Seeking Transparency in Machine Learning Through Optimized Explanations

Focus Period Linköping 2022
Workshop *Hybrid AI – Where data-driven and model-based methods meet*
November 1, 2022

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

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

Outline

Outline

1. Machine Learning for High-Stakes Decision Making
2. Randomized Optimal Classification and Regression Trees
3. Optimized Counterfactual Explanations for Score-Based Classifiers
4. Some thoughts

Transparency

When training a machine learning algorithm,
accuracy of its predictions matters, as does the **transparency**

- **Transparency** is desirable [Freitas, 2014, Rudin et al., 2022], e.g., in **medical diagnosis** [Ustun and Rudin, 2016];
- It is required by regulators for models aiding, e.g., **credit scoring** [Baesens et al., 2003] and **judicial** [Ridgeway, 2013] decisions;
- From 2018 onwards the EU extended this requirement by imposing the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- There is a growing number of **Explainable Artificial Intelligence (XAI)** tools, Ghorbani and Zou [2020], Gunning and Aha [2019], Holter et al. [2018], Miller [2019]
  https://www.darpa.mil/program/explainable-artificial-intelligence
  https://www.microsoft.com/en-us/research/publication/interpretml-a-unified-framework-for-machine-learning-interpretability/
  https://community.fico.com/s/explainable-machine-learning-challenge

Transparency

When training a machine learning algorithm,
accuracy of its predictions matters, as does the **transparency**

- **Transparency** is desirable [Freitas, 2014, Rudin et al., 2022], e.g., in **medical diagnosis** [Ustun and Rudin, 2016];
- It is required by regulators for models aiding, e.g., **credit scoring** [Baesens et al., 2003] and **judicial** [Ridgeway, 2013] decisions;
- From 2018 onwards the EU extended this requirement by imposing the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- There is a growing number of **Explainable Artificial Intelligence (XAI)** tools, Ghorbani and Zou [2020], Gunning and Aha [2019], Holter et al. [2018], Miller [2019]
  https://www.darpa.mil/program/explainable-artificial-intelligence
  https://www.microsoft.com/en-us/research/publication/interpretml-a-unified-framework-for-machine-learning-interpretability/
  https://community.fico.com/s/explainable-machine-learning-challenge

Transparency

When training a machine learning algorithm,
accuracy of its predictions matters, as does the **transparency**

- **Transparency** is desirable [Freitas, 2014, Rudin et al., 2022], e.g., in **medical diagnosis** [Ustun and Rudin, 2016];
- It is required by regulators for models aiding, e.g., **credit scoring** [Baesens et al., 2003] and **judicial** [Ridgeway, 2013] decisions;
- From 2018 onwards the EU extended this requirement by imposing the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- There is a growing number of **Explainable Artificial Intelligence (XAI)** tools, Ghorbani and Zou [2020], Gunning and Aha [2019], Holter et al. [2018], Miller [2019]

  https://www.darpa.mil/program/explainable-artificial-intelligence
  https://www.microsoft.com/en-us/research/publication/interpretml-a-unified-framework-for-machine-learning-interpretability/
  https://community.fico.com/s/explainable-machine-learning-challenge

Transparency

When training a machine learning algorithm,
accuracy of its predictions matters, as does the **transparency**

- **Transparency** is desirable [Freitas, 2014, Rudin et al., 2022], e.g., in **medical diagnosis** [Ustun and Rudin, 2016];
- It is required by regulators for models aiding, e.g., **credit scoring** [Baesens et al., 2003] and **judicial** [Ridgeway, 2013] decisions;
- From 2018 onwards the EU extended this requirement by imposing the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- There is a growing number of **Explainable Artificial Intelligence (XAI)** tools, Ghorbani and Zou [2020], Gunning and Aha [2019], Holter et al. [2018], Miller [2019]

  https://www.darpa.mil/program/explainable-artificial-intelligence
  https://www.microsoft.com/en-us/research/publication/interpretml-a-unified-framework-for-machine-learning-interpretability/
  https://community.fico.com/s/explainable-machine-learning-challenge

Transparency

When training a machine learning algorithm,
accuracy of its predictions matters, as does the **transparency**

- **Transparency** is desirable [Freitas, 2014, Rudin et al., 2022], e.g., in **medical diagnosis** [Ustun and Rudin, 2016];
- It is required by regulators for models aiding, e.g., **credit scoring** [Baesens et al., 2003] and **judicial** [Ridgeway, 2013] decisions;
- From 2018 onwards the EU extended this requirement by imposing the so-called **right-to-explanation** in algorithmic decision making [European Commission, 2020, Goodman and Flaxman, 2017];
- There is a growing number of **Explainable Artificial Intelligence (XAI)** tools, Ghorbani and Zou [2020], Gunning and Aha [2019], Holter et al. [2018], Miller [2019]

  https://www.darpa.mil/program/explainable-artificial-intelligence

  https://www.microsoft.com/en-us/research/publication/interpretml-a-unified-framework-for-machine-learning-interpretability/

  https://community.fico.com/s/explainable-machine-learning-challenge

Enhancing transparency

Focus on the data at hand

- **Sparseness (fewer features):**

  Atamtürk and Gomez [2019], Benítez-Peña et al. [2019, 2020, 2021, 2022], Bertsimas et al. [2016], Blanquero et al. [2021b], Carrizosa et al. [2022c,f], Fountoulakis and Gondzio [2016], Kenney et al. [2021], Maldonado et al. [2014], Rinaldi et al. [2010], Rinaldi and Sciandrone [2010]

Enhancing transparency

Focus on the data at hand

- **Finding prototypes (representative individuals):**
  Carrizosa et al. [2007, 2021d, 2022a,d], Hart [1968], Wilfong [1992]

Enhancing transparency

Focus on the model itself

- **Enhancing interpretability of black-box methods:**

  Support Vector Machines (SVM), Deep Learning (DL) and even Random Forests (RF) are seen as black-boxes, and there have been many efforts to enhance their interpretability

  Bénard et al. [2019], Carrizosa and Romero Morales [2013], Carrizosa et al. [2010, 2011, 2016, 2017, 2021a,b,c, 2022e], Chevaleyre et al. [2013], Golea and Marchand [1993], Lawless et al. [2022], Li et al. [2017], Ustun and Rudin [2016]

Enhancing transparency

Focus on the model itself

- **Building easy-to-understand structures such as rules and trees:**

  Baesens et al. [2003], Blanquero et al. [2021a, 2020, 2022a], Bertsimas and Dunn [2017], Carrizosa et al. [2021d,e], Dash et al. [2018], D'Onofrio et al. [2022], Martens and Provost [2014], Orsenigo and Vercellis [2003, 2004]

High-stakes decision-making

When **high-stakes decisions** are taken, new **demands** on the machine learning algorithm arise:

- **Fairness**, to avoid that **the algorithm discriminates** against sensitive groups, e.g., age, gender, race, religion, socio-economic status, migrants

  Media has reported many of these cases, e.g., Compas, Amazon, A-Levels in the UK, social benefits in The Netherlands

- **Local and counterfactual explanations**, to **understand how the algorithm** arrives at individual predictions and to **give feedback on how the algorithm** would have arrived to the desired prediction

  There is a focus on the impact of the algorithm at the ***individual/instance*** *level, e.g., the convicted person, the online customer, the teenager, the social benefits applicant*

High-stakes decision-making

When **high-stakes decisions** are taken, new **demands** on the machine learning algorithm arise:

- **Fairness**, to avoid that **the algorithm discriminates** against sensitive groups, e.g., age, gender, race, religion, socio-economic status, migrants

  Media has reported many of these cases, e.g., Compas, Amazon, A-Levels in the UK, social benefits in The Netherlands

- **Local and counterfactual explanations**, to **understand how the algorithm** arrives at individual predictions and to **give feedback on how the algorithm** would have arrived to the desired prediction

  There is a focus on the impact of the algorithm at the **individual/instance** *level, e.g., the convicted person, the online customer, the teenager, the social benefits applicant*

High-stakes decision-making

When **high-stakes decisions** are taken, new **demands** on the machine learning algorithm arise:

- **Fairness**, to avoid that **the algorithm discriminates** against sensitive groups, e.g., age, gender, race, religion, socio-economic status, migrants

  Media has reported many of these cases, e.g., Compas, Amazon, A-Levels in the UK, social benefits in The Netherlands

- **Local and counterfactual explanations**, to **understand how the algorithm** arrives at individual predictions and to **give feedback on how the algorithm** would have arrived to the desired prediction

  There is a focus on the impact of the algorithm at the **individual/instance** *level, e.g., the convicted person, the online customer, the teenager, the social benefits applicant*

Fairness

There is a growing literature addressing fairness concerns [Aghaei et al., 2019, Besse et al., 2022, Carrizosa et al., 2022b, Mehrabi et al., 2022, Zafar et al., 2017a,b]

Important!!! It is not enough to check that these sensitive features are not used directly by the model, as they can be used indirectly through other features

For a group of sensitive observations, we may want, for instance, to

- control **accuracy in the sensitive group**, or
- ensure that accuracy in the sensitive group is close to that in the whole group, or
- ensure that predictions in the sensitive group resemble to those in the whole group, e.g., the mean is similar

Fairness

There is a growing literature addressing fairness concerns [Aghaei et al., 2019, Besse et al., 2022, Carrizosa et al., 2022b, Mehrabi et al., 2022, Zafar et al., 2017a,b]

Important!!! It is not enough to check that these sensitive features are not used directly by the model, as they can be used indirectly through other features

For a group of sensitive observations, we may want, for instance, to

- control **accuracy in the sensitive group**, or
- ensure that accuracy in the sensitive group is close to that in the whole group, or
- ensure that predictions in the sensitive group resemble to those in the whole group, e.g., the mean is similar

Fairness

There is a growing literature addressing fairness concerns [Aghaei et al., 2019, Besse et al., 2022, Carrizosa et al., 2022b, Mehrabi et al., 2022, Zafar et al., 2017a,b]

Important!!! It is not enough to check that these sensitive features are not used directly by the model, as they can be used indirectly through other features

For a group of sensitive observations, we may want, for instance, to

- control **accuracy in the sensitive group**, or
- ensure that accuracy in the sensitive group is close to that in the whole group, or
- ensure that predictions in the sensitive group resemble to those in the whole group, e.g., the mean is similar

Fairness

There is a growing literature addressing fairness concerns [Aghaei et al., 2019, Besse et al., 2022, Carrizosa et al., 2022b, Mehrabi et al., 2022, Zafar et al., 2017a,b]

Important!!! It is not enough to check that these sensitive features are not used directly by the model, as they can be used indirectly through other features

For a group of sensitive observations, we may want, for instance, to

- control **accuracy in the sensitive group**, or
- ensure that accuracy in the sensitive group is close to that in the whole group, or
- ensure that predictions in the sensitive group resemble to those in the whole group, e.g., the mean is similar

Fairness

There is a growing literature addressing fairness concerns [Aghaei et al., 2019, Besse et al., 2022, Carrizosa et al., 2022b, Mehrabi et al., 2022, Zafar et al., 2017a,b]

Important!!! It is not enough to check that these sensitive features are not used directly by the model, as they can be used indirectly through other features

For a group of sensitive observations, we may want, for instance, to

- control **accuracy in the sensitive group**, or
- ensure that accuracy in the sensitive group is close to that in the whole group, or
- ensure that predictions in the sensitive group resemble to those in the whole group, e.g., the mean is similar

Local explanations

- Local explanations help to understand the role of each feature in the prediction made for an individual [Ghorbani and Zou, 2020, Gunning and Aha, 2019, Holter et al., 2018, Miller, 2019]

- If the model is linear,

$$y = \alpha + \boldsymbol{\beta}^\top \mathbf{x},$$

  we can easily provide local explanations

  if x_j increases by 1 unit, then y increases by β_j units,

  which does not depend on the individual at hand

- Nowadays, it is common in XAI to provide the explanations from a **surrogate of the black-box model** [Lundberg and Lee, 2017, Lundberg et al., 2020, Ribeiro et al., 2016], while there are fewer approaches that can provide those **by design**

Local explanations

- Local explanations help to understand the role of each feature in the prediction made for an individual [Ghorbani and Zou, 2020, Gunning and Aha, 2019, Holter et al., 2018, Miller, 2019]

- If the model is linear,
$$y = \alpha + \boldsymbol{\beta}^\top \mathbf{x},$$
we can easily provide local explanations
if x_j increases by 1 unit, then y increases by β_j units,
which does not depend on the individual at hand

- Nowadays, it is common in XAI to provide the explanations from a **surrogate of the black-box model** [Lundberg and Lee, 2017, Lundberg et al., 2020, Ribeiro et al., 2016], while there are fewer approaches that can provide those **by design**

Local explanations

- Local explanations help to understand the role of each feature in the prediction made for an individual [Ghorbani and Zou, 2020, Gunning and Aha, 2019, Holter et al., 2018, Miller, 2019]
- If the model is linear,

  $$y = \alpha + \boldsymbol{\beta}^\top \mathbf{x},$$

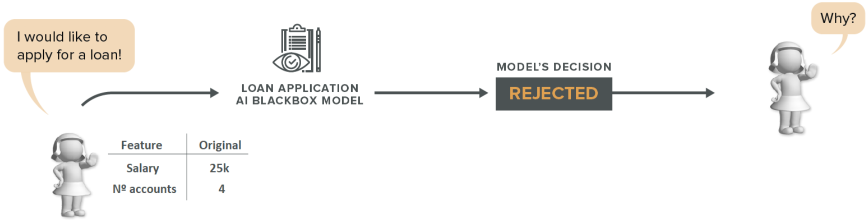
  we can easily provide local explanations

  if x_j increases by 1 unit, then y increases by β_j units,

  which does not depend on the individual at hand
- Nowadays, it is common in XAI to provide the explanations from a **surrogate of the black-box model** [Lundberg and Lee, 2017, Lundberg et al., 2020, Ribeiro et al., 2016], while there are fewer approaches that can provide those **by design**

Counterfactual explanations

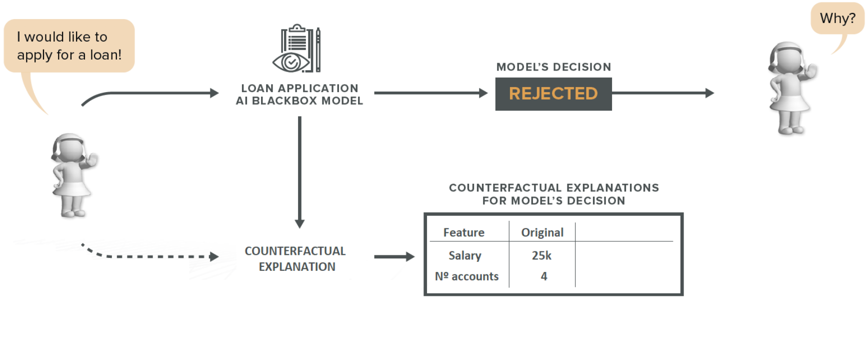
- For a given individual, which features need to change to get a desired prediction



- *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 work in this area is recent [Forel et al., 2022, Guidotti, 2022, Karimi et al., 2020, Maragno et al., 2022, Wachter et al., 2017]

Counterfactual explanations

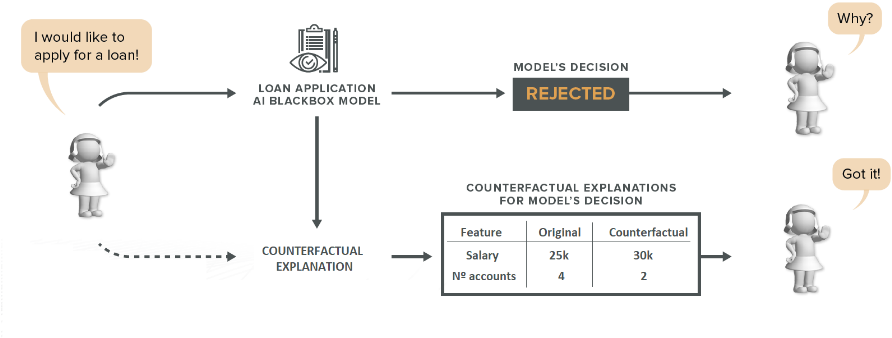
- For a given individual, which features need to change to get a desired prediction



- *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 work in this area is recent [Forel et al., 2022, Guidotti, 2022, Karimi et al., 2020, Maragno et al., 2022, Wachter et al., 2017]

Counterfactual explanations

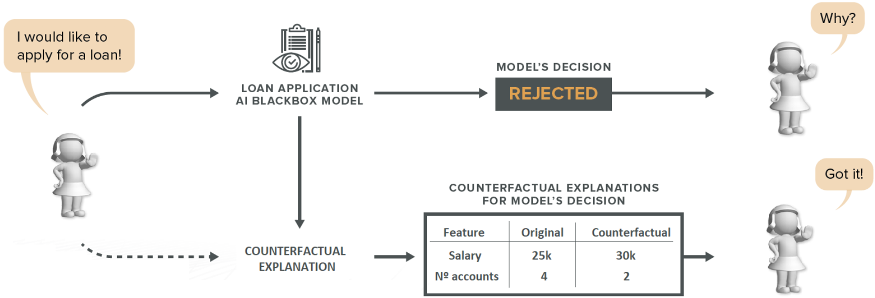
- For a given individual, which features need to change to get a desired prediction



- *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 work in this area is recent [Forel et al., 2022, Guidotti, 2022, Karimi et al., 2020, Maragno et al., 2022, Wachter et al., 2017]

Counterfactual explanations

- For a given individual, which features need to change to get a desired prediction



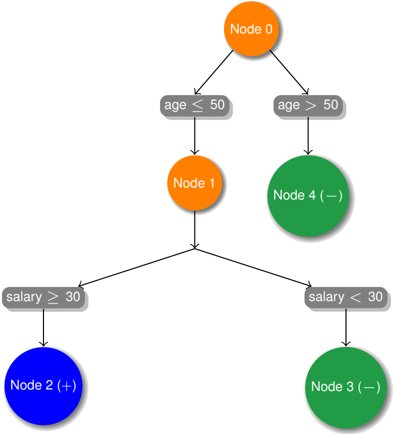
Feature	Original	Counterfactual
Salary	25k	30k
Nº accounts	4	2

- *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 work in this area is recent [Forel et al., 2022, Guidotti, 2022, Karimi et al., 2020, Maragno et al., 2022, Wachter et al., 2017]

Outline

Classification and Regression Trees

'+' (good payers) vs '−' (bad payers)



See our recent review on optimal trees

Carrizosa et al. [2021], Mathematical optimization in classification and regression trees, TOP, 29(1):5-33.
In Open Access

Mixed Integer Linear Optimization

- Aghaei et al. [2020]
- Bertsimas and Dunn [2017]
- Firat et al. [2020]
- Günlük et al. [2021]

Other paradigms

- CP, Verhaeghe et al. [2019]
- DP, Demirović et al. [2022]
- SAT, Narodytska et al. [2018]

Optimal Randomized Classification and Regression Trees

In Blanquero et al. [2020, 2021a, 2022a,b], we propose

Optimal Randomized Classification and Regression Trees:

- We model probabilistic (as opposed to deterministic) splitting rules
- We develop a Continuous Optimization formulation

With:

- Accuracy and sparsity tradeoff
- Tabular and functional data
- Fairness constraints
- Local and counterfactual explanations by design

Optimal Randomized Classification and Regression Trees

In Blanquero et al. [2020, 2021a, 2022a,b], we propose

Optimal Randomized Classification and Regression Trees:

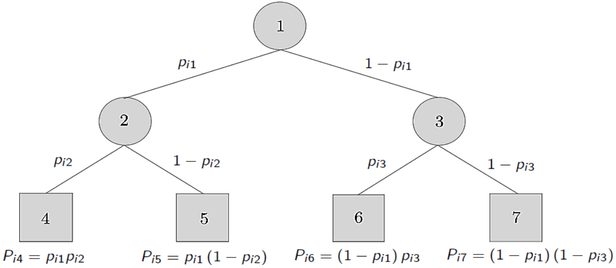
- We model probabilistic (as opposed to deterministic) splitting rules
- We develop a Continuous Optimization formulation

With:

- Accuracy and sparsity tradeoff
- Tabular and functional data
- Fairness constraints
- Local and counterfactual explanations by design

Optimal Randomized Regression Trees

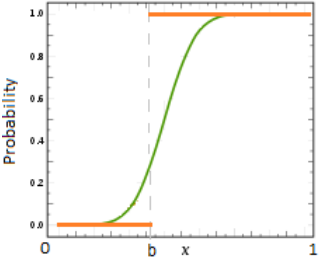
- A sample $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$.
- A maximal binary tree of depth D, with branch $t \in \tau_B$ and leaf $t \in \tau_L$ nodes.



- Oblique splits:
  - a_{jt} coefficient of predictor variable j in the oblique cut at branch node $t \in \tau_B$,
  - μ_t intercept at the oblique cut at branch node $t \in \tau_B$.

Optimal Randomized Regression Trees

- Probabilistic cuts, defined through $F(\cdot)$, the smooth CDF of a univariate continuous random variable



- Probabilities

$$p_{it}\left(\boldsymbol{a}_{\cdot t}, \mu_t\right) = F\left(\frac{1}{p}\sum_{j=1}^{p} a_{jt}x_{ij} - \mu_t\right), \;\; i = 1, \ldots, N, \;\; t \in \tau_B.$$

$$P_{it}\left(\boldsymbol{a}, \boldsymbol{\mu}\right) \equiv \mathbb{P}\left(\boldsymbol{x}_i \in t\right) = \prod_{t_l \in \mathcal{N}_L(t)} p_{it_l}\left(\boldsymbol{a}_{\cdot t_l}, \mu_{t_l}\right) \prod_{t_r \in \mathcal{N}_R(t)} \left(1 - p_{it_r}\left(\boldsymbol{a}_{\cdot t_r}, \mu_{t_r}\right)\right), \;\; i = 1, \ldots, N, \;\; t \in \tau_L.$$

Optimal Randomized Regression Trees (ORRT)

The ORRT model

$$\text{minimize}_{(\boldsymbol{a},\boldsymbol{\mu},\tilde{\boldsymbol{a}},\tilde{\boldsymbol{\mu}})\in\mathbb{R}^{(p+1)(|\tau_B|+|\tau_L|)}} \quad \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t\in\tau_L} P_{it}\left(\boldsymbol{a},\boldsymbol{\mu}\right)\left(\tilde{\boldsymbol{a}}_{\cdot t}^{\top}\boldsymbol{x}_i+\tilde{\mu}_t\right)-y_i\right)^2 \quad \textbf{(MSE)}$$

$$+\lambda^{\text{local}}\sum_{j=1}^{p}\|(\boldsymbol{a}_{j\cdot},\tilde{\boldsymbol{a}}_{j\cdot})\|_1 \quad \textbf{(local sparsity)}$$

$$+\lambda^{\text{global}}\sum_{j=1}^{p}\|(\boldsymbol{a}_{j\cdot},\tilde{\boldsymbol{a}}_{j\cdot})\|_\infty \quad \textbf{(global sparsity)}$$

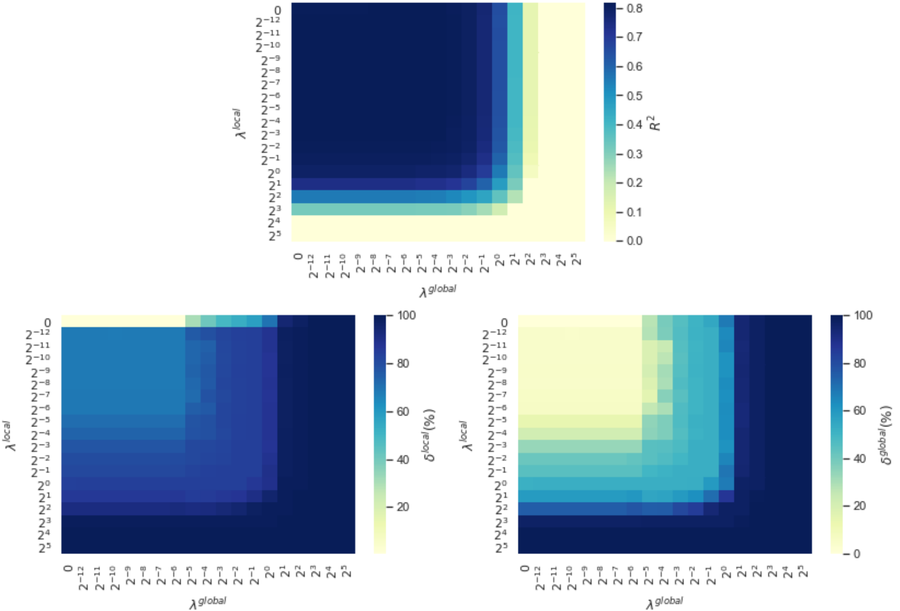
There exists an equivalent nonlinear smooth formulation

This speaks favorably about the explainability of our tree model

There are no decision variables directly linked to the observations

This speaks favorably about the scalability of our approach

Tradeoff between accuracy and sparsity for `ailerons` dataset



Local explanations

Let $(\boldsymbol{a}^*, \boldsymbol{\mu}^*, \tilde{\boldsymbol{a}}^*, \tilde{\boldsymbol{\mu}}^*)$ be the optimal solution. For an incoming individual with predictor vector $\mathbf{x}$, the expected outcome is equal to

$$\mathbf{x} \to \Pi(\mathbf{x}) := \sum_{t \in \tau_L} P_{\mathbf{x}\,t}\,(\boldsymbol{a}^*, \boldsymbol{\mu}^*)\ (\tilde{\boldsymbol{a}}_{\cdot t}^{*\top} \boldsymbol{x}_i + \tilde{\mu}_t^*),$$

where $P_{\mathbf{x}\,t}\,(\cdot,\cdot)$ is defined similarly to $P_{it}\,(\cdot,\cdot)$ with $\mathbf{x}$ replacing $\mathbf{x}_i$.

The *smoothness* of $\Pi(\cdot)$ is crucial to be able to provide **local explanations** to ORRT.

Local explanations

Thus, the matrix of partial derivatives

$$\left(\frac{\partial \Pi}{\partial x_j}(\boldsymbol{x}^0)\right)_{j=1,\ldots,p}$$

gives information on the sensitivity of the outcomes Π around $\boldsymbol{x}^0$.

Illustration of local explanations in ORRT in the `housing` dataset

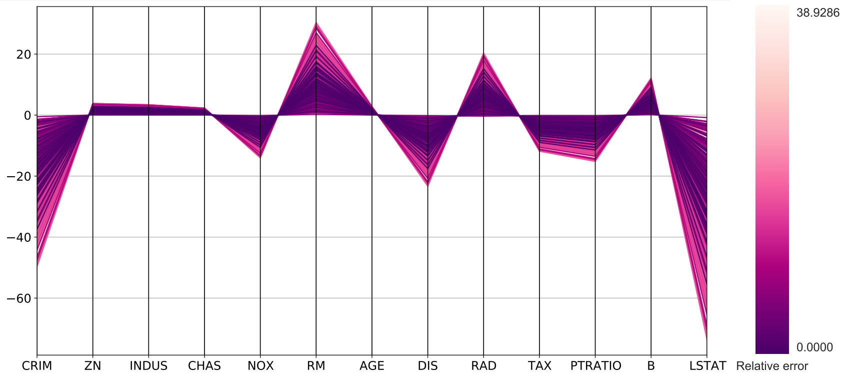
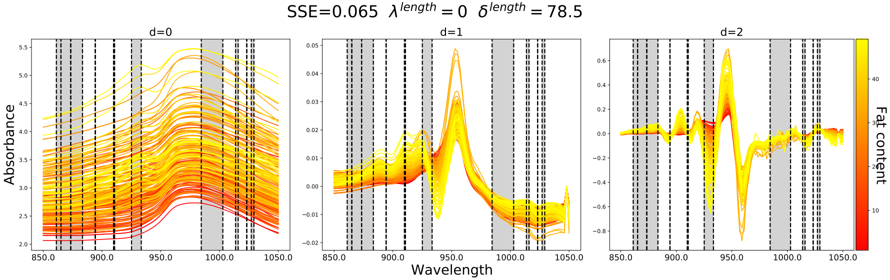
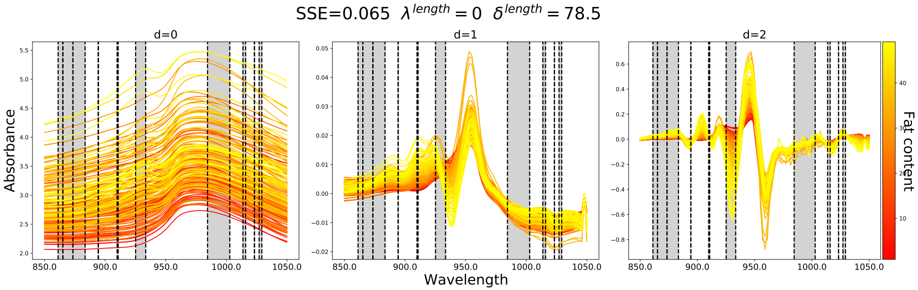


Figure: The `housing` dataset: local explanations for ORRT with $\lambda^{\text{local}} = 0$ and $\lambda^{\text{global}} = \frac{2^2}{13}$, with MSE $= 15.5654$ and $R^2 = 0.8156$.

Detecting critical intervals for functional data with S-ORRT-FD



Detecting critical intervals for functional data with S-ORRT-FD



Outline

Minimum Cost Counterfactual Explanations

The input

- The feature space $\mathcal{X} \subset \mathbb{R}^J$ for a K-class problem
- **A classifier** $\mathcal{M} : \mathcal{X} \longrightarrow \{1, \ldots, K\}$
- An instance $\mathbf{x}^0$ **seeking an explanation** on how to change to $\mathbf{x} \,|\, \mathcal{M}(\mathbf{x}) = k^+$
  - $\mathcal{M}(\mathbf{x}) = k^+$ can mean getting a good credit score, getting social benefits, ...

Minimum Cost Counterfactual Explanations

The problem

- Find $\mathbf{x}$, the counterfactual to $\mathbf{x}^0$, **of minimum cost** such that $\mathbf{x}$ is classified in k^+

Minimum Cost Counterfactual Explanations

The input

- The feature space $\mathcal{X} \subset \mathbb{R}^J$ for a K-class problem
- **A classifier** $\mathcal{M} : \mathcal{X} \longrightarrow \{1, \ldots, K\}$
- An instance $\mathbf{x}^0$ **seeking an explanation** on how to change to $\mathbf{x} \,|\, \mathcal{M}(\mathbf{x}) = k^+$
  - $\mathcal{M}(\mathbf{x}) = k^+$ can mean getting a good credit score, getting social benefits, ...

Minimum Cost Counterfactual Explanations

The problem

- Find $\mathbf{x}$, the counterfactual to $\mathbf{x}^0$, **of minimum cost** such that $\mathbf{x}$ is classified in k^+

Minimum Cost Counterfactual Explanations

The input

- The feature space $\mathcal{X} \subset \mathbb{R}^J$ for a K-class problem
- **A classifier** $\mathcal{M} : \mathcal{X} \longrightarrow \{1, \ldots, K\}$
- An instance $\mathbf{x}^0$ **seeking an explanation** on how to change to $\mathbf{x} \mid \mathcal{M}(\mathbf{x}) = k^+$
  - $\mathcal{M}(\mathbf{x}) = k^+$ can mean getting a good credit score, getting social benefits, ...

Minimum Cost Counterfactual Explanations

The problem

- Find $\mathbf{x}$, the counterfactual to $\mathbf{x}^0$, **of minimum cost** such that $\mathbf{x}$ is classified in k^+

Minimum Cost Counterfactual Explanations

The input

- The feature space $\mathcal{X} \subset \mathbb{R}^J$ for a K-class problem
- **A classifier** $\mathcal{M} : \mathcal{X} \longrightarrow \{1, \ldots, K\}$
- An instance $\mathbf{x}^0$ **seeking an explanation** on how to change to $\mathbf{x} \mid \mathcal{M}(\mathbf{x}) = k^+$
  - $\mathcal{M}(\mathbf{x}) = k^+$ can mean getting a good credit score, getting social benefits, ...

Minimum Cost Counterfactual Explanations

The problem

- Find $\mathbf{x}$, the counterfactual to $\mathbf{x}^0$, **of minimum cost** such that $\mathbf{x}$ is classified in k^+

Minimum Cost Counterfactual Explanations

The input

- The feature space $\mathcal{X} \subset \mathbb{R}^J$ for a K-class problem
- **A classifier** $\mathcal{M} : \mathcal{X} \longrightarrow \{1, \ldots, K\}$
- An instance $\mathbf{x}^0$ **seeking an explanation** on how to change to $\mathbf{x} \mid \mathcal{M}(\mathbf{x}) = k^+$
  - $\mathcal{M}(\mathbf{x}) = k^+$ can mean getting a good credit score, getting social benefits, ...

Minimum Cost Counterfactual Explanations

The problem

- Find $\mathbf{x}$, the counterfactual to $\mathbf{x}^0$, **of minimum cost** such that $\mathbf{x}$ is classified in k^+

Minimum Cost Counterfactual Explanations

In Carrizosa et al. [2021a, 2022a], we propose

a unified approach to counterfactual explanations for **score-based classifiers** such as **Logistic Regression, Random Forests, Support Vector Machines, or XGBoost**

- Controlling sparsity
- Modeling, e.g., actionability and plausibility constraints
- Dealing with both tabular as well as functional data
- Individual and collective explanations

Minimum Cost Counterfactual Explanations

In Carrizosa et al. [2021a, 2022a], we propose

a unified approach to counterfactual explanations for **score-based classifiers** such as **Logistic Regression, Random Forests, Support Vector Machines, or XGBoost**

- Controlling sparsity
- Modeling, e.g., actionability and plausibility constraints
- Dealing with both tabular as well as functional data
- Individual and collective explanations

Minimum Cost Counterfactual Explanations

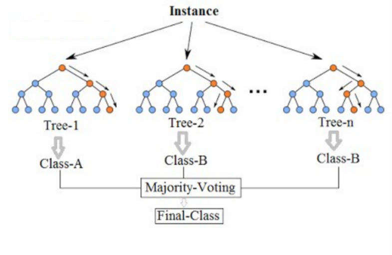
Counterfactual explanation for $\boldsymbol{x}^0$ to be classified in class k^+

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

where

- $f_k : \mathbb{R}^J \to \mathbb{R}$ is the **score function** of classifier $\mathcal{M}$ for class $k = 1, \ldots, K$
- $\mathcal{X}^0 \subset \mathbb{R}^J$ actionability and plausibility constraints
  - polyhedron with some integer coordinates
- a cost function $C(\cdot, \cdot) : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$
  - $\ell_0, \ell_1, \ell_2, \ldots$

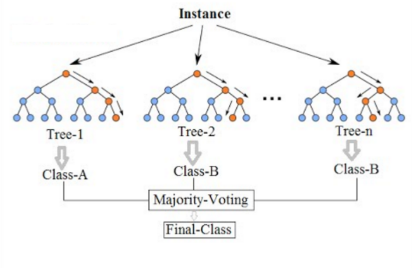
The score-based classifier is an Additive Tree Model



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, $s \in \text{Left}(l, t) \cup \text{Right}(l, t)$
- $\mathcal{L}^t_k$ subset of leaves in t whose output is class $k = 1, \ldots, K$

The score-based classifier is an Additive Tree Model



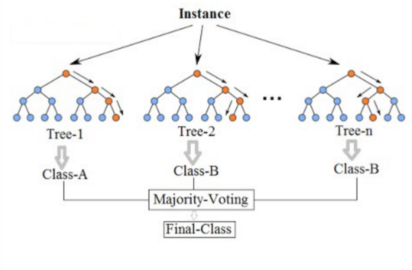
Decision variables

- $\boldsymbol{x} \in \mathbb{R}^J$ 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

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

The score-based classifier is an Additive Tree Model



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

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

$C(\boldsymbol{x}, \boldsymbol{x^0}) = \lambda_0\, \ell_0(\boldsymbol{x} - \boldsymbol{x^0}) + \lambda_2\, \ell_2^2(\boldsymbol{x} - \boldsymbol{x^0})$

- An equivalent Mixed Integer Convex Quadratic Model with linear constraints
- If $\lambda_2 = 0$, an equivalent MILP formulation

Counterfactual Explanations for ATM models

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

$C(\boldsymbol{x}, \boldsymbol{x^0}) = \lambda_0 \, \ell_0(\boldsymbol{x} - \boldsymbol{x^0}) + \lambda_2 \, \ell_2^2(\boldsymbol{x} - \boldsymbol{x^0})$

- An equivalent Mixed Integer Convex Quadratic Model with linear constraints
- If $\lambda_2 = 0$, an equivalent MILP formulation

Counterfactual Explanations for ATM models

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

$C(\boldsymbol{x}, \boldsymbol{x^0}) = \lambda_0\, \ell_0(\boldsymbol{x} - \boldsymbol{x^0}) + \lambda_2\, \ell_2^2(\boldsymbol{x} - \boldsymbol{x^0})$

- An equivalent Mixed Integer Convex Quadratic Model with linear constraints
- If $\lambda_2 = 0$, an equivalent MILP formulation

Counterfactual Explanations for ATM models

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

$C(\boldsymbol{x}, \boldsymbol{x^0}) = \lambda_0\, \ell_0(\boldsymbol{x} - \boldsymbol{x^0}) + \lambda_2\, \ell_2^2(\boldsymbol{x} - \boldsymbol{x^0})$

- An equivalent Mixed Integer Convex Quadratic Model with linear constraints
- If $\lambda_2 = 0$, an equivalent MILP formulation

Numerical illustration for `housing` dataset

Counterfactual explanation for 1 instance

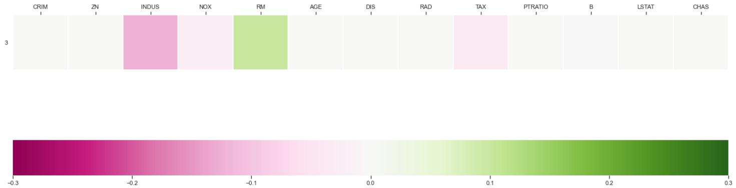


Figure: Random Forest for the `housing` dataset: To $k^+ = +1$ with $C = 0.01\ell_0 + \ell_2^2$

Numerical illustration for `housing` dataset

Counterfactual explanation for 10 instances

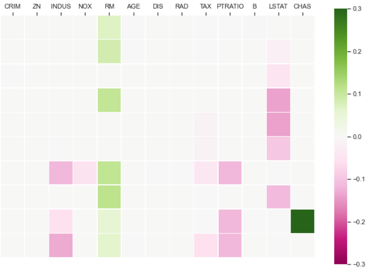


Figure: Random Forest for the `housing` dataset: To $k^+ = +1$ with $C = 0.01\ell_0 + \ell_2^2$

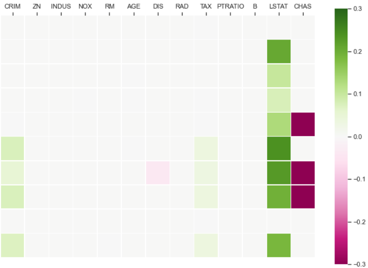


Figure: Random Forest for the `housing` dataset: To $k^+ = -1$ with $C = 0.01\ell_0 + \ell_2^2$

Counterfactual explanations for a collective of individuals

Counterfactual explanations for a collective of individuals

- A collective of individuals, **each of them requires a counterfactual explanation**
- If the problem is separable, then use the single-instance model (in previous slides)
- The problem is not separable, e.g., when controlling ℓ_0^{global}, i.e., the sparsity across all counterfactual explanations

  useful for the modeler to detect important features to classifier $\mathcal{M}$

- **If the problem is not separable**, we propose a novel formulation in which the linking constraints are modeled

Counterfactual explanations for a collective of individuals

Counterfactual explanations for a collective of individuals

- A collective of individuals, **each of them requires a counterfactual explanation**
- If the problem is separable, then use the single-instance model (in previous slides)
- The problem is not separable, e.g., when controlling ℓ_0^{global}, i.e., the sparsity across all counterfactual explanations

  useful for the modeler to detect important features to classifier $\mathcal{M}$

- **If the problem is not separable**, we propose a novel formulation in which the linking constraints are modeled

Counterfactual explanations for a collective of individuals

Counterfactual explanations for a collective of individuals

- A collective of individuals, **each of them requires a counterfactual explanation**
- If the problem is separable, then use the single-instance model (in previous slides)
- The problem is not separable, e.g., when controlling ℓ_0^{global}, i.e., the sparsity across all counterfactual explanations
  - **useful for the modeler to detect important features to classifier $\mathcal{M}$**
- **If the problem is not separable**, we propose a novel formulation in which the linking constraints are modeled

Counterfactual explanations for a collective of individuals

Counterfactual explanations for a collective of individuals

- A collective of individuals, **each of them requires a counterfactual explanation**
- If the problem is separable, then use the single-instance model (in previous slides)
- The problem is not separable, e.g., when controlling ℓ_0^{global}, i.e., the sparsity across all counterfactual explanations
    - **useful for the modeler to detect important features to classifier $\mathcal{M}$**
- **If the problem is not separable**, we propose a novel formulation in which the linking constraints are modeled

Numerical illustration for `housing` dataset

Counterfactual explanation for 10 instances

SEPARABLE

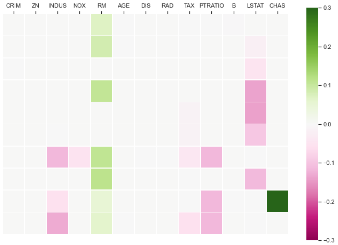


Figure: Random Forest for the `housing` dataset: To $k^+ = +1$ with $C = 0.01\ell_0 + \ell_2^2$, separable case

NON-SEPARABLE

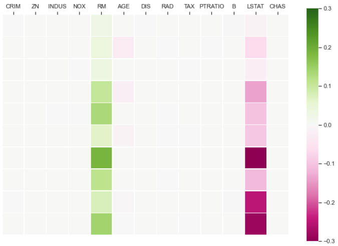
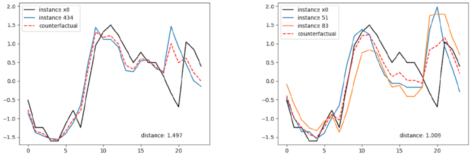


Figure: Random Forest for the `housing` dataset: To $k^+ = +1$ with $C = 0.1\ell_0^{\text{global}} + \ell_2^2$, non-separable case

Functional data and counterfactual explanations

- Counterfactual explanations: convex combinations of prototypes
- Cost: $\lambda_0 \ell_0 + \lambda_{\mathrm{DTW}}$DTW, where DTW stands for Dynamic Time Warping

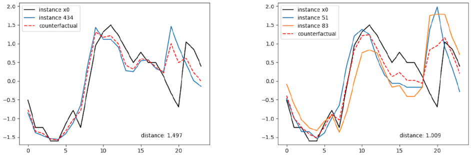


(a) $B^{\max} = 1$ (b) $B^{\max} = 2$

Figure: Random Forest for the `ItalyPowerDemand` dataset: To $k^+ = +1$ with $C =$ DTW. Different values of $B^{\max}$, i.e., the number of prototypes used for the convex combination, have been imposed.

Functional data and counterfactual explanations

- Counterfactual explanations: convex combinations of prototypes
- Cost: $\lambda_0 \ell_0 + \lambda_{\mathrm{DTW}}\mathrm{DTW}$, where DTW stands for Dynamic Time Warping



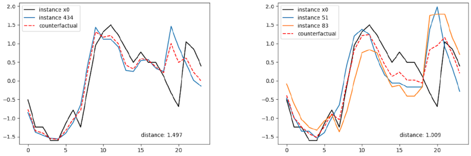
(a) $B^{\max} = 1$

(b) $B^{\max} = 2$

Figure: Random Forest for the `ItalyPowerDemand` dataset: To $k^{+} = +1$ with $C = \mathrm{DTW}$. Different values of $B^{\max}$, i.e., the number of prototypes used for the convex combination, have been imposed.

Functional data and counterfactual explanations

- Counterfactual explanations: convex combinations of prototypes
- Cost: $\lambda_0 \ell_0 + \lambda_{\mathrm{DTW}}$DTW, where DTW stands for Dynamic Time Warping

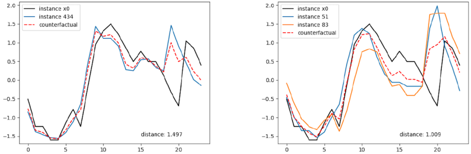


(a) $B^{\max} = 1$ (b) $B^{\max} = 2$

Figure: Random Forest for the `ItalyPowerDemand` dataset: To $k^+ = +1$ with $C =$ DTW. Different values of $B^{\max}$, i.e., the number of prototypes used for the convex combination, have been imposed.

Functional data and counterfactual explanations

- Counterfactual explanations: convex combinations of prototypes
- Cost: $\lambda_0 \ell_0 + \lambda_{\mathrm{DTW}}\mathrm{DTW}$, where DTW stands for Dynamic Time Warping



(a) $B^{\max} = 1$ (b) $B^{\max} = 2$

Figure: Random Forest for the `ItalyPowerDemand` dataset: To $k^{+} = +1$ with $C = \mathrm{DTW}$. Different values of $B^{\max}$, i.e., the number of prototypes used for the convex combination, have been imposed.

Outline

Some thoughts

- Transparency by design that can model the loss in accuracy
- Counterfactual explanations to understand the machine learning model
- Counterfactual explanations to understand decision making models

You are kindly invited to

imUS VIRTUAL SEMINAR

SEMINAR SERIES

Machine Learning NeEDS

Mathematical Optimization

Organizers: Emilio Carrizosa, Thomas Halskov, Kseniia Kurishchenko, Cristina Molero del Río, Jasone Ramírez-Ayerbe and Dolores Romero Morales

https://congreso.us.es/mlneedsmo/

Universidad de Sevilla · CBS Copenhagen Business School Handelshøjskolen · imus instituto de matemáticas universidad de sevilla · NeEDS · Unidad IMAT

Thank you very much!

E: drm.eco@cbs.dk **H**: doloresromero.com **T**: @DoloresRomeroM

RG: https://www.researchgate.net/profile/Dolores-Romero-Morales

References I

S. Aghaei, M.J. Azizi, and P. Vayanos. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 1418–1426, 2019.

S. Aghaei, A. Gomez, and P. Vayanos. Learning optimal classification trees: Strong max-flow formulations. *arXiv preprint arXiv:2002.09142*, 2020.

A. Atamtürk and A. Gomez. Rank-one convexification for sparse regression. *arXiv preprint arXiv:1901.10334*, 2019.

B. Baesens, R. Setiono, C. Mues, and J. Vanthienen. Using neural network rule extraction and decision tables for credit-risk evaluation. *Management Science*, 49(3):312–329, 2003.

C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. SIRUS: making random forests interpretable. *arXiv preprint arXiv:1908.06852*, 2019.

S. Benítez-Peña, R. Blanquero, E. Carrizosa, and P. Ramírez-Cobo. Cost-sensitive feature selection for Support Vector Machines. *Computers & Operations Research*, 106:169–178, 2019.

S. Benítez-Peña, P. Bogetoft, and D. Romero Morales. Feature selection in data envelopment analysis: A mathematical optimization approach. *Omega*, 96: 102068, 2020.

S. Benítez-Peña, E. Carrizosa, V. Guerrero, M. Dolores Jiménez-Gamero, B. Martín-Barragán, C. Molero-Río, P. Ramírez-Cobo, D. Romero Morales, and M. Remedios Sillero-Denamiel. On sparse ensemble methods: An application to short-term predictions of the evolution of COVID-19. *European Journal of Operational Research*, 295(2):648–663, 2021.

S. Benítez-Peña, P. Bogetoft, and D. Romero Morales. Joint clustering and feature selection in data envelopment analysis. Technical report, Copenhagen Business School, Denmark, `https://www.researchgate.net/publication/363113389_Joint_Clustering_and_Feature_Selection_in_Data_Envelopment_Analysis`, 2022.

D. Bertsimas and J. Dunn. Optimal classification trees. *Machine Learning*, 106(7):1039–1082, 2017.

D. Bertsimas, A. King, and R. Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.

Philippe Besse, Eustasio del Barrio, Paula Gordaliza, Jean-Michel Loubes, and Laurent Risser. A survey of bias in machine learning through the prism of statistical parity. *The American Statistician*, 76(2):188–198, 2022.

R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. Sparsity in optimal randomized classification trees. *European Journal of Operational Research*, 284(1):255 – 272, 2020.

R. Blanquero, E. Carrizosa, C. Molero-Río, and Romero Morales. Optimal randomized classification trees. *Computers and Operations Research*, 132: 105281, 2021a.

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

R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. On sparse optimal regression trees. *European Journal of Operational Research*, 299(3): 1045–1054, 2022a.

References II

R. Blanquero, E. Carrizosa, C. Molero-Río, and D. Romero Morales. On optimal regression trees to detect critical intervals for multivariate functional data. Technical report, IMUS, Sevilla, Spain, https://www.researchgate.net/publication/360396613_On_optimal_regression_trees_to_detect_critical_intervals_for_multivariate_functional_data, 2022b.

E. Carrizosa and D. Romero Morales. Supervised classification and mathematical optimization. *Computers and Operations Research*, 40(1):150–165, 2013.

E. Carrizosa, B. Martín-Barragán, D. Romero Morales, and F. Plastria. On the selection of the globally optimal prototype subset for nearest-neighbor classification. *INFORMS Journal on Computing*, 19(3):470–479, 2007.

E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Binarized support vector machines. *INFORMS Journal on Computing*, 22(1):154–167, 2010.

E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Detecting relevant variables and interactions in supervised classification. *European Journal of Operational Research*, 213(1):260–269, 2011.

E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Strongly agree or strongly disagree?: Rating features in Support Vector Machines. *Information Sciences*, 329:256–273, 2016.

E. Carrizosa, A. Nogales-Gómez, and D. Romero Morales. Clustering categories in support vector machines. *Omega*, 66:28–37, 2017.

E. Carrizosa, M.J. Ramírez Ayerbe, and D. Romero Morales. Generating collective counterfactual explanations in score-based classification via mathematical optimization. Technical report, IMUS, Sevilla, Spain, https://www.researchgate.net/publication/353073138_Generating_Collective_Counterfactual_Explanations_in_Score-Based_Classification_via_Mathematical_Optimization, 2021a.

E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. A binarization approach to model interactions between categorical predictors in generalized linear models. Technical report, Copenhagen Business School, Denmark, https://www.researchgate.net/publication/350755054_A_binarization_approach_to_model_interactions_between_categorical_predictors_in_Generalized_Linear_Models, 2021b.

E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. On clustering categories of categorical predictors in generalized linear models. *Experts Systems With Applications*, 182:115245, 2021c.

E. Carrizosa, K. Kurishchenko, A. Marín, and D. Romero Morales. On clustering and interpreting with rules by means of mathematical optimization. Technical report, Copenhagen Business School, Denmark, https://www.researchgate.net/publication/354208780_On_Clustering_and_Interpreting_with_Rules_by_Means_of_Mathematical_Optimization/stats, 2021d.

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

E. Carrizosa, M.J. Ramírez Ayerbe, and D. Romero Morales. A new model for counterfactual analysis for functional data. Technical report, IMUS, Sevilla, Spain, https://www.researchgate.net/publication/363539291_A_New_Model_for_Counterfactual_Analysis_for_Functional_Data, 2022a.

References III

E. Carrizosa, M. Galvis Restrepo, and D. Romero Morales. Improving the fairness of linear models in supervised classification by feature shrinkage. Technical report, Copenhagen Business School, Denmark, `https://www.researchgate.net/publication/358614960_Improving_the_fairness_of_linear_models_in_supervised_classification_by_feature_shrinkage`, 2022b.

E. Carrizosa, V. Guerrero, and D. Romero Morales. On mathematical optimization for clustering categories in contingency tables. *Forthcoming in Advances in Data Analysis and Classification*, 2022c.

E. Carrizosa, K. Kurishchenko, A. Marín, and D. Romero Morales. Interpreting clusters by prototype optimization. *Omega*, 107:102543, 2022d.

E. Carrizosa, K. Kurishchenko, and D. Romero Morales. On sparse random forests. Technical report, Copenhagen Business School, Denmark, In preparation, 2022e.

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

Y. Chevaleyre, F. Koriche, and J.-D. Zucker. Rounding methods for discrete linear classification. In S. Dasgupta and D. McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, volume 28, pages 651–659. JMLR Workshop and Conference Proceedings, 2013.

S. Dash, O. Günlük, and D. Wei. Boolean decision rules via column generation. In *Advances in Neural Information Processing Systems*, pages 4655–4665, 2018.

E. Demirović, A. Lukina, E. Hebrard, J. Chan, J. Bailey, C. Leckie, K.i Ramamohanarao, and P.J. Stuckey. Murtree: Optimal decision trees via dynamic programming and search. *Journal of Machine Learning Research*, 23(26):1–47, 2022.

F. D'Onofrio, G. Grani, M. Monaci, and L. Palagi. Margin optimal classification trees. *https://arxiv.org/abs/2210.10567*, 2022.

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

M. Firat, G. Crognier, A.F. Gabor, C.A.J. Hurkens, and Y. Zhang. Column generation based heuristic for learning classification trees. *Computers & Operations Research*, 116:104866, 2020.

Alexandre Forel, Axel Parmentier, and Thibaut Vidal. Robust counterfactual explanations for random forests, 2022. URL `https://arxiv.org/abs/2205.14116`.

K. Fountoulakis and J. Gondzio. A second-order method for strongly convex ℓ_1-regularization problems. *Mathematical Programming*, 156(1):189–219, 2016.

A.A. Freitas. Comprehensible classification models: a position paper. *ACM SIGKDD Explorations Newsletter*, 15(1):1–10, 2014.

A. Ghorbani and J. Zou. Neuron Shapley: Discovering the Responsible Neurons. *arXiv preprint arXiv:2002.09815*, 2020.

M. Golea and M. Marchand. On learning perceptrons with binary weights. *Neural Computation*, 5(5):767–782, 1993.

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

References IV

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. *Forthcoming in Data Mining and Knowledge Discovery*, pages 1–55, 2022.

O. Günlük, J. Kalagnanam, M. Li, M. Menickelly, and K. Scheinberg. Optimal Decision Trees for Categorical Data via Integer Programming. *Journal of Global Optimization*, 81:233–260, 2021.

D. Gunning and D.W. Aha. DARPA's Explainable Artificial Intelligence Program. *AI Magazine*, 40(2):44–58, 2019.

P.E. Hart. The condensed nearest neighbor rule. *IEEE Transactions on Information Theory*, 14:515–516, 1968.

S. Holter, O. Gomez, and E. Bertini. *FICO Explainable Machine Learning Challenge*. `https://community.fico.com/s/explainable-machine-learning-challenge`, 2018.

A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: definitions, formulations, solutions, and prospects. *arXiv preprint arXiv:2010.04050*, 2020.

A. Kenney, F. Chiaromonte, and G. Felici. MIP-BOOST: Efficient and effective l_0 feature selection for linear regression. *Journal of Computational and Graphical Statistics*, 30(3):566–577, 2021.

C. Lawless, J. Kalagnanam, L.M Nguyen, D. Phan, and C. Reddy. Interpretable clustering via multi-polytope machines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7309–7316, 2022.

O. Li, H. Liu, C. Chen, and C. Rudin. Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions, 2017.

S.M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1):2522–5839, 2020.

S. Maldonado, J. Pérez, R. Weber, and M. Labbé. Feature selection for support vector machines via mixed integer linear programming. *Information Sciences*, 279:163–175, 2014.

Donato Maragno, Tabea E Röber, and Ilker Birbil. Counterfactual explanations using optimization with constraint learning. *arXiv preprint arXiv:2209.10997*, 2022.

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

N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54:1–35, 2022.

T. Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019.

N. Narodytska, A. Ignatiev, F. Pereira, and J. Marques-Silva. Learning Optimal Decision Trees with SAT. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 1362–1368, 2018.

References V

C. Orsenigo and C. Vercellis. Multivariate classification trees based on minimum features discrete support vector machines. *IMA Journal of Management Mathematics*, 14(3):221–234, 2003.

C. Orsenigo and C. Vercellis. Discrete support vector decision trees via tabu search. *Computational Statistics and Data Analysis*, 47(2):311–322, 2004.

M.T. Ribeiro, S. Singh, and C. 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.

G. Ridgeway. The pitfalls of prediction. *National Institute of Justice Journal*, 271:34–40, 2013.

F. Rinaldi and M. Sciandrone. Feature selection combining linear support vector machines and concave optimization. *Optimization Methods and Software*, 25(1):117–128, 2010.

F. Rinaldi, F. Schoen, and M. Sciandrone. Concave programming for minimizing the zero-norm over polyhedral sets. *Computational Optimization and Applications*, 46(3):467–486, 2010.

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

B. Ustun and C. Rudin. Supersparse linear integer models for optimized medical scoring systems. *Machine Learning*, 102(3):349–391, 2016.

H. Verhaeghe, S. Nijssen, G. Pesant, C.-G. Quimper, and P. Schaus. Learning optimal decision trees using constraint programming. In *The 25th International Conference on Principles and Practice of Constraint Programming (CP2019)*, 2019.

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

G. Wilfong. Nearest neighbor problems. *International Journal of Computational Geometry and Applications*, 2(4):383–416, 1992.

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, 2017a.

M.B. Zafar, I. Valera, M. Gomez Rodriguez, and K.P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pages 1171–1180, 2017b.