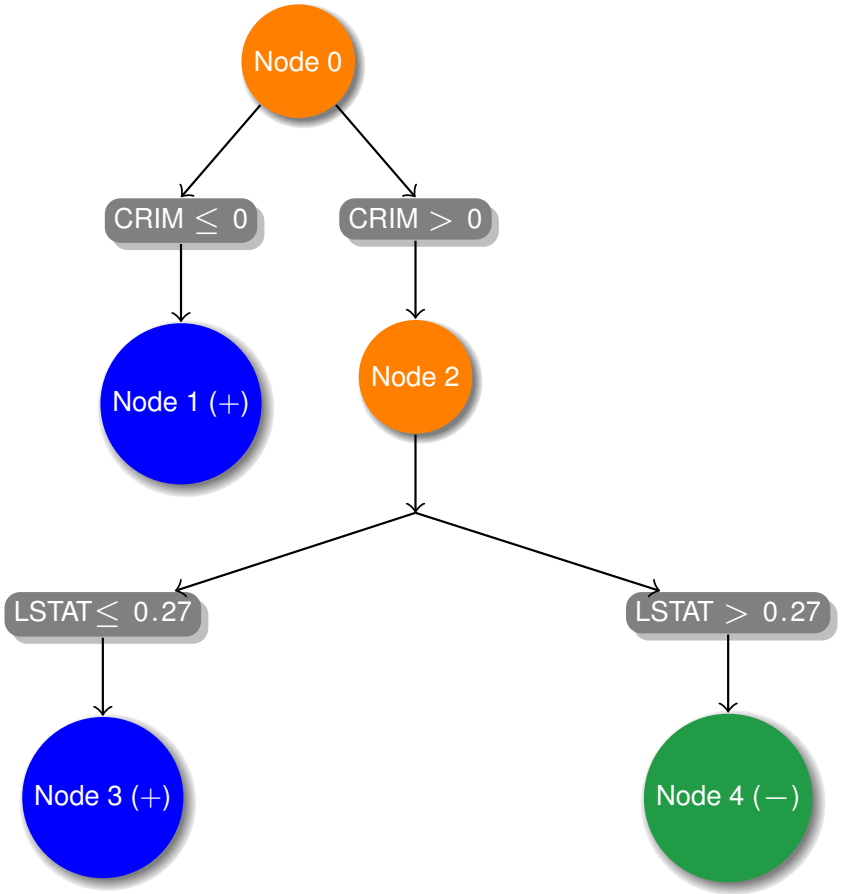


'+' (high price) vs '−' (low price)



Additive Tree Models

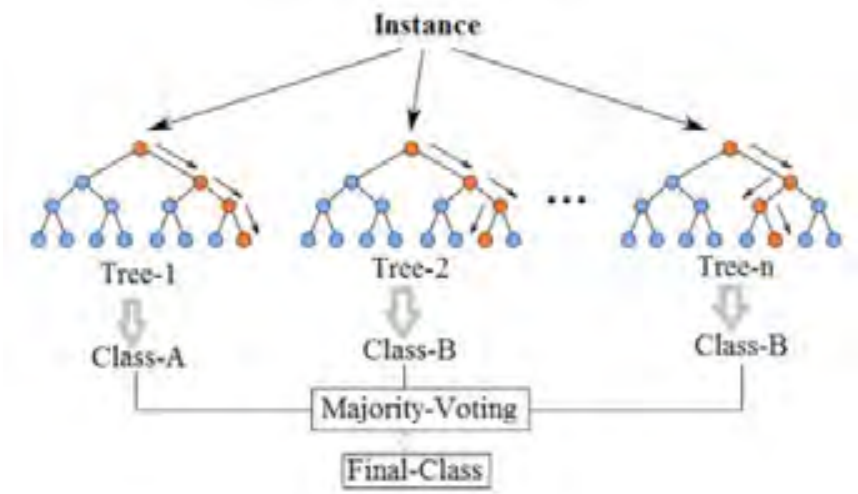
'+' (high price) vs '−' (low price)



Data from tree t, $t = 1, \ldots, T$, in the ATM

- weight $w^t \geq 0$
- set of leaves $\mathcal{L}^t$
- sets of splits Left(t, l) and Right(t, l) for $l \in \mathcal{L}^t$
- threshold value c_s and feature used $v(s)$ in each split node s, $s \in \text{Left}(l, t) \cup \text{Right}(l, t)$
- $\mathcal{L}_k^t$ subset of leaves in t whose output is class $k = 1, \ldots, K$

Additive Tree Models



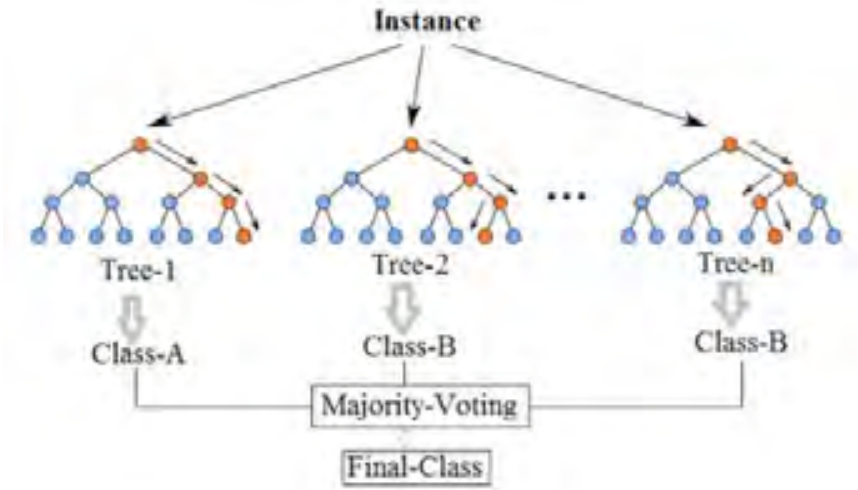
Decision variables

- $\boldsymbol{x} \in \mathbb{R}^J$ counterfactual
- $z_l^t \in \{0, 1\}$ indicates whether the counterfactual instance $\boldsymbol{x}$ ends in leaf $l \in \mathcal{L}_t$ or not, $t = 1, \ldots, T$

Score function for class k

$$\sum_{t=1}^{T} w^t \cdot \left\{ \begin{array}{cc} 1 & \text{if } \boldsymbol{x} \text{ predicted in class } k \text{ in tree } t \\ 0 & \text{otherwise} \end{array} \right\}$$

Additive Tree Models



Decision variables

- $\boldsymbol{x} \in \mathbb{R}^J$ counterfactual
- $z_l^t \in \{0, 1\}$ indicates whether the counterfactual instance $\boldsymbol{x}$ ends in leaf $l \in \mathcal{L}_t$ or not, $t = 1, \ldots, T$

Score function for class k

$$\sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_k^t} z_l^t$$

Counterfactual Explanations for ATM models

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{z}} \quad & C(\boldsymbol{x}, \boldsymbol{x^0}) \\
\text{s.t} \quad & x_{v(s)} - M_1(1 - z_l^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T \\
& x_{v(s)} + M_2(1 - z_l^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T \\
& \sum_{l \in \mathcal{L}^t} z_l^t = 1 \qquad \forall t = 1, \ldots, T \\
& \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_{k^*}} z_l^t \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_k} z_l^t \qquad \forall k = 1, \ldots, K \quad k \neq k^* \\
& z_l^t \in \{0,1\} \qquad \forall l \in \mathcal{L}^t \qquad \forall t = 1, \ldots, T \\
& \boldsymbol{x} \in \mathcal{X}^0
\end{aligned}
$$

The value of M_1 and M_2 can be tightened for each split.

Counterfactual Explanations for ATM models

$$
\begin{aligned}
\min_{\boldsymbol{x},\boldsymbol{z}} \quad & C(\boldsymbol{x}, \boldsymbol{x^0}) \\
\text{s.t} \quad & x_{v(s)} - M_1(1 - z_l^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T \\
& x_{v(s)} + M_2(1 - z_l^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T \\
& \sum_{l \in \mathcal{L}^t} z_l^t = 1 \qquad \forall t = 1, \ldots, T \\
& \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_{k^*}} z_l^t \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_k} z_l^t \qquad \forall k = 1, \ldots, K \quad k \neq k^* \\
& z_l^t \in \{0,1\} \qquad \forall l \in \mathcal{L}^t \qquad \forall t = 1, \ldots, T \\
& \boldsymbol{x} \in \mathcal{X}^0
\end{aligned}
$$

The value of M_1 and M_2 can be tightened for each split.

A particular case

$$C(\boldsymbol{x}, \boldsymbol{x}^0) = \lambda_0 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0 + \lambda_2 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2$$

- ℓ_0: Number of features changed.
- ℓ_2: Sum of the squared deviations

Counterfactual Explanations for ATM models

$$\min_{\boldsymbol{x}, \boldsymbol{z}} \quad \lambda_0 \sum_{j=1}^{J} \xi_j + \lambda_2 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2$$

$$\text{s.t} \quad x_{v(s)} - M_1(1 - z_l^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l, t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$x_{v(s)} + M_2(1 - z_l^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l, t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$\sum_{l \in \mathcal{L}^t} z_l^t = 1 \qquad \forall t = 1, \ldots, T$$

$$\sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_{k^*}^t} z_l^t \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_k^t} z_l^t \qquad \forall k = 1, \ldots, K \quad k \neq k^*$$

$$-M_3 \xi_j \leq x_j^0 - x_j \leq M_3 \xi_j \quad j = 1, \ldots, J$$

$$\xi_j \in \{0, 1\} \quad j = 1, \ldots, J$$

$$z_l^t \in \{0, 1\} \qquad \forall l \in \mathcal{L}^t \qquad \forall t = 1, \ldots, T$$

$$\boldsymbol{x} \in \mathcal{X}^0$$

- $\lambda_2 = 0$: MILP model
- otherwise: Mixed Integer Convex Quadratic Model with linear constraints
- initial solution easy to calculate

A particular case

$$C(\boldsymbol{x}, \boldsymbol{x}^0) = \lambda_0 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_0 + \lambda_2 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2$$

- ℓ_0: Number of features changed.
- ℓ_2: Sum of the squared deviations

Counterfactual Explanations for ATM models

$$\min_{\boldsymbol{x},\boldsymbol{z}} \quad \lambda_0 \sum_{j=1}^{J} \xi_j + \lambda_2 \|\boldsymbol{x}^0 - \boldsymbol{x}\|_2^2$$

$$\text{s.t} \quad x_{v(s)} - M_1(1 - z_l^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l, t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$x_{v(s)} + M_2(1 - z_l^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l, t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$\sum_{l \in \mathcal{L}^t} z_l^t = 1 \qquad \forall t = 1, \ldots, T$$

$$\sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_{k^*}^t} z_l^t \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_k^t} z_l^t \qquad \forall k = 1, \ldots, K \quad k \neq k^*$$

$$-M_3 \xi_j \leq x_j^0 - x_j \leq M_3 \xi_j \quad j = 1, \ldots, J$$

$$\xi_j \in \{0, 1\} \quad j = 1, \ldots, J$$

$$z_l^t \in \{0, 1\} \qquad \forall l \in \mathcal{L}^t \qquad \forall t = 1, \ldots, T$$

$$\boldsymbol{x} \in \mathcal{X}^0$$

- $\lambda_2 = 0$: MILP model
- otherwise: Mixed Integer Convex Quadratic Model with linear constraints
- initial solution easy to calculate

Computational Experiments

- Boston Housing Dataset [Harrison Jr and Rubinfeld, 1978]

Variable	Definition	Type
CRIM	per capita crime rate by town	numerical
ZN	proportion of residential land zoned for lots over 25,000 sq.ft	numerical
INDUS	proportion of non-retail business acres per town	numerical
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	binary
NOX	nitric oxides concentration (parts per 10 million)	numerical
RM	average number of rooms per dwelling	numerical
AGE	proportion of owner-occupied units built prior to 1940	numerical
DIS	weighted distances to five Boston employment centres	numerical
RAD	index of accessibility to radial highways	numerical
TAX	full-value property-tax rate per $10,000	numerical
PTRATIO	pupil-teacher ratio by town	numerical
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town	numerical
LSTAT	% lower status of the population	numerical
MEDV	$+1$ if Median value of owner-occupied homes in $1000's over the 50th percentile, -1 otherwise	binary target

Table: Description of the features of the Boston housing dataset

- Model: RF with $T = 500$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]
- Time limit of 250 sec

Computational Experiments

- Boston Housing Dataset [Harrison Jr and Rubinfeld, 1978]

Variable	Definition	Type
CRIM	per capita crime rate by town	numerical
ZN	proportion of residential land zoned for lots over 25,000 sq.ft	numerical
INDUS	proportion of non-retail business acres per town	numerical
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	binary
NOX	nitric oxides concentration (parts per 10 million)	numerical
RM	average number of rooms per dwelling	numerical
AGE	proportion of owner-occupied units built prior to 1940	numerical
DIS	weighted distances to five Boston employment centres	numerical
RAD	index of accessibility to radial highways	numerical
TAX	full-value property-tax rate per $10,000	numerical
PTRATIO	pupil-teacher ratio by town	numerical
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town	numerical
LSTAT	% lower status of the population	numerical
MEDV	$+1$ if Median value of owner-occupied homes in $1000's over the 50th percentile, -1 otherwise	binary target

Table: Description of the features of the Boston housing dataset

- Model: RF with $T = 500$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]
- Time limit of 250 sec

Computational Experiments

- Boston Housing Dataset [Harrison Jr and Rubinfeld, 1978]

Variable	Definition	Type
CRIM	per capita crime rate by town	numerical
ZN	proportion of residential land zoned for lots over 25,000 sq.ft	numerical
INDUS	proportion of non-retail business acres per town	numerical
CHAS	Charles River dummy variable (1 if tract bounds river; 0 otherwise)	binary
NOX	nitric oxides concentration (parts per 10 million)	numerical
RM	average number of rooms per dwelling	numerical
AGE	proportion of owner-occupied units built prior to 1940	numerical
DIS	weighted distances to five Boston employment centres	numerical
RAD	index of accessibility to radial highways	numerical
TAX	full-value property-tax rate per $10,000	numerical
PTRATIO	pupil-teacher ratio by town	numerical
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town	numerical
LSTAT	% lower status of the population	numerical
MEDV	$+1$ if Median value of owner-occupied homes in $1000's over the 50th percentile, -1 otherwise	binary target

Table: Description of the features of the Boston housing dataset

- Model: RF with $T = 500$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]
- Time limit of 250 sec

Counterfactual explanation for 1 instance

Figure: Counterfactual instance from $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$

Computational experiments

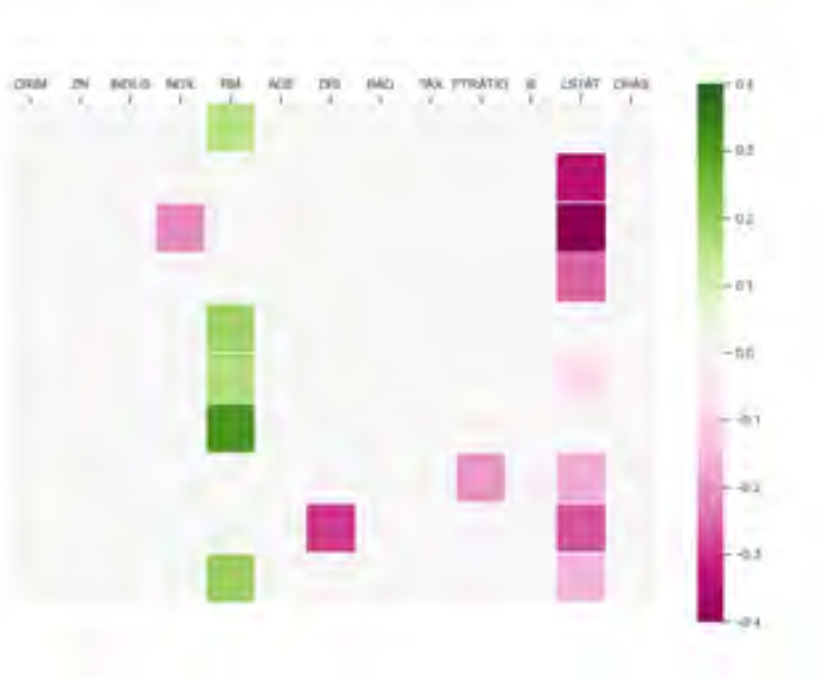


Figure: From class $k = -1$ to $k^* = +1$ with $C = \ell_0$

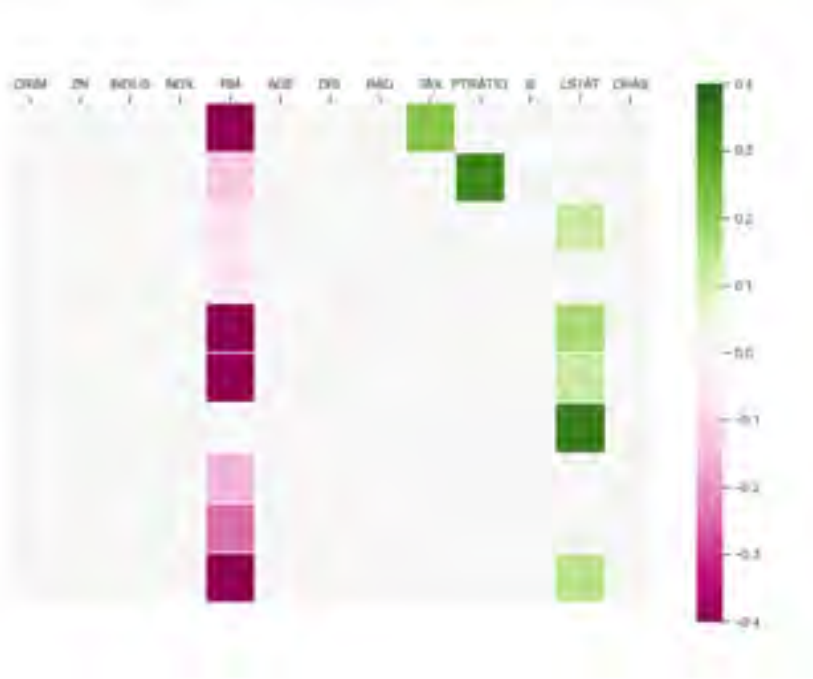


Figure: From class $k = +1$ to $k^* = -1$ with $C = \ell_0$

Computational experiments

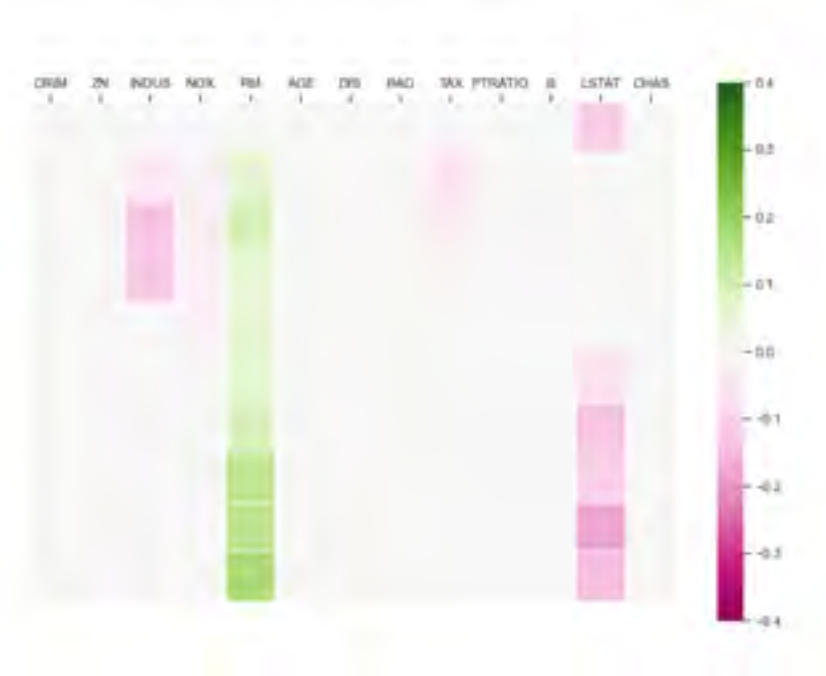


Figure: From class $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$

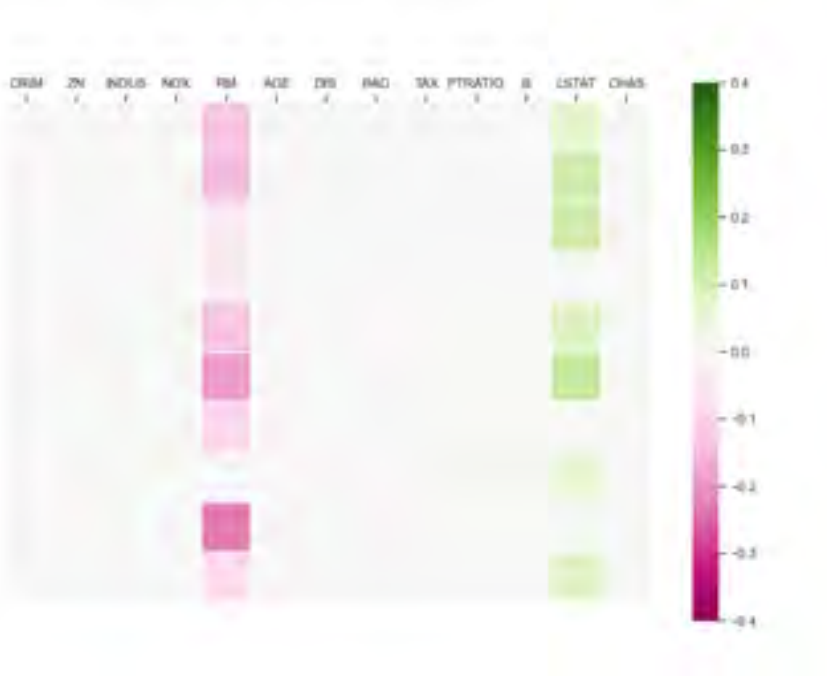


Figure: From class $k = +1$ to $k^* = -1$ with $C = 0.01\ell_0 + \ell_2^2$

Computational experiments

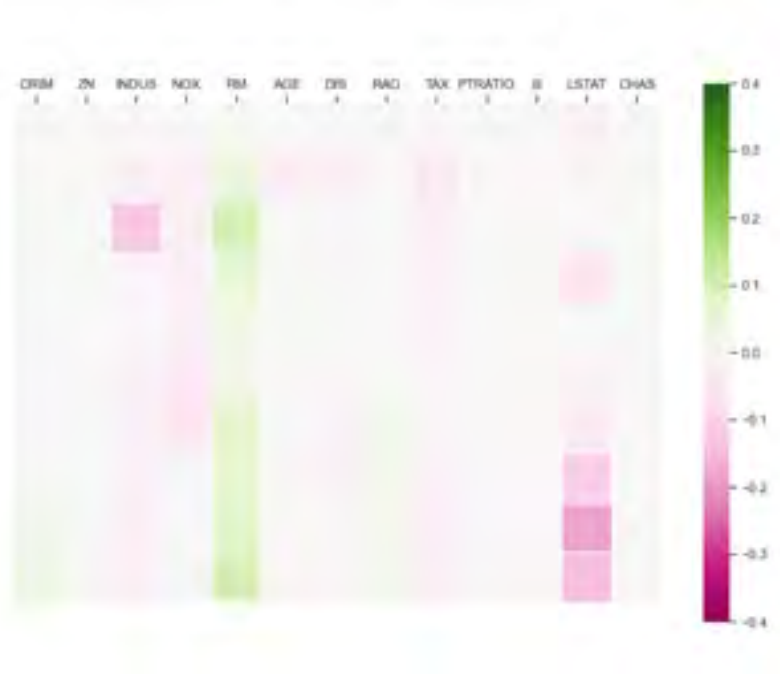


Figure: From class $k = -1$ to $k^* = +1$ with $C = \ell_2^2$

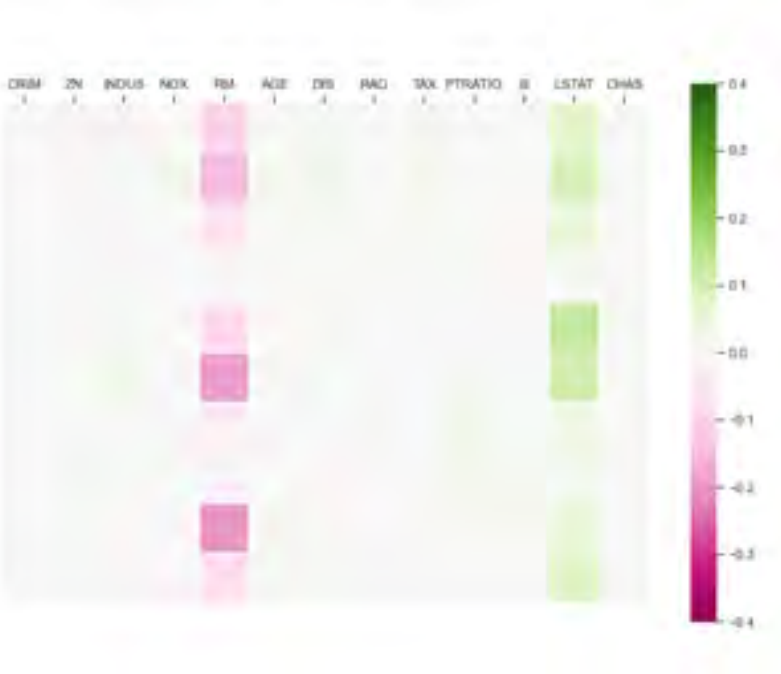


Figure: From class $k = +1$ to $k^* = -1$ with $C = \ell_2^2$

Counterfactual Explanations for linear models

Given a multi-classification linear model:

- Logistic Regression (LR)
- Support Vector Machine (SVM)

Score function:

$$w^k \boldsymbol{x} + b^k$$

Counterfactual Explanations for linear models

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x^0}) \\ \text{s.t.} \quad & w^{k^*} \boldsymbol{x} + b^{k^*} \geq w^k \boldsymbol{x} + b^k \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

$C(\boldsymbol{x}, \boldsymbol{x}^0) = \lambda_0 \|\boldsymbol{x^0} - \boldsymbol{x}\|_0 + \lambda_2 \|\boldsymbol{x^0} - \boldsymbol{x}\|_2^2$

- $\lambda_2 = 0$: MILP model
- otherwise: Mixed Integer Convex Quadratic Model with linear constraints

Computational Experiments

Counterfactuals for LR

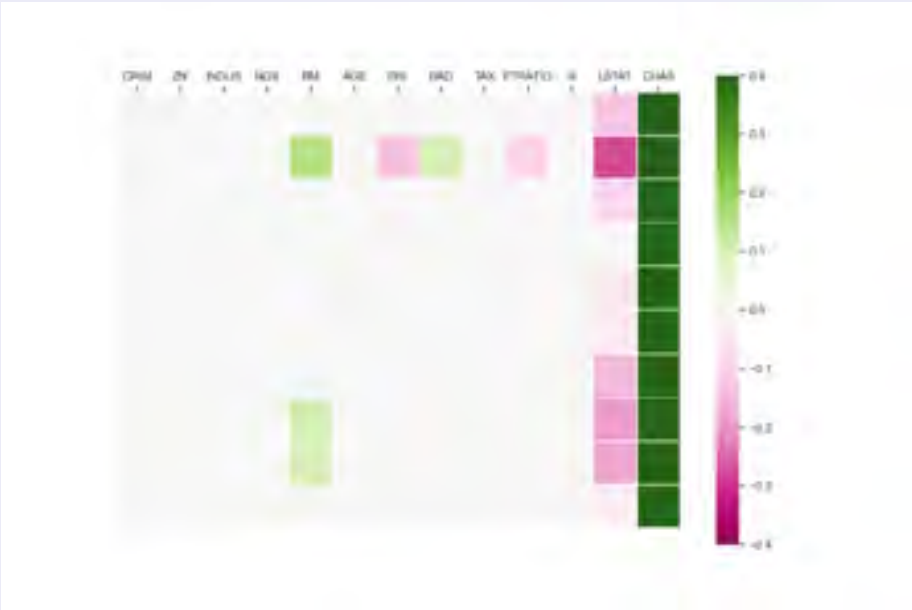


Figure: From $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$

Counterfactuals for SVM

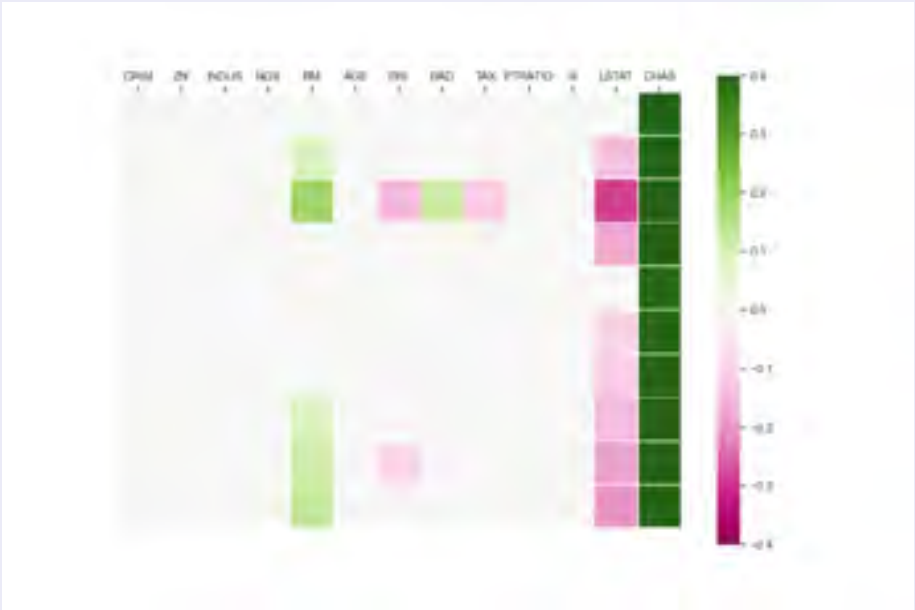


Figure: From $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$

Computational Experiments

Counterfactuals for LR

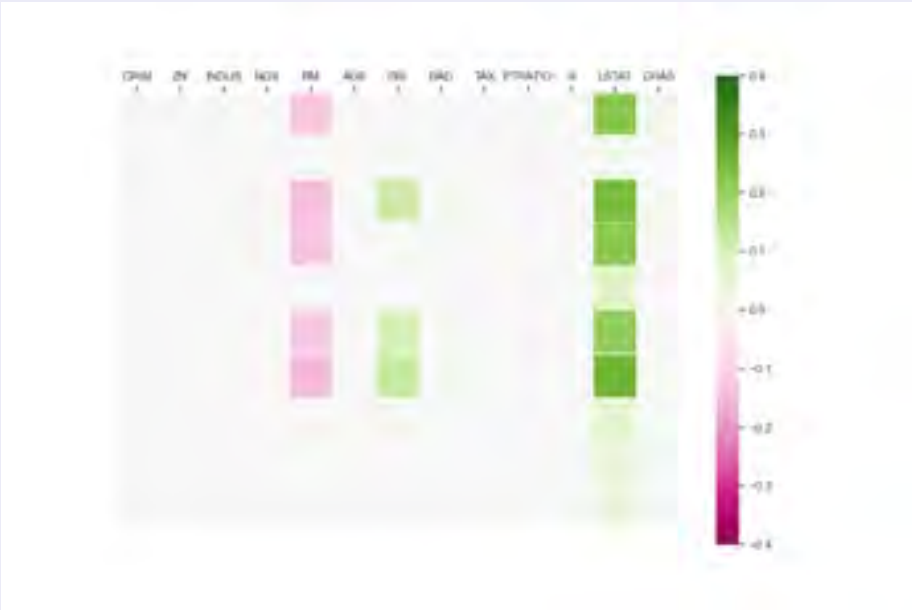


Figure: From $k = +1$ to $k^* = -1$ with $C = 0.01\ell_0 + \ell_2^2$

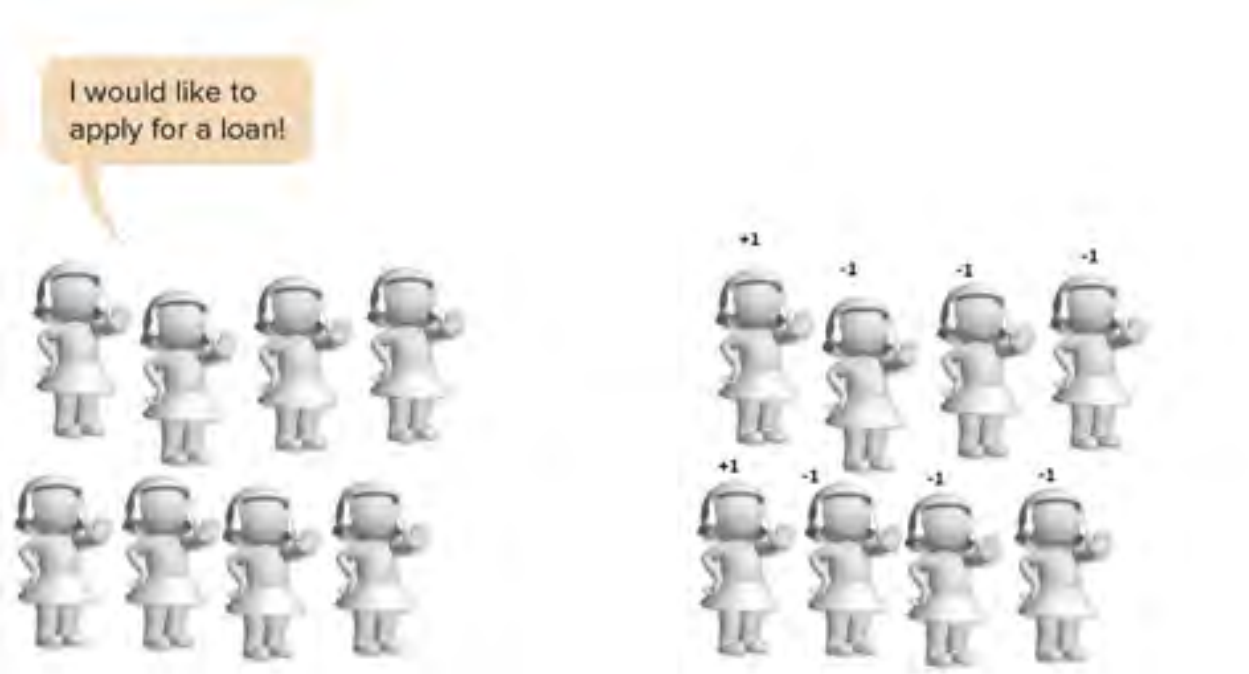
Counterfactuals for SVM

Figure: From $k = +1$ to $k^* = -1$ with $C = 0.01\ell_0 + \ell_2^2$

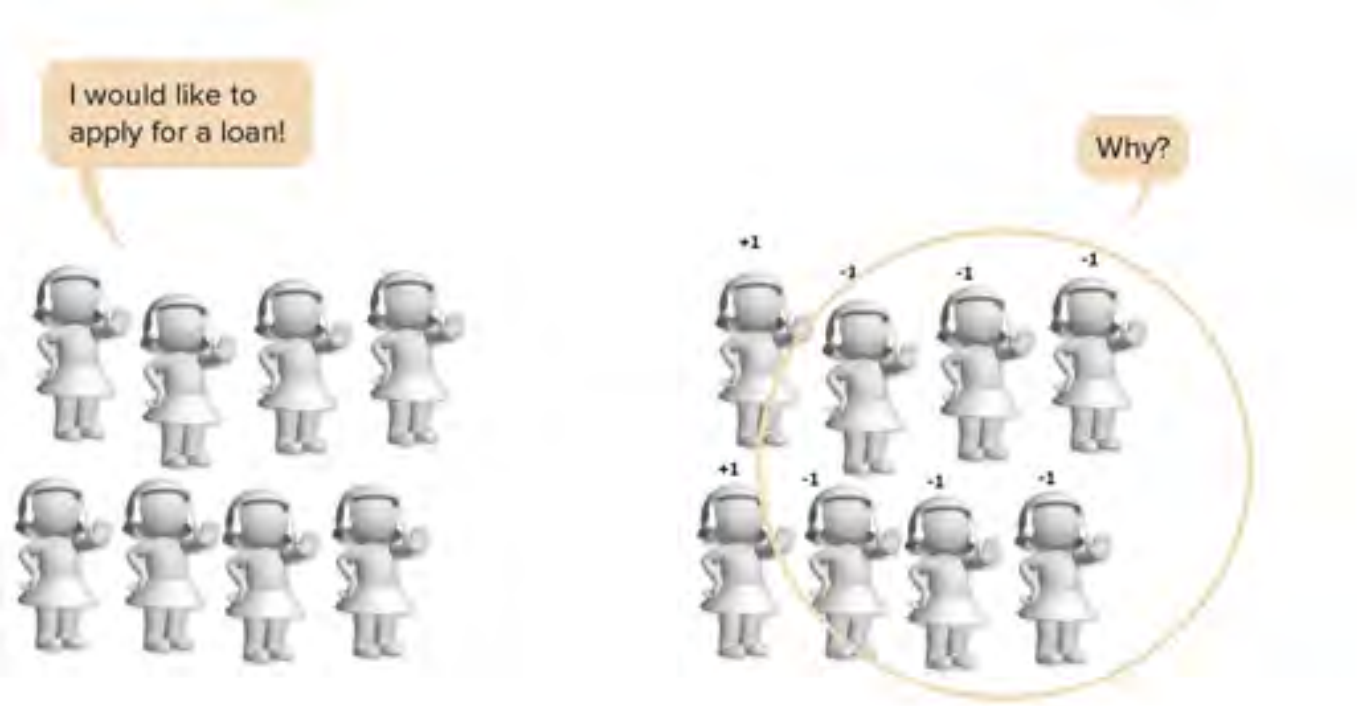
Outline

Group-Level Counterfactual Explanations

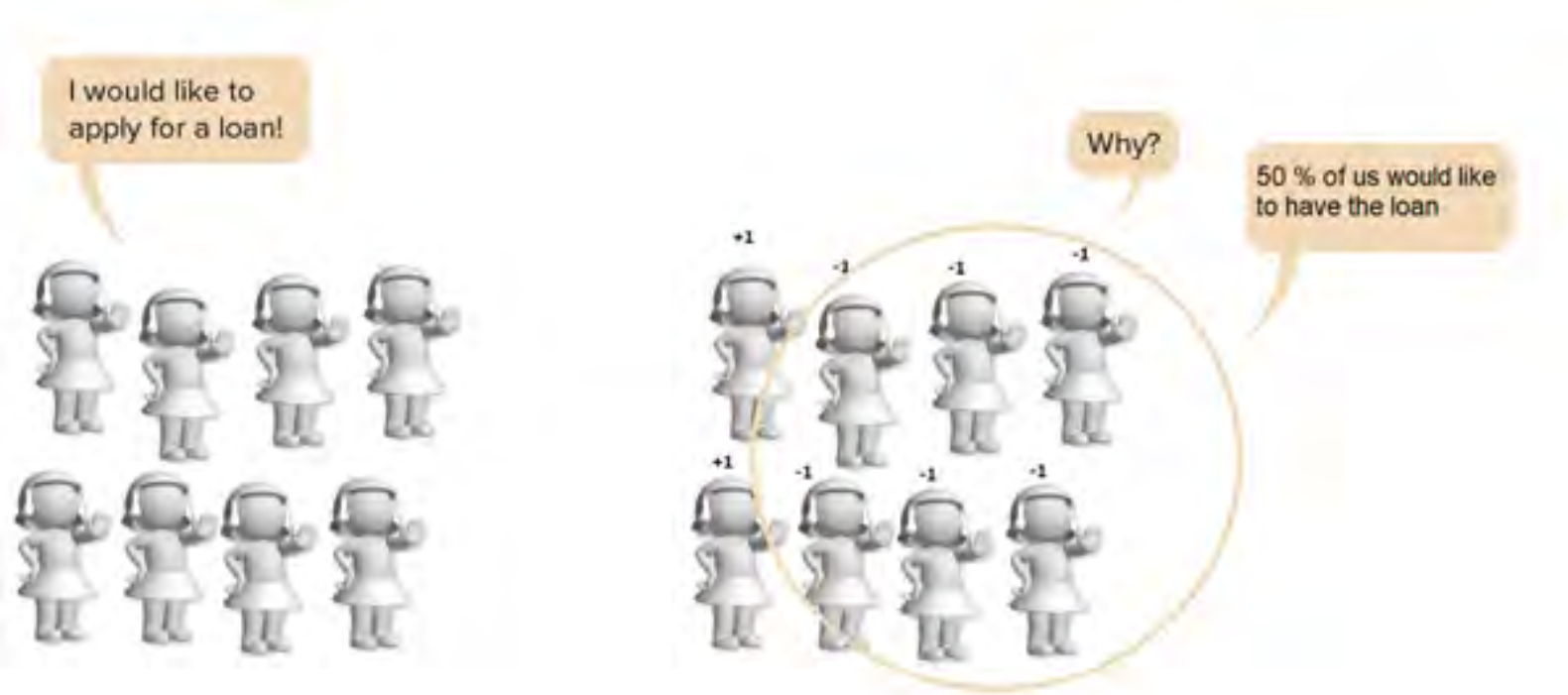
Group-Level Counterfactual Explanations



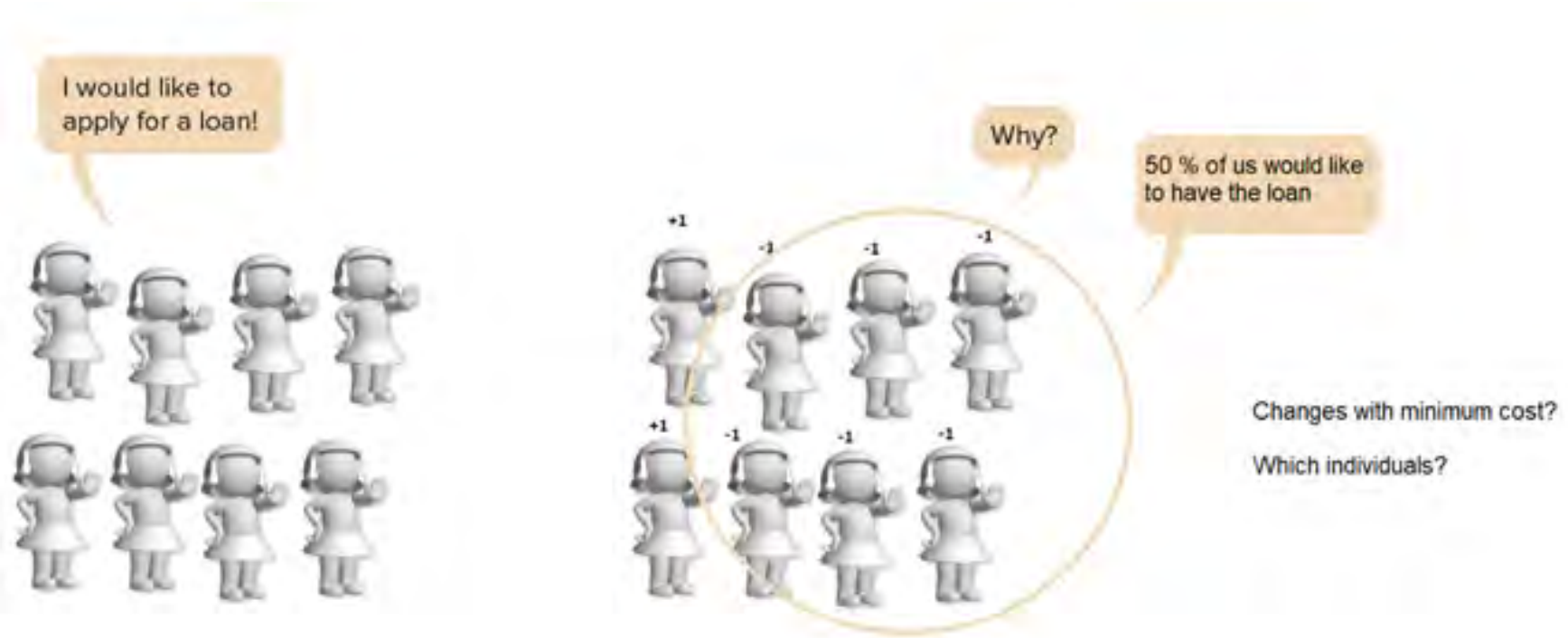
Group-Level Counterfactual Explanations



Group-Level Counterfactual Explanations



Group-Level Counterfactual Explanations



Group-Level Counterfactual Explanations

Given:

- Group of instances: $\underline{\boldsymbol{x}}^{\mathbf{0}} = (\boldsymbol{x}_1^0, \ldots, \boldsymbol{x}_l^0)$
- Subset of indices: $\mathcal{I}^* \subset \{1, \ldots, l\}$

Group-level Counterfactual Explanations

$$
\begin{cases}
\min_{\underline{\boldsymbol{x}}, \mathcal{I}^*} & C(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}}^{\mathbf{0}}) \\
\text{s.t.} & f_{k^*}(\boldsymbol{x}_i) \geq f_k(\boldsymbol{x}_i) \quad \forall k = 1, \ldots, K \quad k \neq k^* \quad \forall i \in \mathcal{I}^* \\
& \boldsymbol{x}_i = \boldsymbol{x}_i^0 \quad \forall i \notin \mathcal{I}^* \\
& \boldsymbol{x_i} \in \mathcal{X}^0 \quad i = 1, \ldots, l
\end{cases}
$$

Group-Level Counterfactual Explanations

Optimization problem for group-level counterfactuals

$$
\begin{aligned}
\min_{\underline{\boldsymbol{x}},\boldsymbol{z},\boldsymbol{y}} \quad & C(\underline{\boldsymbol{x}}^{\mathbf{0}}, \underline{\boldsymbol{x}}) \\
\text{s.t} \quad & x_{i,v(s)} - M_1(1 - z^t_{l,i}) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l,t) \quad \forall l \in \mathcal{L}^t \quad t = 1,\ldots,T \quad i = 1\ldots,I \\
& x_{i,v(s)} + M_2(1 - z^t_{l,i}) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l,t) \quad \forall l \in \mathcal{L}^t \quad t = 1,\ldots,T \quad i = 1\ldots,I \\
& \sum_{l \in \mathcal{L}^t} z^t_{l,i} = 1 \qquad t = 1,\ldots,T \quad i = 1,\ldots,I \\
& \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_{k*}} z^t_{l,i} \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}^t_{k}} y_i z^t_{l,i} \qquad \forall k = 1,\ldots,K \quad k \neq k^* \quad i = 1,\ldots,I \\
& \sum_{i=1}^{I} y_i = |\mathcal{I}^*| \\
& (1 - y_i)(\boldsymbol{x}_i^0 - \boldsymbol{x}_i) = 0 \quad i = 1,\ldots,I \\
& z^t_{l,i} \in \{0,1\} \qquad \forall l \in \mathcal{L}^t \qquad t = 1,\ldots,T \quad i = 1,\ldots,I \\
& y_i \in \{0,1\} \quad i = 1,\ldots,I \\
& \boldsymbol{x}_i \in \mathcal{X}^0 \quad i = 1,\ldots,I
\end{aligned}
$$

Particular case:

$$
C(\underline{\boldsymbol{x}}, \underline{\boldsymbol{x}}^{\mathbf{0}}) = \lambda_0 \sum_{j=1}^{J} \| \max_i |x_{ij} - x^0_{ij}| \|_0 + \lambda_2 \|\underline{\boldsymbol{x}} - \underline{\boldsymbol{x}}^0\|_2^2
$$

Group-Level Counterfactual Explanations

Optimization problem for group-level counterfactuals

$$\min_{\underline{\boldsymbol{x}},\boldsymbol{z},\boldsymbol{y},\boldsymbol{\xi}} \quad \lambda_0 \sum_{j=1}^{J} \xi_j + \lambda_2 \|\underline{\boldsymbol{x}}^{\mathbf{0}} - \underline{\boldsymbol{x}}\|_2^2$$

$$\text{s.t} \quad x_{i,v(s)} - M_1(1 - z_{l,i}^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l,t) \quad \forall l \in \mathcal{L}^t \quad t = 1,\ldots,T \quad i = 1\ldots,I$$

$$x_{i,v(s)} + M_2(1 - z_{l,i}^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l,t) \quad \forall l \in \mathcal{L}^t \quad t = 1,\ldots,T \quad i = 1\ldots,I$$

$$\sum_{l \in \mathcal{L}^t} z_{l,i}^t = 1 \qquad t = 1,\ldots,T \quad i = 1,\ldots,I$$

$$\sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_{k*}^t} z_{l,i}^t \geq \sum_{t=1}^{T} w^t \sum_{l \in \mathcal{L}_k^t} u_{l,i}^t \qquad \forall k = 1,\ldots,K \quad k \neq k^* \quad i = 1,\ldots,I$$

$$u_{l,i}^t \leq y_i \quad i = 1,\ldots,I, l \in \mathcal{L}_k^t \quad t = 1,\ldots,T \quad k = 1,\ldots,K$$

$$u_{l,i}^t \leq z_{l,i}^t \quad i = 1,\ldots,I \quad l \in \mathcal{L}_k^t \quad t = 1,\ldots,T \quad k = 1,\ldots,K$$

$$u_{l,i}^t \geq y_i + z_{l,i}^t - 1 \quad i = 1,\ldots,I \quad l \in \mathcal{L}_k^t \quad t = 1,\ldots,T \quad k = 1,\ldots,K$$

$$\sum_{i=1}^{I} y_i = |\mathcal{I}^*|$$

$$-M_3 \xi_{ij} \leq x_{ij}^0 - x_{ij} \leq M_3 \xi_{ij} \quad i = 1,\ldots I, \quad j = 1,\ldots,J$$

$$\xi_j \geq \xi_{ij} \quad i = 1,\ldots,I, \quad j = 1,\ldots J$$

$$\xi_j, \xi_{ij} \in \{0,1\} \quad i,j = 1,\ldots,J$$

$$z_{l,i}^t \in \{0,1\} \qquad \forall l \in \mathcal{L}^t \qquad t = 1,\ldots,T \quad i = 1,\ldots,I$$

$$y_i \in \{0,1\} \quad i = 1,\ldots,I$$

$$\boldsymbol{x}_i \in \mathcal{X}^0 \quad i = 1,\ldots,I$$

Computational Experiments

Instances with a criminality (CRIM) above the median value

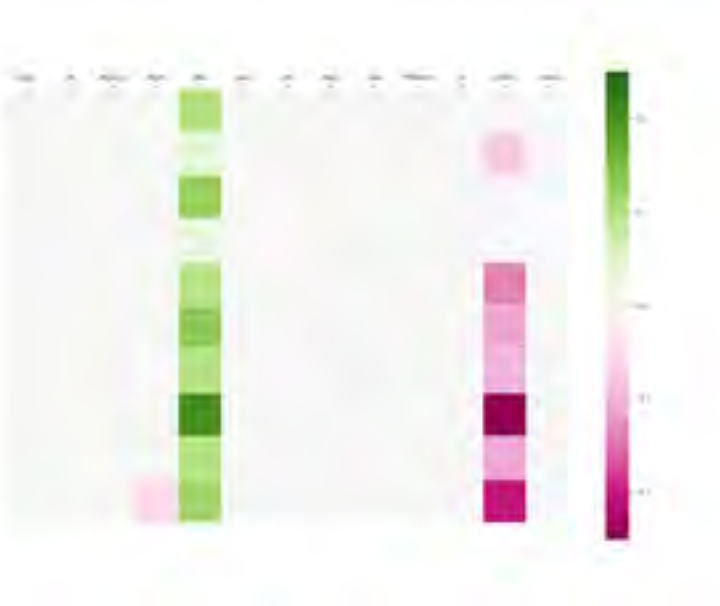


Figure: From class $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$ and $|\mathcal{I}^*| = 10$

Figure: From $k = -1$ to $k^* = +1$ with $C = 0.01\ell_0 + \ell_2^2$ and $|\mathcal{I}^*| = 5$

Outline

Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^*

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & C(\boldsymbol{x}, \boldsymbol{x}^0) \\ \text{s.t.} \quad & f_{k^*}(\boldsymbol{x}) \geq f_k(\boldsymbol{x}) \quad \forall k = 1, \ldots, K \quad k \neq k^* \\ & \boldsymbol{x} \in \mathcal{X}^0 \end{aligned}$$

Define:

- Plausability constraints:

  Using B prototypes: $\boldsymbol{x}^b \in \mathcal{X}^*$, i.e., $f(\boldsymbol{x}^b) = k^*$

$$\boldsymbol{x} = \alpha_0 \boldsymbol{x}^0 + \sum_{b=1}^{B} \alpha_b \boldsymbol{x}^b$$

- Distance: Euclidean, Dynamic Time Warping (DTW), ℓ_0

Counterfactual Explanations for functional data

Optimization problem for counterfactuals with functional data

$$\min_{\boldsymbol{x},\boldsymbol{z},\boldsymbol{\xi},\boldsymbol{u}} \quad \lambda_0 \sum_{j=1}^{J} \xi_j + \lambda_2 \sum_{j=1}^{J} d_l^2(x_j^0, x_j)$$

$$\text{s.t.} \quad x_{v(s)}(t_s) - M_1(1 - z_l^t) + \epsilon \leq c_s \quad \forall s \in \text{Left}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$x_{v(s)}(t_s) + M_2(1 - z_l^t) - \epsilon \geq c_s \quad \forall s \in \text{Right}(l,t) \quad \forall l \in \mathcal{L}^t \quad \forall t = 1, \ldots, T$$

$$\sum_{l \in \mathcal{L}^t} z_l^t = 1 \qquad \forall t = 1, \ldots, T$$

$$\sum_{t=1}^{T} \sum_{l \in \mathcal{L}_{k^*}^t} w^t z_l^t \geq \sum_{t=1}^{T} \sum_{l \in \mathcal{L}_k^t} w^t z_l^t \qquad \forall k = 1, \ldots, K \quad k \neq k^*$$

$$\boldsymbol{x} = \alpha_0 \boldsymbol{x}^0 + \sum_{b=1}^{B} \alpha_b \boldsymbol{x}^b$$

$$\sum_{b=0}^{B} \alpha_b = 1$$

$$-M_3 \xi_j \leq x_j^0(t) - x_j(t) \leq M_3 \xi_j \quad j = 1, \ldots, J \quad \forall t$$

$$-M_4 u_b \leq \alpha_b \leq M_4 u_b \quad b = 1, \ldots, B$$

$$\sum_{b=1}^{B} u_b = B^*$$

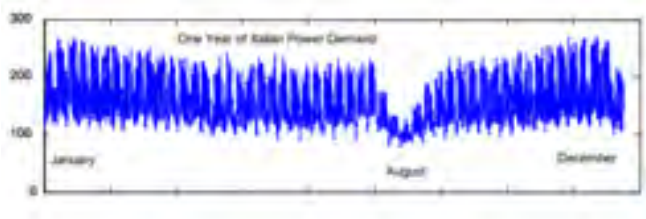
$$u_b, \xi_j, z_l^t \in \{0, 1\}$$

$$\boldsymbol{x}^b \in \mathcal{X}^* \quad \forall b = 1, \ldots, B$$

$$\boldsymbol{x} \in \mathcal{X}^0.$$

Computational Experiments

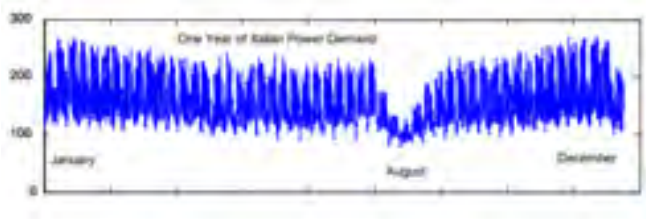
- Dataset:ItalyPowerDemand [Keogh et al., 2006]
  - 1 functional feature: power demand in 6 months
  - target: -1 October to March, $+1$ April to September



- Model: RF with $T = 200$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]

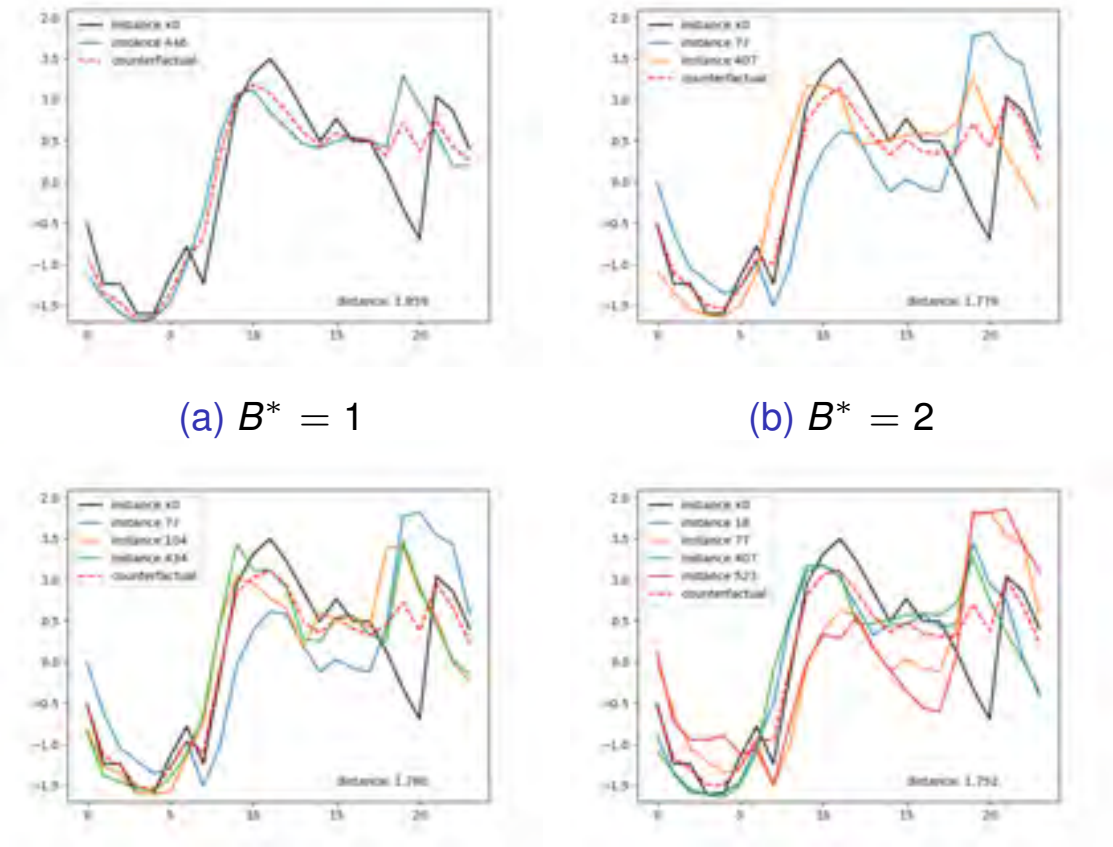
Computational Experiments

- Dataset:ItalyPowerDemand [Keogh et al., 2006]
  - 1 functional feature: power demand in 6 months
  - target: -1 October to March, $+1$ April to September



- Model: RF with $T = 200$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]

Computational experiments



(a) $B^* = 1$ (b) $B^* = 2$

(c) $B^* = 3$ (d) $B^* = 4$

Figure: Counterfactuals explanations for one instance of the ItalyPowerDemand data set who has been predicted with a Random Forest in $k^0 = -1$ and it is imposed $k^* = +1$. Different B^* prototypes have been imposed. It has been used the euclidean distance.

Computational Experiments

- Dataset: NATOPS [Ghouaiel et al., 2017]
  - 24 functional features: X, Y, and Z coordinates of the left and right hand, wrist, thumb, and elbows
  - target: -1 "All clear", $+1$ "Not clear"

- Model: RF with $T = 200$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]

Computational Experiments

- Dataset: NATOPS [Ghouaiel et al., 2017]
  - 24 functional features: X, Y, and Z coordinates of the left and right hand, wrist, thumb, and elbows
  - target: -1 "All clear", $+1$ "Not clear"

- Model: RF with $T = 200$ trees and maximum depth 4
- Pyomo optimization modelling language [Hart et al., 2011, 2017] in Python 3.7
- MILP solver Gurobi 9.0 [Gurobi Optimization, 2021]

Computational Experiments

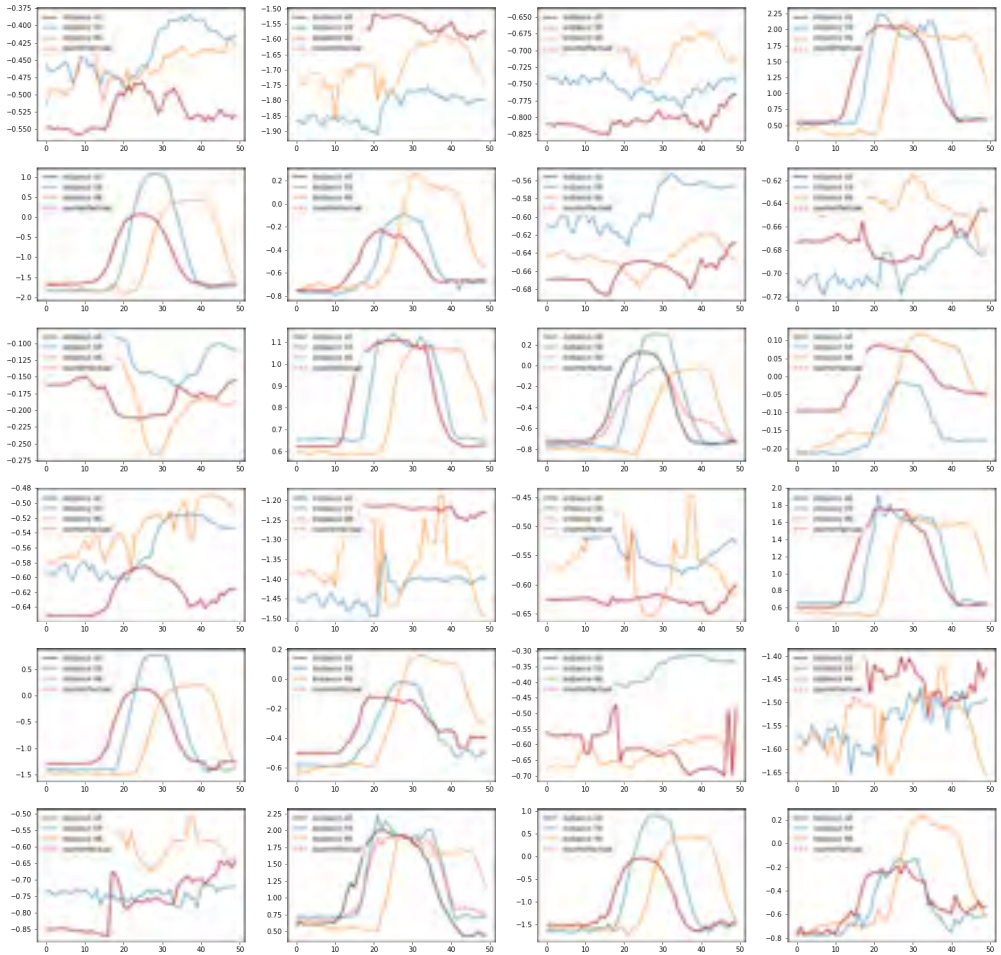
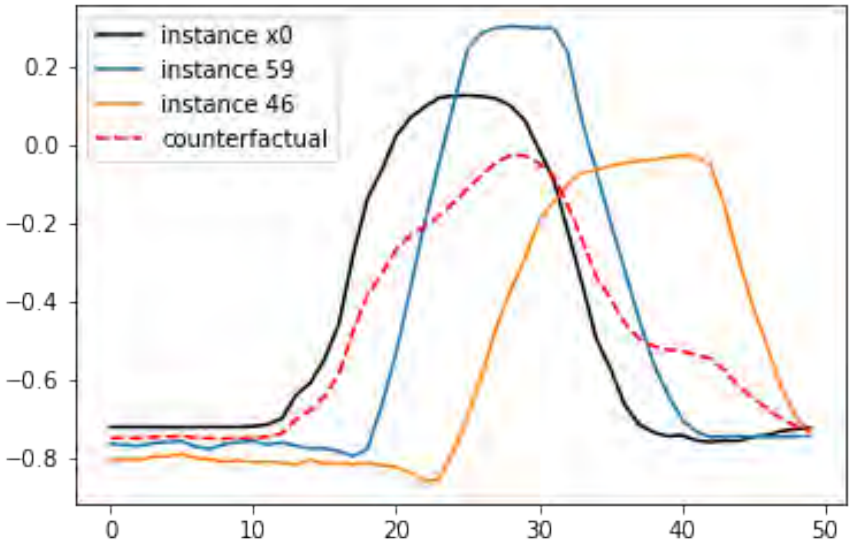
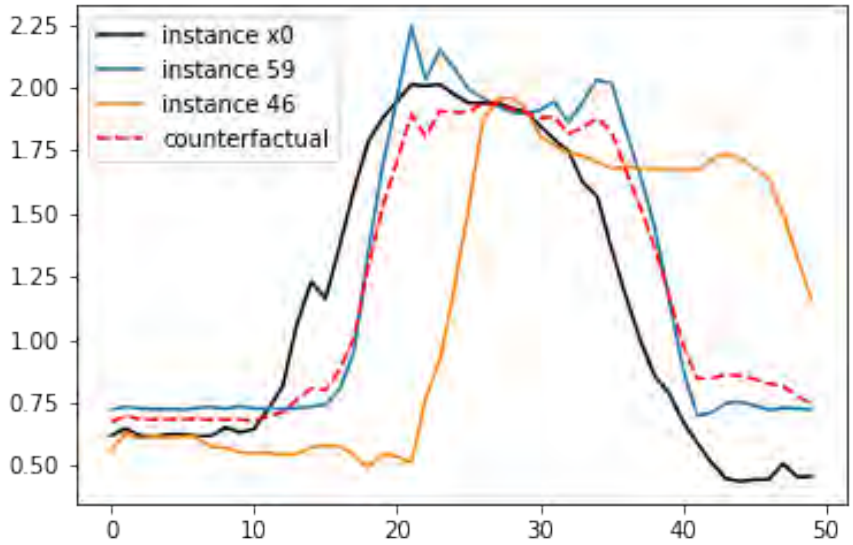


Figure: Counterfactuals explanations for one instance of the NATOPS data set who has been predicted with a Random Forest in $k^0 = +1$ and it is imposed $k^* = -1$. $B^* = 2$ prototypes has been imposed. Cost function: $C = \ell_0 + 0.005\ell_2$.

Computational Experiments



(a) Feature 11



(b) Feature 22

Figure: Changed features in the counterfactual explanation for one instance of the NATOPS data set who was predicted with a Random Forest in $k^0 = +1$ and is imposed $k^* = -1.B^* = 2$ prototypes has been imposed. Cost function: $C = \ell_0 + 0.005\ell_2$

Outline

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Summary

Today

We propose a unified approach to Counterfactual Explanations:

- by means of Mathematical Optimization
- Applicable to diverse score-based classifiers:
  - ATM like RF, XGBoost...
  - Linear models like LR, SVM...
- At two levels:
  - Individual
  - Group
- Tabular and Functional data

Future Work

- Improvements to the formulation
- Uncertainty in $\boldsymbol{x}^0$
- Robustness across models
- Casual constraints

Future Work

- Improvements to the formulation
- Uncertainty in $\boldsymbol{x}^0$
- Robustness across models
- Casual constraints

Future Work

- Improvements to the formulation
- Uncertainty in $\boldsymbol{x}^0$
- Robustness across models
- Casual constraints

Future Work

- Improvements to the formulation
- Uncertainty in $\boldsymbol{x}^0$
- Robustness across models
- Casual constraints

References I

Emilio Carrizosa, Jasone Ramırez-Ayerbe, and Dolores Romero Morales. Generating counterfactual explanations in score-based classification via mathematical optimization, 2021. URL https://www.researchgate.net/publication/353073138_Generating_Counterfactual_Explanations_in_Score-Based_Classification_via_Mathematical_Optimization.

Zhicheng Cui, Wenlin Chen, Yujie He, and Yixin Chen. Optimal action extraction for random forests and boosted trees. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 179–188, 2015.

Susanne Dandl, Christoph Molnar, Martin Binder, and Bernd Bischl. Multi-objective counterfactual explanations. In *International Conference on Parallel Problem Solving from Nature*, pages 448–469. Springer, 2020.

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

Nehla Ghouaiel, Pierre-François Marteau, and Marc Dupont. Continuous pattern detection and recognition in stream-a benchmark for online gesture recognition. *International Journal of Applied Pattern Recognition*, 4(2):146–160, 2017.

B. Goodman and S. Flaxman. European Union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3):50–57, 2017.

Riccardo Guidotti, Anna Monreale, Fosca Giannotti, Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Factual and counterfactual explanations for black box decision making. *IEEE Intelligent Systems*, 34(6):14–23, 2019.

LLC Gurobi Optimization. Gurobi optimizer reference manual, 2021. URL http://www.gurobi.com.

David Harrison Jr and Daniel L Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of environmental economics and management*, 5 (1):81–102, 1978.

William E Hart, Jean-Paul Watson, and David L Woodruff. Pyomo: modeling and solving mathematical programs in python. *Mathematical Programming Computation*, 3(3):219–260, 2011.

William E. Hart, Carl D. Laird, Jean-Paul Watson, David L. Woodruff, Gabriel A. Hackebeil, Bethany L. Nicholson, and John D. Siirola. *Pyomo–optimization modeling in python*, volume 67. Springer Science & Business Media, second edition, 2017.

Eamonn Keogh, Li Wei, Xiaopeng Xi, Stefano Lonardi, Jin Shieh, and Scott Sirowy. Intelligent icons: Integrating lite-weight data mining and visualization into gui operating systems. In *Sixth International Conference on Data Mining (ICDM'06)*, pages 912–916. IEEE, 2006.

Tony Lindgren, Panagiotis Papapetrou, Isak Samsten, and Lars Asker. Example-based feature tweaking using random forests. In *2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI)*, pages 53–60. IEEE, 2019.

Ana Lucic, Harrie Oosterhuis, Hinda Haned, and Maarten de Rijke. FOCUS: Flexible optimizable counterfactual explanations for tree ensembles. *arXiv preprint arXiv:1911.12199*, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

References II

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

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

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 344–350, 2020.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " 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, 2016.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31:841, 2017.

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

Thank you for your attention!

mrayerbe@us.es

All this research is available at:

https://www.researchgate.net/profile/Jasone-Ramirez-Ayerbe