Collective LIME: Enhancing the explainability of the explainer

Dolores Romero Morales, Copenhagen Business School

The 5th EUROYoung Workshop
Naples, October 16, 2025

EURO Young
THE ASSOCIATION OF
EUROPEAN OPERATIONAL
RESEARCH SOCIETIES

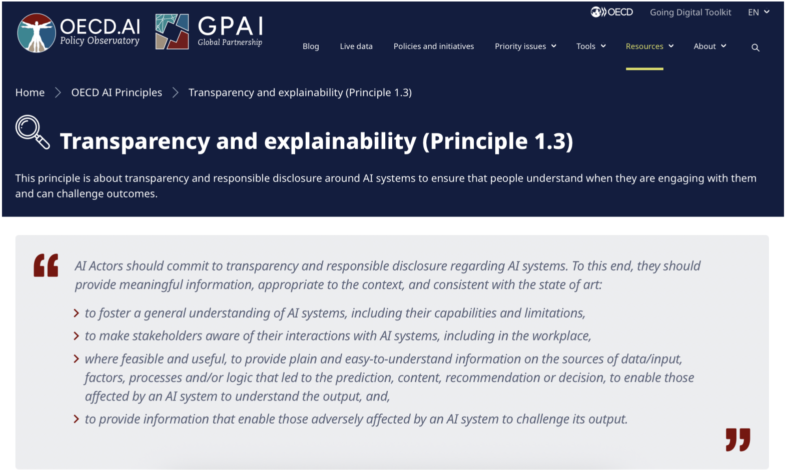
Thank you to the work of the **EUROYoung** Forum!

Outline

- Introduction
- The CLIME methodology
- Conclusions

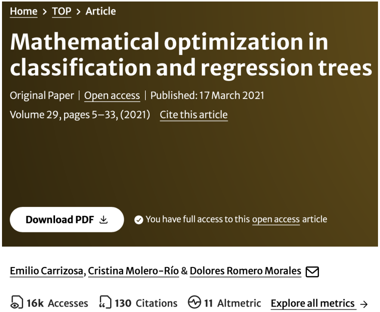
Introduction

When training a machine learning model, **accuracy** of its predictions matters, as does its **transparency** (Baesens et al., 2003; Panigutti et al., 2023; Rudin et al., 2022)



Introduction

Generally, linear models (e.g., medical scoring systems (Ustun and Rudin, 2016)) are considered to be easy-to-understand, as well as rule-based models (Carrizosa et al., 2021b)



Home > TOP > Article

Mathematical optimization in classification and regression trees

Original Paper | Open access | Published: 17 March 2021

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

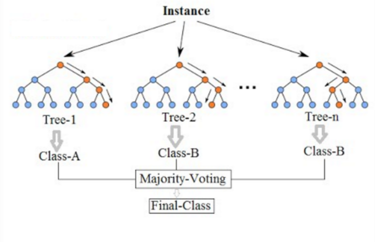
Download PDF

You have full access to this open access article

Emilio Carrizosa, Cristina Molero-Río & Dolores Romero Morales

16k Accesses 130 Citations 11 Altmetric Explore all metrics →

More complex models such as Random Forests (Breiman, 2001) XGBoost (Chen and Guestrin, 2016) and Deep Learning (Goodfellow et al., 2016) are seen as **black boxes**



Introduction

- Even linear models lose transparency when their **complexity** increases, as measured, e.g., by # features. Classic methodologies such as LASSO (Tibshirani, 1996) or Best Subset selection (Bertsimas et al., 2016; Hazimeh and Mazumder, 2020) aim at selecting a small **subset of features** that give a good accuracy
- Similarly, rule-based models lose transparency when the # rules increases (Carrizosa et al., 2025b)

Introduction

- Even linear models lose transparency when their **complexity** increases, as measured, e.g., by # features. Classic methodologies such as LASSO (Tibshirani, 1996) or Best Subset selection (Bertsimas et al., 2016; Hazimeh and Mazumder, 2020) aim at selecting a small **subset of features** that give a good accuracy
- Similarly, rule-based models lose transparency when the # rules increases (Carrizosa et al., 2025b)

Sparsity in Generalized Linear Models for categorical data in Carrizosa et al. (2021a)

An example of a categorical feature

BE BG CZ DK DE EE SE


ELSEVIER
Expert Systems with Applications
Volume 182, 15 November 2021, 115245

On clustering categories of categorical predictors in generalized linear models

Emilio Carrizosa [a], Marcela Galvis Restrepo [b], Dolores Romero Morales [b]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.eswa.2021.115245 Get rights and content

Enhancing the transparency of (generalized) linear models

Sparsity in Generalized Linear Models for categorical data in Carrizosa et al. (2021a)

Clustering categories into red and blue clusters

ELSEVIER

Expert Systems with Applications
Volume 182, 15 November 2021, 115245

On clustering categories of categorical predictors in generalized linear models

Emilio Carrizosa [a], Marcela Galvis Restrepo [b], Dolores Romero Morales [b]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.eswa.2021.115245

Get rights and content

Sparsity in Generalized Linear Models for categorical data in Carrizosa et al. (2021a)

yields a sparser representation of the variable


ELSEVIER
Expert Systems with Applications
Volume 182, 15 November 2021, 115245

On clustering categories of categorical predictors in generalized linear models

Emilio Carrizosa [a], Marcela Galvis Restrepo [b], Dolores Romero Morales [b]

Show more

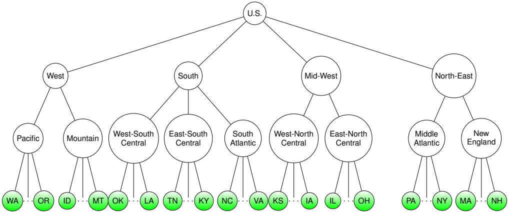
Add to Mendeley Share Cite

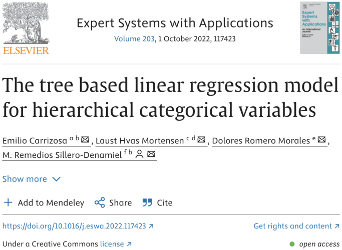
https://doi.org/10.1016/j.eswa.2021.115245

Get rights and content

Sparsity for Linear Models for hierarchical data in Carrizosa et al. (2022b)

An example of a hierarchical categorical feature




ELSEVIER

Expert Systems with Applications
Volume 203, 1 October 2022, 117423

The tree based linear regression model for hierarchical categorical variables

Emilio Carrizosa [a b], Laust Hvas Mortensen [c d], Dolores Romero Morales [e], M. Remedios Sillero-Denamiel [f b]

Show more

+ Add to Mendeley Share Cite

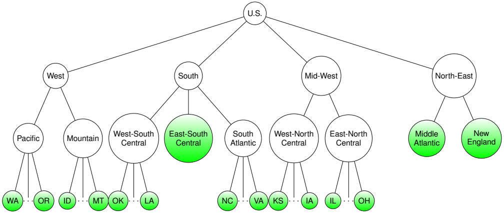
https://doi.org/10.1016/j.eswa.2022.117423 Get rights and content

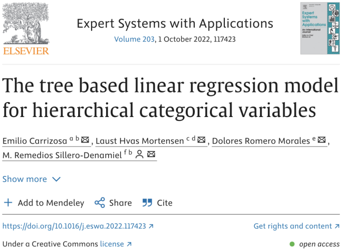
Under a Creative Commons license open access

Enhancing the transparency of (generalized) linear models

Sparsity for Linear Models for hierarchical data in Carrizosa et al. (2022b)

Clustering cat. yields a sparser representation





ELSEVIER

Expert Systems with Applications

Volume 203, 1 October 2022, 117423

The tree based linear regression model for hierarchical categorical variables

Emilio Carrizosa [a b], Laust Hvas Mortensen [c d], Dolores Romero Morales [e], M. Remedios Sillero-Denamiel [f b]

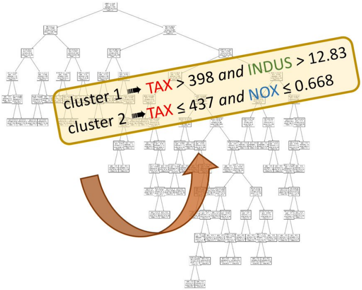
Show more

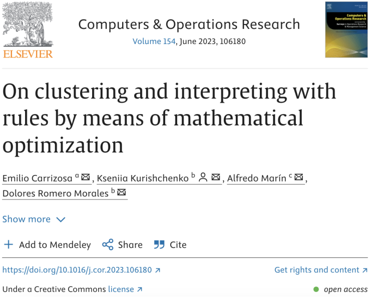
Add to Mendeley Share Cite

https://doi.org/10.1016/j.eswa.2022.117423

Get rights and content

If-then rules to explain clusters in Carrizosa et al. (2023)




Computers & Operations Research
Volume 154, June 2023, 106180

On clustering and interpreting with rules by means of mathematical optimization

Emilio Carrizosa [a], Kseniia Kurishchenko [b], Alfredo Marín [c], Dolores Romero Morales [b]

Show more

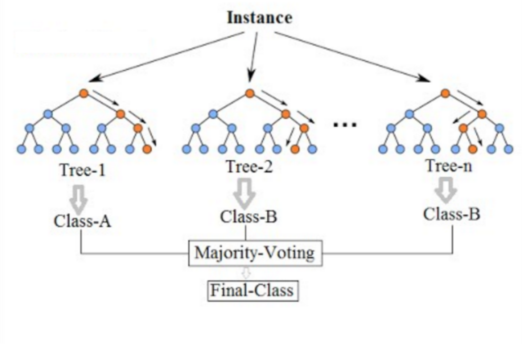
Add to Mendeley Share Cite

https://doi.org/10.1016/j.cor.2023.106180 Get rights and content

Under a Creative Commons license open access

Building surrogate interpretable models from opaque ones[1]

A surrogate optimal tree from a tree ensemble in Piccialli et al. (2024)




ELSEVIER

European Journal of Operational Research

Volume 317, Issue 2, 1 September 2024, Pages 273-285

Supervised feature compression based on counterfactual analysis

Veronica Piccialli [a], Dolores Romero Morales [b], Cecilia Salvatore [c]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.ejor.2023.11.019

Get rights and content

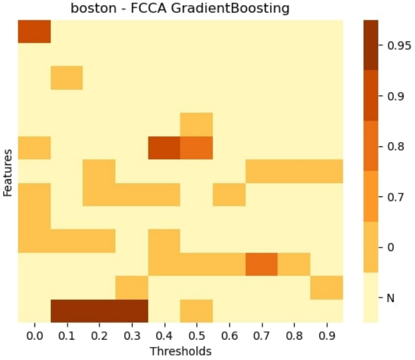
Under a Creative Commons license

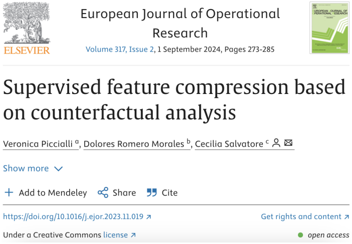
open access

[1] Bénard et al. (2021); Carrizosa et al. (2010, 2011, 2016, 2017, 2022a); Chevaleyre et al. (2013); Di Teodoro et al. (2024); Emine et al. (2024); Golea and Marchand (1993); Li et al. (2017); Martens et al. (2007); Piccialli et al. (2024); Vidal and Schiffer (2020)

Building surrogate interpretable models from opaque ones

A surrogate optimal tree from a tree ensemble in Piccialli et al. (2024)




ELSEVIER
European Journal of Operational Research
Volume 317, Issue 2, 1 September 2024, Pages 273-285

Supervised feature compression based on counterfactual analysis

Veronica Piccialli [a], Dolores Romero Morales [b], Cecilia Salvatore [c]

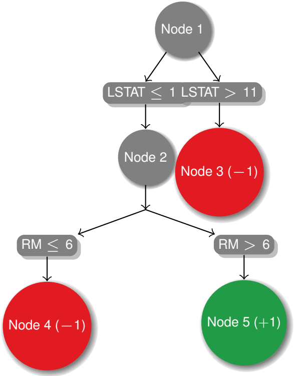
Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.ejor.2023.11.019 Get rights and content

Building surrogate interpretable models from opaque ones

A surrogate optimal tree from a tree ensemble in Piccialli et al. (2024)



ELSEVIER
European Journal of Operational Research
Volume 317, Issue 2, 1 September 2024, Pages 273-285

Supervised feature compression based on counterfactual analysis

Veronica Piccialli [a], Dolores Romero Morales [b], Cecilia Salvatore [c]

Show more

Add to Mendeley Share Cite

https://doi.org/10.1016/j.ejor.2023.11.019

Get rights and content

open access

Growing number of Explainable AI tools

Intelligent machines are asked to explain how their minds work

Researchers tackle problem that threatens to hold up adoption of advanced AI

A deep-learning machine on display at an artificial intelligence trade show in Tokyo last month © EPA

Richard Waters in San Francisco JULY 10 2017 53

Researchers at Parc, a laboratory with links to some of Silicon Valley's biggest breakthroughs, have just taken on a particularly thorny challenge: teaching intelligent machines to explain, in human terms, how their minds work.

The project, one of several sponsored by the US Defense Advanced Research Projects Agency (Darpa), is part of the search for an answer to one of the hardest problems in artificial intelligence.

Given an opaque model, explain it locally

Given an ML model $\mathcal{M}$, explain the prediction made for $\mathbf{x}_0$

XAI approaches (Bach et al., 2015; Burkart and Huber, 2021; Goldstein et al., 2015) can be grouped into finding

- the **important variables** in the prediction of $\mathcal{M}$ for $\mathbf{x}_0$, or
- a **surrogate and interpretable** model, with high fidelity to $\mathcal{M}$ around $\mathbf{x}_0$, or
- a **counterfactual instance** to $\mathbf{x}_0$ but with a desired prediction by $\mathcal{M}$

- Plethora of feature importance metrics, including SHapley Additive exPlanations (aka, SHAP) Lundberg and Lee (2017)
- Local Interpretable Model-Agnostic Explanations (aka, LIME) Ribeiro et al. (2016)
- Counterfactual Explanations Carrizosa et al. (2024); Martens and Provost (2014); Wachter et al. (2017)

Given an opaque model, explain it locally

Given an ML model $\mathcal{M}$, explain the prediction made for $\mathbf{x}_0$

XAI approaches (Bach et al., 2015; Burkart and Huber, 2021; Goldstein et al., 2015) can be grouped into finding

- the **important variables** in the prediction of $\mathcal{M}$ for $\mathbf{x}_0$, or
- a **surrogate and interpretable** model, with high fidelity to $\mathcal{M}$ around $\mathbf{x}_0$, or
- a **counterfactual instance** to $\mathbf{x}_0$ but with a desired prediction by $\mathcal{M}$

- Plethora of feature importance metrics, including SHapley Additive exPlanations (aka, SHAP) Lundberg and Lee (2017)
- Local Interpretable Model-Agnostic Explanations (aka, LIME) Ribeiro et al. (2016)
- Counterfactual Explanations Carrizosa et al. (2024); Martens and Provost (2014); Wachter et al. (2017)

Given an opaque model, explain it locally

Given an ML model $\mathcal{M}$, explain the prediction made for $\mathbf{x}_0$

XAI approaches (Bach et al., 2015; Burkart and Huber, 2021; Goldstein et al., 2015) can be grouped into finding

- the **important variables** in the prediction of $\mathcal{M}$ for $\mathbf{x}_0$, or
  - Plethora of feature importance metrics, including SHapley Additive exPlanations (aka, SHAP) Lundberg and Lee (2017)
- a **surrogate and interpretable** model, with high fidelity to $\mathcal{M}$ around $\mathbf{x}_0$, or
  - Local Interpretable Model-Agnostic Explanations (aka, LIME) Ribeiro et al. (2016)
- a **counterfactual instance** to $\mathbf{x}_0$ but with a desired prediction by $\mathcal{M}$
  - Counterfactual Explanations Carrizosa et al. (2024); Martens and Provost (2014); Wachter et al. (2017)

Given an opaque model, explain it locally

Given an ML model $\mathcal{M}$, explain the prediction made for $\mathbf{x}_0$

XAI approaches (Bach et al., 2015; Burkart and Huber, 2021; Goldstein et al., 2015) can be grouped into finding

- the **important variables** in the prediction of $\mathcal{M}$ for $\mathbf{x}_0$, or
- a **surrogate and interpretable** model, with high fidelity to $\mathcal{M}$ around $\mathbf{x}_0$, or
- a **counterfactual instance** to $\mathbf{x}_0$ but with a desired prediction by $\mathcal{M}$

- Plethora of feature importance metrics, including SHapley Additive exPlanations (aka, SHAP) Lundberg and Lee (2017)
- Local Interpretable Model-Agnostic Explanations (aka, LIME) Ribeiro et al. (2016)
- Counterfactual Explanations Carrizosa et al. (2024); Martens and Provost (2014); Wachter et al. (2017)

Given an opaque model, explain it locally

Given an ML model $\mathcal{M}$, explain the prediction made for $\mathbf{x}_0$

XAI approaches (Bach et al., 2015; Burkart and Huber, 2021; Goldstein et al., 2015) can be grouped into finding

- the **important variables** in the prediction of $\mathcal{M}$ for $\mathbf{x}_0$, or
- a **surrogate and interpretable** model, with high fidelity to $\mathcal{M}$ around $\mathbf{x}_0$, or
- a **counterfactual instance** to $\mathbf{x}_0$ but with a desired prediction by $\mathcal{M}$

- Plethora of feature importance metrics, including SHapley Additive exPlanations (aka, SHAP) Lundberg and Lee (2017)
- Local Interpretable Model-Agnostic Explanations (aka, LIME) Ribeiro et al. (2016)
- Counterfactual Explanations Carrizosa et al. (2024); Martens and Provost (2014); Wachter et al. (2017)

LIME in a nutshell

Given model $\mathcal{M}$ and $\mathbf{x}_0$, LIME works as follows (Ribeiro et al., 2016)

- Construct perturbations around $\mathbf{x}_0$, generating a set of instances
- $\mathcal{M}$ is used to get the response for each instance
- Each instance is weighted according to their proximity to $\mathbf{x}_0$
- An interpretable surrogate model, usually a linear one, is fitted

LIME in a nutshell

Given model $\mathcal{M}$ and $\mathbf{x}_0$, LIME works as follows (Ribeiro et al., 2016)

- Construct perturbations around $\mathbf{x}_0$, generating a set of instances
- $\mathcal{M}$ is used to get the response for each instance
- Each instance is weighted according to their proximity to $\mathbf{x}_0$
- An interpretable surrogate model, usually a linear one, is fitted

LIME in a nutshell

Given model $\mathcal{M}$ and $\mathbf{x}_0$, LIME works as follows (Ribeiro et al., 2016)

- Construct perturbations around $\mathbf{x}_0$, generating a set of instances
- $\mathcal{M}$ is used to get the response for each instance
- Each instance is weighted according to their proximity to $\mathbf{x}_0$
- An interpretable surrogate model, usually a linear one, is fitted

LIME in a nutshell

Given model $\mathcal{M}$ and $\mathbf{x}_0$, LIME works as follows (Ribeiro et al., 2016)

- Construct perturbations around $\mathbf{x}_0$, generating a set of instances
- $\mathcal{M}$ is used to get the response for each instance
- Each instance is weighted according to their proximity to $\mathbf{x}_0$
- An interpretable surrogate model, usually a linear one, is fitted

LIME in a nutshell

Given model $\mathcal{M}$ and $\mathbf{x}_0$, LIME works as follows (Ribeiro et al., 2016)

- Construct perturbations around $\mathbf{x}_0$, generating a set of instances
- $\mathcal{M}$ is used to get the response for each instance
- Each instance is weighted according to their proximity to $\mathbf{x}_0$
- An interpretable surrogate model, usually a linear one, is fitted

Outline

- Introduction
- The CLIME methodology
- Conclusions

Collective Local Interpretable Model-Agnostic Explanations

We propose CLIME (Carrizosa et al., 2025a)

- Today, a collective framework for LIME, hereafter **Collective Local Interpretable Model-Agnostic Explanations (CLIME)**
- For a **collection** of instances, we build a surrogate model around each of them that is interpretable and locally accurate
- Our collective framework enables control over global properties of the explanations such as **global feature selection**, i.e., across the surrogate models

Collective Local Interpretable Model-Agnostic Explanations

We propose CLIME (Carrizosa et al., 2025a)

- Today, a collective framework for LIME, hereafter **Collective Local Interpretable Model-Agnostic Explanations (CLIME)**
- For a **collection** of instances, we build a surrogate model around each of them that is interpretable and locally accurate
- Our collective framework enables control over global properties of the explanations such as **global feature selection**, i.e., across the surrogate models

Collective Local Interpretable Model-Agnostic Explanations

We propose CLIME (Carrizosa et al., 2025a)

- Today, a collective framework for LIME, hereafter **Collective Local Interpretable Model-Agnostic Explanations (CLIME)**
- For a **collection** of instances, we build a surrogate model around each of them that is interpretable and locally accurate
- Our collective framework enables control over global properties of the explanations such as **global feature selection**, i.e., across the surrogate models

The surrogate model (and the decision) in CLIME

- Feature space $\mathcal{X} \subseteq \mathbb{R}^p$, and response space $\mathcal{Y} \subseteq \mathbb{R}$
- The prediction function $y : \mathbb{R}^p \to \mathcal{Y}$ associated with model $\mathcal{M}$
- The prediction function associated with surrogate model $\hat{y}$ that explains y

Generalized Linear Models (GLMs) as interpretable surrogate

- **Generalized Linear Models (GLMs)** (Nelder and Wedderburn, 1972) as interpretable surrogate models
- The simplest case is a linear model with vector of **coefficients** $\beta(\mathbf{x}) \in \mathbb{R}^p$

$$\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

so that $\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) \approx y(\tilde{\mathbf{x}})$. The fidelity of $\hat{y}$ to y around $\mathbf{x}$ can be measured, e.g., by

$$(\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) - y(\tilde{\mathbf{x}}))^2$$

- The **decision in CLIME** is $\beta : \mathbb{R}^p \to \mathbb{R}^p$ yielding the coefficients $\beta(\mathbf{x})$ at instance $\mathbf{x}$

The surrogate model (and the decision) in CLIME

- Feature space $\mathcal{X} \subseteq \mathbb{R}^p$, and response space $\mathcal{Y} \subseteq \mathbb{R}$
- The prediction function $y : \mathbb{R}^p \to \mathcal{Y}$ associated with model $\mathcal{M}$
- The prediction function associated with surrogate model $\hat{y}$ that explains y

Generalized Linear Models (GLMs) as interpretable surrogate

- **Generalized Linear Models (GLMs)** (Nelder and Wedderburn, 1972) as interpretable surrogate models
- The simplest case is a linear model with vector of **coefficients** $\beta(\mathbf{x}) \in \mathbb{R}^p$

$$\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

so that $\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) \approx y(\tilde{\mathbf{x}})$. The fidelity of $\hat{y}$ to y around $\mathbf{x}$ can be measured, e.g., by

$$(\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) - y(\tilde{\mathbf{x}}))^2$$

- The **decision in CLIME** is $\beta : \mathbb{R}^p \to \mathbb{R}^p$ yielding the coefficients $\beta(\mathbf{x})$ at instance $\mathbf{x}$

The surrogate model (and the decision) in CLIME

- Feature space $\mathcal{X} \subseteq \mathbb{R}^p$, and response space $\mathcal{Y} \subseteq \mathbb{R}$
- The prediction function $y : \mathbb{R}^p \to \mathcal{Y}$ associated with model $\mathcal{M}$
- The prediction function associated with surrogate model $\hat{y}$ that explains y

Generalized Linear Models (GLMs) as interpretable surrogate

- **Generalized Linear Models (GLMs)** (Nelder and Wedderburn, 1972) as interpretable surrogate models
- The simplest case is a linear model with vector of **coefficients** $\beta(\mathbf{x}) \in \mathbb{R}^p$

$$\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

so that $\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) \approx y(\tilde{\mathbf{x}})$. The fidelity of $\hat{y}$ to y around $\mathbf{x}$ can be measured, e.g., by

$$(\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) - y(\tilde{\mathbf{x}}))^2$$

- The **decision in CLIME** is $\beta : \mathbb{R}^p \to \mathbb{R}^p$ yielding the coefficients $\beta(\mathbf{x})$ at instance $\mathbf{x}$

The surrogate model (and the decision) in CLIME

- Feature space $\mathcal{X} \subseteq \mathbb{R}^p$, and response space $\mathcal{Y} \subseteq \mathbb{R}$
- The prediction function $y : \mathbb{R}^p \to \mathcal{Y}$ associated with model $\mathcal{M}$
- The prediction function associated with surrogate model $\hat{y}$ that explains y

Generalized Linear Models (GLMs) as interpretable surrogate

- **Generalized Linear Models (GLMs)** (Nelder and Wedderburn, 1972) as interpretable surrogate models
- The simplest case is a linear model with vector of **coefficients** $\beta(\mathbf{x}) \in \mathbb{R}^p$

$$\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

so that $\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) \approx y(\tilde{\mathbf{x}})$. The fidelity of $\hat{y}$ to y around $\mathbf{x}$ can be measured, e.g., by

$$(\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) - y(\tilde{\mathbf{x}}))^2$$

- The **decision in CLIME** is $\beta : \mathbb{R}^p \to \mathbb{R}^p$ yielding the coefficients $\beta(\mathbf{x})$ at instance $\mathbf{x}$

The surrogate model (and the decision) in CLIME

- Feature space $\mathcal{X} \subseteq \mathbb{R}^p$, and response space $\mathcal{Y} \subseteq \mathbb{R}$
- The prediction function $y : \mathbb{R}^p \to \mathcal{Y}$ associated with model $\mathcal{M}$
- The prediction function associated with surrogate model $\hat{y}$ that explains y

Generalized Linear Models (GLMs) as interpretable surrogate

- **Generalized Linear Models (GLMs)** (Nelder and Wedderburn, 1972) as interpretable surrogate models
- The simplest case is a linear model with vector of **coefficients** $\beta(\mathbf{x}) \in \mathbb{R}^p$

$$\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

  so that $\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) \approx y(\tilde{\mathbf{x}})$. The fidelity of $\hat{y}$ to y around $\mathbf{x}$ can be measured, e.g., by

$$(\hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}) - y(\tilde{\mathbf{x}}))^2$$

- The **decision in CLIME** is $\beta : \mathbb{R}^p \to \mathbb{R}^p$ yielding the coefficients $\beta(\mathbf{x})$ at instance $\mathbf{x}$

A visualization of the output of CLIME

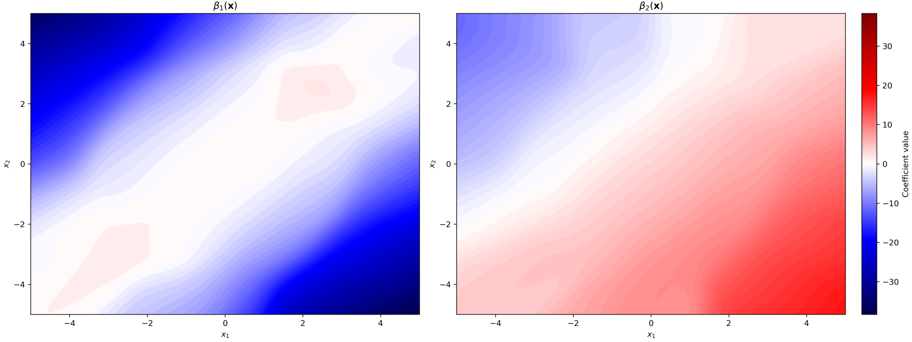


Figure: Visualizing $\beta_1(\mathbf{x})$ and $\beta_2(\mathbf{x})$ for a toy dataset, $y = (x_2 - x_1)^2 + 3x_1$.

The surrogate model (and the decision) in CLIME

- Although LIME mostly uses linear models independently of the nature of $\mathcal{Y}$, linear models are not suitable, for instance, in Supervised Classification
- GLMs are rich enough to deal with different types of **response** variable

$$g(\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}]) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

where g is the so-called link function. Conversely,

$$\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}] = g^{-1}(\beta(\mathbf{x})^\top \tilde{\mathbf{x}})$$

- With this, by choosing the right g and fidelity, we can handle

Linear reg	**Logistic reg**	**Poisson reg**
for continuous response	for binary response	for count data response
g identity function	g logistic function	g exponential function

The surrogate model (and the decision) in CLIME

- Although LIME mostly uses linear models independently of the nature of $\mathcal{Y}$, linear models are not suitable, for instance, in Supervised Classification
- GLMs are rich enough to deal with different types of **response** variable

$$g(\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}]) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

where g is the so-called link function. Conversely,

$$\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}] = g^{-1}(\beta(\mathbf{x})^\top \tilde{\mathbf{x}})$$

- With this, by choosing the right g and fidelity, we can handle

Linear reg	**Logistic reg**	**Poisson reg**
for continuous response	for binary response	for count data response
g identity function	g logistic function	g exponential function

The surrogate model (and the decision) in CLIME

- Although LIME mostly uses linear models independently of the nature of $\mathcal{Y}$, linear models are not suitable, for instance, in Supervised Classification
- GLMs are rich enough to deal with different types of **response** variable

$$g(\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}]) = \beta(\mathbf{x})^\top \tilde{\mathbf{x}},$$

where g is the so-called link function. Conversely,

$$\mathbb{E}[Y \mid \mathbf{X} = \tilde{\mathbf{x}}] = g^{-1}(\beta(\mathbf{x})^\top \tilde{\mathbf{x}})$$

- With this, by choosing the right g and fidelity, we can handle

Linear reg	**Logistic reg**	**Poisson reg**
for continuous response	for binary response	for count data response
g identity function	g logistic function	g exponential function

The objective function in CLIME

As in LIME, we have a *local* measure of error:

$$\delta(\beta(\mathbf{x}), \mathbf{x}) := \int \omega(\mathbf{x}, \tilde{\mathbf{x}}) \ell(y(\tilde{\mathbf{x}}), \hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}))\, d\mathbf{P}(\tilde{\mathbf{x}})$$

- Distribution $\mathbf{P}$ for ground truth of instances
  e.g., discrete uniform on a set of instances, or mixture of Normal distributions
- Weighting function $\omega : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$

$$e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2} \qquad e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2} \qquad \mathbb{1}_{\{\|\mathbf{x}-\tilde{\mathbf{x}}\| \le \epsilon\}}$$

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For linear regression $\ell(s, t) = (s - t)^2$, and for other GLMs the negative of the corresponding log-likelihood.

The objective function in CLIME

As in LIME, we have a *local* measure of error:

$$\delta(\beta(\mathbf{x}), \mathbf{x}) := \int \omega(\mathbf{x}, \tilde{\mathbf{x}}) \ell(y(\tilde{\mathbf{x}}), \hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}))\, d\mathbf{P}(\tilde{\mathbf{x}})$$

- Distribution **P** for ground truth of instances
  e.g., discrete uniform on a set of instances, or mixture of Normal distributions
- Weighting function $\omega : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$

  $$e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2} \qquad e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2} \qquad \mathbb{1}_{\{\|\mathbf{x}-\tilde{\mathbf{x}}\| \leq \epsilon\}}$$

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For linear regression $\ell(s, t) = (s - t)^2$, and for other GLMs the negative of the corresponding log-likelihood.

The objective function in CLIME

As in LIME, we have a *local* measure of error:

$$\delta(\beta(\mathbf{x}), \mathbf{x}) := \int \omega(\mathbf{x}, \tilde{\mathbf{x}})\ell(y(\tilde{\mathbf{x}}), \hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}))\, d\mathbf{P}(\tilde{\mathbf{x}})$$

- Distribution **P** for ground truth of instances
  e.g., discrete uniform on a set of instances, or mixture of Normal distributions
- Weighting function $\omega : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$

$$e^{-\gamma\|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2} \qquad e^{-\gamma\|\mathbf{x}-\tilde{\mathbf{x}}\|_2} \qquad \mathbb{1}_{\{\|\mathbf{x}-\tilde{\mathbf{x}}\|\le\epsilon\}}$$

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For linear regression $\ell(s, t) = (s - t)^2$, and for other GLMs the negative of the corresponding log-likelihood.

The objective function in CLIME

As in LIME, we have a *local* measure of error:

$$\delta(\beta(\mathbf{x}), \mathbf{x}) := \int \omega(\mathbf{x}, \tilde{\mathbf{x}}) \ell(y(\tilde{\mathbf{x}}), \hat{y}(\beta(\mathbf{x}), \tilde{\mathbf{x}}))\, d\mathbf{P}(\tilde{\mathbf{x}})$$

- Distribution **P** for ground truth of instances
  e.g., discrete uniform on a set of instances, or mixture of Normal distributions
- Weighting function $\omega : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}_+$

  $$e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2^2} \qquad e^{-\gamma \|\mathbf{x}-\tilde{\mathbf{x}}\|_2} \qquad \mathbb{1}_{\{\|\mathbf{x}-\tilde{\mathbf{x}}\| \leq \epsilon\}}$$

- Loss function $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$. For linear regression $\ell(s, t) = (s - t)^2$, and for other GLMs the negative of the corresponding log-likelihood.

The objective function in CLIME

The objective function in CLIME measures the error incurred *globally*

$$\Delta(\beta) := \int \delta(\beta(\mathbf{x}), \mathbf{x}) \, d\mathbf{Q}(\mathbf{x}),$$

where **Q** is the distribution controlling the relevance of each explanation

The optimization model behind CLIME

The global optimization problem associated with CLIME reads as follows

$$\min_{\beta \in \mathcal{B}} \Delta(\beta),$$

where $\mathcal{B}$ is the feasible region for β

The objective function in CLIME

The objective function in CLIME measures the error incurred *globally*

$$\Delta(\beta) := \int \delta(\beta(\mathbf{x}), \mathbf{x}) \, d\mathbf{Q}(\mathbf{x}),$$

where **Q** is the distribution controlling the relevance of each explanation

The optimization model behind CLIME

The global optimization problem associated with CLIME reads as follows

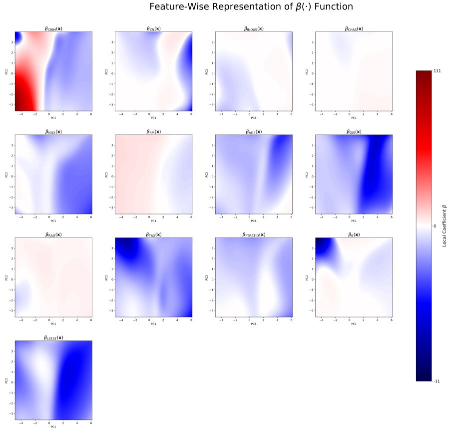
$$\min_{\beta \in \mathcal{B}} \Delta(\beta),$$

where $\mathcal{B}$ is the feasible region for β

Theoretical properties of CLIME

Our framework addresses some drawbacks of LIME pointed in the literature (Garreau and von Luxburg, 2020; Sepulveda et al., 2025; Tiukhova et al., 2024; Zhou and Wang, 2021). Indeed, if $\mathcal{B}$ consists of all functions, the following proposition can be shown:

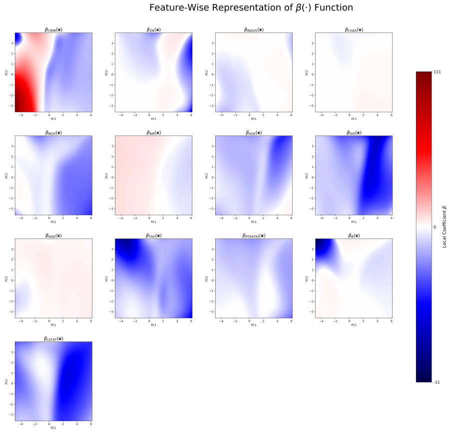
Proposition Under some technical conditions, if ω is continuous (resp., Lipschitz continuous) then the optimal solution of CLIME is continuous (resp., Lipschitz continuous)



Theoretical properties of CLIME

Our framework addresses some drawbacks of LIME pointed in the literature (Garreau and von Luxburg, 2020; Sepulveda et al., 2025; Tiukhova et al., 2024; Zhou and Wang, 2021). Indeed, if $\mathcal{B}$ consists of all functions, the following proposition can be shown:

Proposition Under some technical conditions, if ω is continuous (resp., Lipschitz continuous) then the optimal solution of CLIME is continuous (resp., Lipschitz continuous)



Cost-Sensitive Feature Selection

Our methodology is designed to be able to control global properties

Model $\mathcal{M}$ may have used all p features, but, with a **stakeholder** point of view, we would like to know whether there are surrogate models using fewer features with a good error. For this, we now impose **global sparsity** to CLIME

This means that we would like to choose a subset of features $\mathcal{J} \subseteq \{1, \ldots, p\}$, such that the surrogate models can only use features in $\mathcal{J}$

Cost-Sensitive Feature Selection

Our methodology is designed to be able to control global properties

Model $\mathcal{M}$ may have used all p features, but, with a **stakeholder** point of view, we would like to know whether there are surrogate models using fewer features with a good error. For this, we now impose **global sparsity** to CLIME

This means that we would like to choose a subset of features $\mathcal{J} \subseteq \{1, \ldots, p\}$, such that the surrogate models can only use features in $\mathcal{J}$

Cost-Sensitive Feature Selection

Our methodology is designed to be able to control global properties

Model $\mathcal{M}$ may have used all p features, but, with a **stakeholder** point of view, we would like to know whether there are surrogate models using fewer features with a good error. For this, we now impose **global sparsity** to CLIME

This means that we would like to choose a subset of features $\mathcal{J} \subseteq \{1, \ldots, p\}$, such that the surrogate models can only use features in $\mathcal{J}$

The Cost-Sensitive Feature Selection Problem

The objective function of the Feature Selection Problem reads as follows $\pi(\mathcal{J}) := \min_{\beta_{\mathcal{J}}} \Delta_{\mathcal{J}}(\beta_{\mathcal{J}})$, with

- $\beta_{\mathcal{J}}$, $\mathbf{x}_{\mathcal{J}}$, $\tilde{\mathbf{x}}_{\mathcal{J}}$ be the respective projections of β, $\mathbf{x}$, $\tilde{\mathbf{x}}$
- $\delta_{\mathcal{J}}(\beta_{\mathcal{J}}(\mathbf{x}_{\mathcal{J}}), \mathbf{x}_{\mathcal{J}}) := \int \omega_{\mathcal{J}}(\mathbf{x}_{\mathcal{J}}, \tilde{\mathbf{x}}_{\mathcal{J}}) \ell(y(\tilde{\mathbf{x}}), \hat{y}_{\mathcal{J}}(\beta_{\mathcal{J}}(\mathbf{x}_{\mathcal{J}}), \tilde{\mathbf{x}}_{\mathcal{J}})) d\mathbf{P}(\tilde{\mathbf{x}})$
- $\Delta_{\mathcal{J}}(\beta_{\mathcal{J}}) := \int \delta(\beta_{\mathcal{J}}(\mathbf{x}_{\mathcal{J}}), \mathbf{x}_{\mathcal{J}}) \, d\mathbf{Q}(\mathbf{x}_{\mathcal{J}})$

Since the feature selection is cost-sensitive:

- Let c_j be the cost assigned to feature j (Carrizosa et al., 2008; Turney, 1995)
- Let C be the total budget

$$\boxed{\begin{array}{ll} \min\limits_{\mathcal{J}} & \pi(\mathcal{J}) \\ \text{s.t.} & \sum\limits_{j \in \mathcal{J}} c_j \leq C \end{array}}$$

Solution approach

Assume that we have a solution $\mathcal{J}^{\text{cr}}$ for the cost-sensitive feature selection problem. For instance, $\mathcal{J}^{\text{cr}} = \emptyset$, $\mathcal{J}^{\text{cr}} = \{1, \ldots, p\}$ or $\mathcal{J}^{\text{cr}}$ obtained with a greedy approach such as in stepwise feature selection procedures (Hastie et al., 2009)

Let V_j^{cr} be a measure of the **contribution of feature j to π** around $\mathcal{J}^{\text{cr}}$. The linearization at $\mathcal{J} = \mathcal{J}^{\text{cr}}$ yields the following Knapsack Problem

$$\begin{aligned} \max_{\mathcal{J}} \quad & \sum_{j \in \mathcal{J}} V_j^{\text{cr}} \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}} c_j \leq C \end{aligned}$$

How to measure V_j^{cr}

$$V_j^{\text{cr}} := \begin{cases} \pi(\mathcal{J}^{\text{cr}}) - \pi(\mathcal{J}^{\text{cr}} \cup \{j\}), & j \notin \mathcal{J}^{\text{cr}}, \\ \pi(\mathcal{J}^{\text{cr}} \setminus \{j\}) - \pi(\mathcal{J}^{\text{cr}}), & j \in \mathcal{J}^{\text{cr}} \end{cases}$$

Algorithm Iterative Algorithm For Feature Selection

Input:
  Initial feature set $\mathcal{J}^{\text{start}}$
  Total budget C
Output: Selected feature set $\mathcal{J}^{\text{inc}}$
Initialize:
Incumbent feature set $\mathcal{J}^{\text{inc}} \leftarrow \mathcal{J}^{\text{start}}$
Current feature set $\mathcal{J}^{\text{cr}} \leftarrow \mathcal{J}^{\text{inc}}$
while stopping criteria not met **do**
  for $j = 1$ **to** p **do**
    Calculate V_j^{cr}
  end for
  Solve the Knapsack Problem with V_j^{cr}, for all j
  Update feature set $\mathcal{J}^{\text{cr}}$ to its optimal solution
  if $\pi(\mathcal{J}^{\text{cr}}) < \pi(\mathcal{J}^{\text{inc}})$ **then**
    $\mathcal{J}^{\text{inc}} \leftarrow \mathcal{J}^{\text{cr}}$ ▷ Update incumbent solution
  end if
end while
return $\mathcal{J}^{\text{inc}}$ ▷ Return the best set found

Solution approach

Assume that we have a solution $\mathcal{J}^{\text{cr}}$ for the cost-sensitive feature selection problem. For instance, $\mathcal{J}^{\text{cr}} = \emptyset$, $\mathcal{J}^{\text{cr}} = \{1, \ldots, p\}$ or $\mathcal{J}^{\text{cr}}$ obtained with a greedy approach such as in stepwise feature selection procedures (Hastie et al., 2009)

Let V_j^{cr} be a measure of the **contribution of feature j to π** around $\mathcal{J}^{\text{cr}}$. The linearization at $\mathcal{J} = \mathcal{J}^{\text{cr}}$ yields the following Knapsack Problem

$$\max_{\mathcal{J}} \quad \sum_{j \in \mathcal{J}} V_j^{\text{cr}}$$
$$\text{s.t.} \quad \sum_{j \in \mathcal{J}} c_j \leq C$$

How to measure V_j^{cr}

$$V_j^{\text{cr}} := \begin{cases} \pi(\mathcal{J}^{\text{cr}}) - \pi(\mathcal{J}^{\text{cr}} \cup \{j\}), & j \notin \mathcal{J}^{\text{cr}}, \\ \pi(\mathcal{J}^{\text{cr}} \setminus \{j\}) - \pi(\mathcal{J}^{\text{cr}}), & j \in \mathcal{J}^{\text{cr}} \end{cases}$$

Algorithm Iterative Algorithm For Feature Selection

1: **Input:**
 Initial feature set $\mathcal{J}^{\text{start}}$
 Total budget C
2: **Output:** Selected feature set $\mathcal{J}^{\text{inc}}$
3: **Initialize:**
4: Incumbent feature set $\mathcal{J}^{\text{inc}} \leftarrow \mathcal{J}^{\text{start}}$
5: Current feature set $\mathcal{J}^{\text{cr}} \leftarrow \mathcal{J}^{\text{inc}}$
6: **while** stopping criteria not met **do**
7: **for** $j = 1$ **to** p **do**
8:   Calculate V_j^{cr}
9: **end for**
10: Solve the Knapsack Problem with V_j^{cr}, for all j
11: Update feature set $\mathcal{J}^{\text{cr}}$ to its optimal solution
12: **if** $\pi(\mathcal{J}^{\text{cr}}) < \pi(\mathcal{J}^{\text{inc}})$ **then**
13:   $\mathcal{J}^{\text{inc}} \leftarrow \mathcal{J}^{\text{cr}}$ ▷ Update incumbent solution
14: **end if**
15: **end while**
16: **return** $\mathcal{J}^{\text{inc}}$ ▷ Return the best set found

Solution approach

Assume that we have a solution $\mathcal{J}^{\text{cr}}$ for the cost-sensitive feature selection problem. For instance, $\mathcal{J}^{\text{cr}} = \emptyset$, $\mathcal{J}^{\text{cr}} = \{1, \ldots, p\}$ or $\mathcal{J}^{\text{cr}}$ obtained with a greedy approach such as in stepwise feature selection procedures (Hastie et al., 2009)

Let V_j^{cr} be a measure of the **contribution of feature j to π** around $\mathcal{J}^{\text{cr}}$. The linearization at $\mathcal{J} = \mathcal{J}^{\text{cr}}$ yields the following Knapsack Problem

$$\boxed{\begin{aligned} \max_{\mathcal{J}} \quad & \sum_{j \in \mathcal{J}} V_j^{\text{cr}} \\ \text{s.t.} \quad & \sum_{j \in \mathcal{J}} c_j \leq C \end{aligned}}$$

How to measure V_j^{cr}

$$V_j^{\text{cr}} := \begin{cases} \pi(\mathcal{J}^{\text{cr}}) - \pi(\mathcal{J}^{\text{cr}} \cup \{j\}), & j \notin \mathcal{J}^{\text{cr}}, \\ \pi(\mathcal{J}^{\text{cr}} \setminus \{j\}) - \pi(\mathcal{J}^{\text{cr}}), & j \in \mathcal{J}^{\text{cr}} \end{cases}$$

Algorithm Iterative Algorithm For Feature Selection

```
Input:
    Initial feature set J^start
    Total budget C
Output: Selected feature set J^inc
Initialize:
Incumbent feature set J^inc ← J^start
Current feature set J^cr ← J^inc
while stopping criteria not met do
    for j = 1 to p do
        Calculate V_j^cr
    end for
    Solve the Knapsack Problem with V_j^cr, for all j
    Update feature set J^cr to its optimal solution
    if π(J^cr) < π(J^inc) then
        J^inc ← J^cr                      ▷ Update incumbent solution
    end if
end while
return J^inc                              ▷ Return the best set found
```

Solution approach

Assume that we have a solution $\mathcal{J}^{\mathrm{cr}}$ for the cost-sensitive feature selection problem. For instance, $\mathcal{J}^{\mathrm{cr}} = \emptyset$, $\mathcal{J}^{\mathrm{cr}} = \{1, \ldots, p\}$ or $\mathcal{J}^{\mathrm{cr}}$ obtained with a greedy approach such as in stepwise feature selection procedures (Hastie et al., 2009)

Let V_j^{cr} be a measure of the **contribution of feature j to π** around $\mathcal{J}^{\mathrm{cr}}$. The linearization at $\mathcal{J} = \mathcal{J}^{\mathrm{cr}}$ yields the following Knapsack Problem

$$
\begin{aligned}
\max_{\mathcal{J}} \quad & \sum_{j \in \mathcal{J}} V_j^{\mathrm{cr}} \\
\text{s.t.} \quad & \sum_{j \in \mathcal{J}} c_j \leq C
\end{aligned}
$$

How to measure V_j^{cr}

$$
V_j^{\mathrm{cr}} := \begin{cases} \pi(\mathcal{J}^{\mathrm{cr}}) - \pi(\mathcal{J}^{\mathrm{cr}} \cup \{j\}), & j \notin \mathcal{J}^{\mathrm{cr}}, \\ \pi(\mathcal{J}^{\mathrm{cr}} \setminus \{j\}) - \pi(\mathcal{J}^{\mathrm{cr}}), & j \in \mathcal{J}^{\mathrm{cr}} \end{cases}
$$

Algorithm Iterative Algorithm For Feature Selection

Input:
  Initial feature set $\mathcal{J}^{\mathrm{start}}$
  Total budget C
Output: Selected feature set $\mathcal{J}^{\mathrm{inc}}$
Initialize:
Incumbent feature set $\mathcal{J}^{\mathrm{inc}} \leftarrow \mathcal{J}^{\mathrm{start}}$
Current feature set $\mathcal{J}^{\mathrm{cr}} \leftarrow \mathcal{J}^{\mathrm{inc}}$
while stopping criteria not met **do**
  for $j = 1$ **to** p **do**
    Calculate V_j^{cr}
  end for
  Solve the Knapsack Problem with V_j^{cr}, for all j
  Update feature set $\mathcal{J}^{\mathrm{cr}}$ to its optimal solution
  if $\pi(\mathcal{J}^{\mathrm{cr}}) < \pi(\mathcal{J}^{\mathrm{inc}})$ **then**
    $\mathcal{J}^{\mathrm{inc}} \leftarrow \mathcal{J}^{\mathrm{cr}}$ ▷ Update incumbent solution
  end if
end while
return $\mathcal{J}^{\mathrm{inc}}$ ▷ Return the best set found

Datasets and CLIME ingredients

The ML model $\mathcal{M}$ to be explained is a Random Forest

Dataset	**Response**	**# Observations**	**# Features**	**Surrogate Model**
Boston Housing Regression	Continuous	506	13	Linear Reg
Communities and Crime	Continuous	810	103	Linear Reg
Law School	Continuous	20,800	17	Linear Reg
Boston Housing Classification	Binary	506	13	Logistic Reg
DebTrivedi	Count data	4,406	21	Poisson Reg

Table: Summary of benchmark datasets

P	**Q**	ω	ℓ
Discrete uniform on dataset	Discrete uniform on dataset	$\omega(\mathbf{x}_0, \tilde{\mathbf{x}}) = e^{-\frac{1}{2}\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_2^2}$	see surrogate model

Table: CLIME ingredients

CLIME: Linear Regression

y	**P**	**Q**	ω	$\hat{y}$
RF	Discrete uniform on dataset	Discrete uniform on dataset	$\omega(\mathbf{x}_0, \tilde{\mathbf{x}}) = e^{-\frac{1}{2}\lVert \mathbf{x}_0 - \tilde{\mathbf{x}} \rVert_2^2}$	Linear Reg

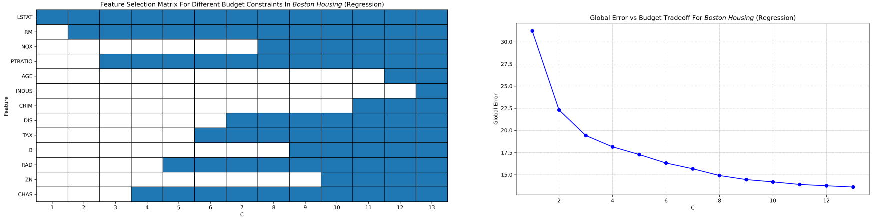
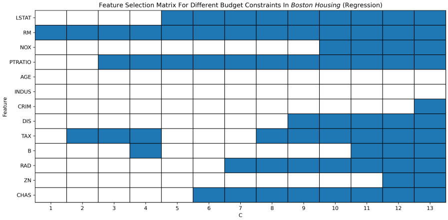


Figure: `Boston Housing` dataset (regression) with equal feature costs: Features chosen for all possible values of budget C (left), and global error vs budget (right)

CLIME: Linear Regression



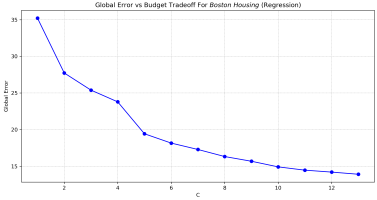


Figure: `Boston Housing` dataset (regression) with non-equal feature costs: Features chosen by Algorithm 1 for different values of budget *C* (left), and global error vs budget (right)

CLIME: Linear Regression

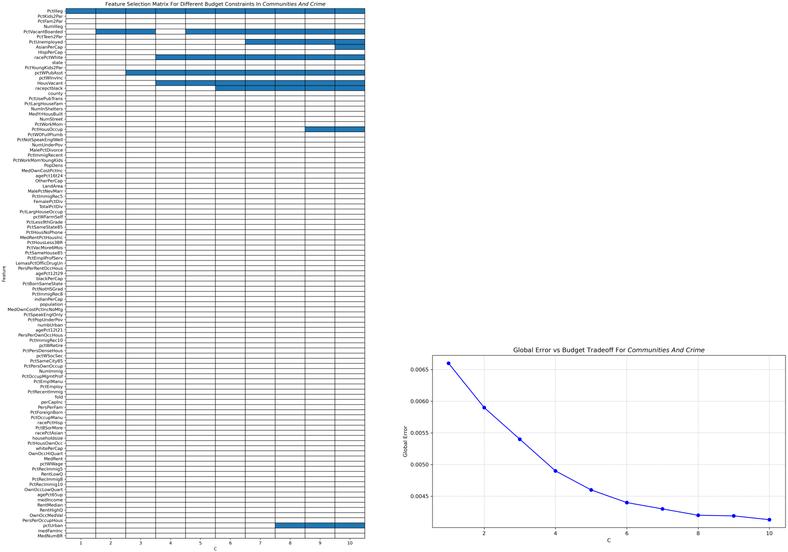
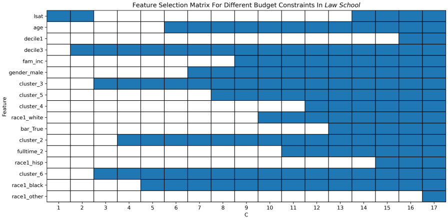


Figure: `Communities and Crime` dataset (regression) with equal feature costs: Features chosen by Algorithm 1 for budget $C \in \{1, 2, \dots, 10\}$ (left), and global error vs budget (right)

CLIME: Linear Regression



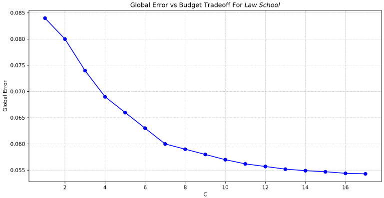
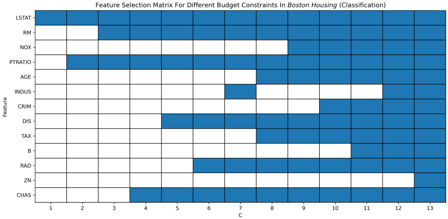


Figure: `Law School` (regression) dataset with equal feature costs: Features chosen by Algorithm 1 for all values of budget C (left), and global error vs budget (right)

CLIME: Classification

y	**P**	**Q**	ω	$\hat{y}$
RF	Discrete uniform on dataset	Discrete uniform on dataset	$\omega(\mathbf{x}_0, \tilde{\mathbf{x}}) = e^{-\frac{1}{2}\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_2^2}$	Logistic Reg



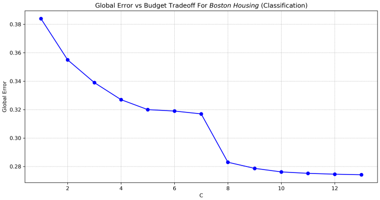
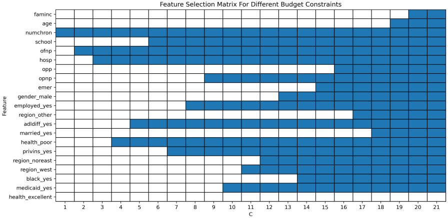


Figure: `Boston Housing` dataset (classification) with equal feature costs: Features chosen for all possible values of budget C (left), and global error vs budget (right)

CLIME: Poisson Regression

y	**P**	**Q**	ω	$\hat{y}$
RF	Discrete uniform on dataset	Discrete uniform on dataset	$\omega(\mathbf{x}_0, \tilde{\mathbf{x}}) = e^{-\frac{1}{2}\|\mathbf{x}_0 - \tilde{\mathbf{x}}\|_2^2}$	Poisson Reg



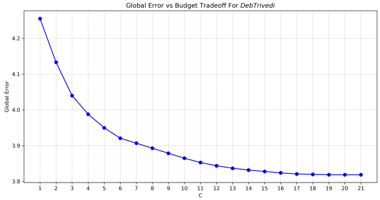


Figure: `Debtrivedi` (regression with count data) dataset with equal feature costs: Features chosen for all possible values of budget C (left), and global error vs budget (right)

Outline

- Introduction
- The CLIME methodology
- Conclusions

Conclusions

- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)

Conclusions

- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)

Conclusions

- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)

Conclusions

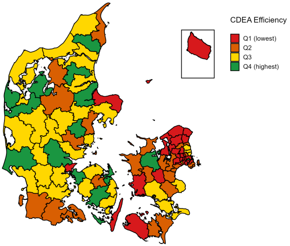
- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)

Conclusions

- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)

Conclusions

- Global optimization formulation to enhance the explainability of ML models
- With a stakeholder point of view, we control global sparsity
- Very same methodology for different types of response variables through GLMs
- Currently, dealing with complex data (Blanquero et al., 2019, 2023) and other types of explainable surrogates (Carrizosa et al., 2021b)
- Explainability in other decision-making domains (De Bock et al., 2024), such as performance benchmarking (Benítez-Peña et al., 2020; Bogetoft et al., 2024)



EURO Online Seminar Series on OR and ML

EOSS OR AND ML

EURO Online Seminar Series on OR and ML

EOSS
OR AND ML

EURO Online Seminar Series on OR and ML

YOUNG EURO OSS on Operational Research and Machine Learning

October 20, 2025, 16.30 CET

Marina Cuesta
Universidad Carlos III de Madrid,
Spain

Solène Delannoy-Pavy
Ecole Nationale des Ponts et Chaussées,
France

Qi Wang
University of Michigan,
USA

EOSS OR AND ML

To receive updates register here

For more information follow us on

EOSS OR AND ML

CBS

UNIVERSIDAD DE SEVILLA

Thank you very much!

References I

S. Bach, G. Binder, A.and Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One*, 10(7):e0130140, 2015.

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: Stable and interpretable rule set for classification. *Electronic Journal of Statistics*, 15(1):427–505, 2021.

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

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

R. Blanquero, E. Carrizosa, A. Jiménez-Cordero, and B. Martín-Barragán. Variable selection in classification for multivariate functional data. *Information Sciences*, 481:445–462, 2019.

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

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

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

N. Burkart and M.F. Huber. A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317, 2021.

E. Carrizosa, B. Martín-Barragán, and D. Romero Morales. Multi-group support vector machines with measurement costs: A biobjective approach. *Discrete Applied Mathematics*, 156:950–966, 2008.

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. Galvis Restrepo, and D. Romero Morales. On clustering categories of categorical predictors in generalized linear models. *Expert Systems With Applications*, 182:115245, 2021a.

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

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

References II

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, 2022b.

E. Carrizosa, K. Kurishchenko, A. Marín, and D. Romero Morales. On clustering and interpreting with rules by means of mathematical optimization. *Computers & Operations Research*, 154:106180, 2023.

E. Carrizosa, J. Ramírez-Ayerbe, and D. Romero Morales. Mathematical optimization modelling for group counterfactual explanations. *European Journal of Operational Research*, 319(2):399–412, 2024.

E. Carrizosa, T. Halskov, and D. Romero Morales. Collective LIME: Making black boxes explainable and sparse. Technical report, Copenhagen Business School, Denmark, https://www.researchgate.net/publication/394413617_Collective_LIME_Making_black_boxes_explainable_and_sparse, 2025a.

E. Carrizosa, K. Kurishchenko, and D. Romero Morales. On enhancing the explainability and fairness of tree ensembles. *European Journal of Operational Research*, 323(2):599–608, 2025b.

T. Chen and C. Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

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.

K.W. De Bock, K. Coussement, A. De Caigny, R. Słowiński, B. Baesens, R.N. Boute, T.-M. Choi, D. Delen, M. Kraus, S. Lessmann, S. Maldonado, D. Martens, M. Óskarsdóttir, C. Vairetti, W. Verbeke, and R. Weber. Explainable AI for Operational Research: A defining framework, methods, applications, and a research agenda. *European Journal of Operational Research*, 317(2):249–272, 2024.

G. Di Teodoro, M. Monaci, and L. Palagi. Unboxing tree ensembles for interpretability: a hierarchical visualization tool and a multivariate optimal re-built tree. *EURO Journal on Computational Optimization*, 12:100084, 2024.

Y. Emine, A. Forel, I. Malek, and T. Vidal. Free lunch in the forest: Functionally-identical pruning of boosted tree ensembles. *arXiv preprint arXiv:2408.16167*, 2024.

D. Garreau and U. von Luxburg. Explaining the Explainer: A First Theoretical Analysis of LIME. In S. Chiappa and R. Calandra, editors, *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 1287–1296, 2020.

A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *Journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

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

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, New York, 2nd edition, 2009.

References III

H. Hazimeh and R. Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5): 1517–1537, 2020.

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.

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

D. Martens, B. Baesens, T.V. Gestel, and J. Vanthienen. Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, 183(3):1466–1476, 2007.

J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 135(3):370–384, 1972.

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

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

M.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.

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.

E. Sepulveda, F. Vandervorst, B. Baesens, and T. Verdonck. Enhancing explainability in real-world scenarios: Towards a robust stability measure for local interpretability. *Expert Systems with Applications*, 274:126922, 2025.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

E. Tiukhova, P. Vemuri, N. López Flores, A.S. Islind, M. Óskarsdóttir, S. Poelmans, B. Baesens, and M. Snoeck. Explainable learning analytics: Assessing the stability of student success prediction models by means of explainable ai. *Decision Support Systems*, 182:114229, 2024.

P.D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

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

T. Vidal and M. Schiffer. Born-again tree ensembles. In *International Conference on Machine Learning*, pages 9743–9753. PMLR, 2020.

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. Zhou, Z.and Hooker and F. Wang. S-LIME: Stabilized-LIME for model explanation. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2429–2438, 2021.