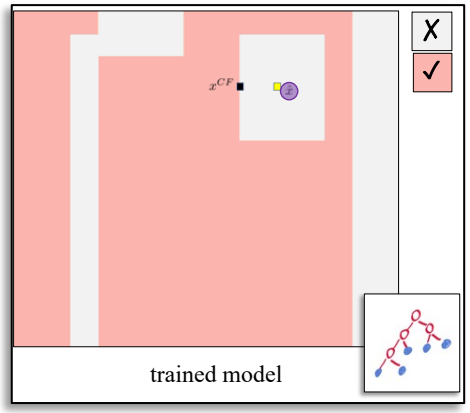
Trustworthy AI for Business and Society

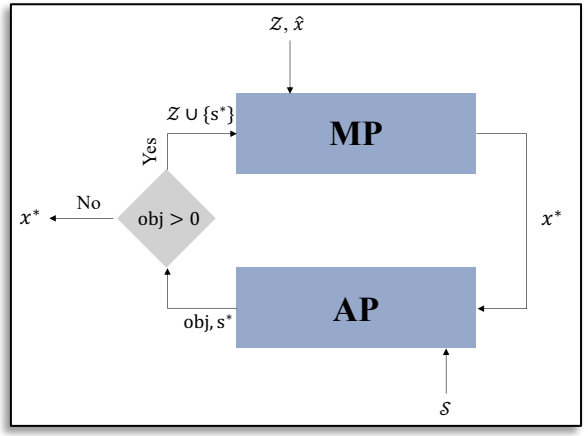
Ilker Birbil

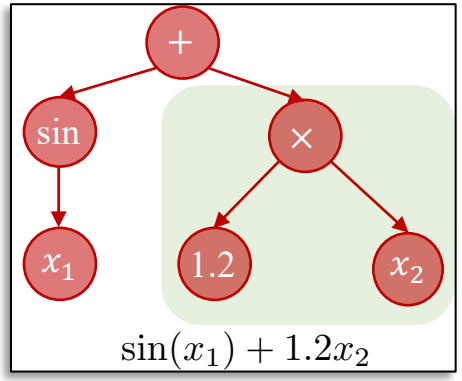
Other XAI Methods

Big Picture

- Counterfactual Explanations
  - Constraint Learning
  - Robust Optimization
- Symbolic Regression
  - Genetic Programming
  - Linear Optimization







$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{n} v_i \\
\text{subject to} \quad & w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} - v_i \leq u_{i0} && i = 1, \ldots, n, \\
& w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} + v_i \geq u_{i0} && i = 1, \ldots, n, \\
& v_i \geq 0 && i = 1, \ldots, n, \\
& w_j \in \mathbb{R} && j = 0, \ldots, p.
\end{aligned}
$$

Counterfactual Explanations

Counterfactual Explanations

COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

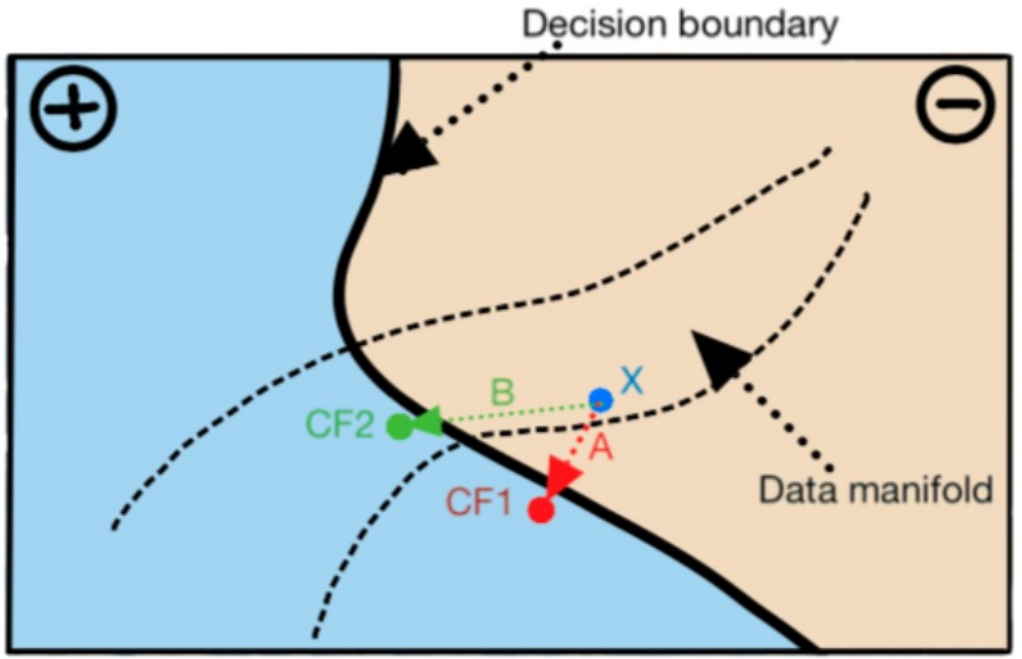
Harvard Journal of Law & Technology, 2018

Explanation in artificial intelligence: Insights from the social sciences

Tim Miller

School of Computing and Information Systems, University of Melbourne, Melbourne, Australia

Artificial Intelligence, 2019



(Verma et al., 2020)

Counterfactual Explanations for Machine Learning: A Review

Sahil Verma
University of Washington
Arthur AI
vsahil@cs.washington.edu

John Dickerson
Arthur AI
University of Maryland
john@arthur.ai

Keegan Hines
Arthur AI
keegan@arthur.ai

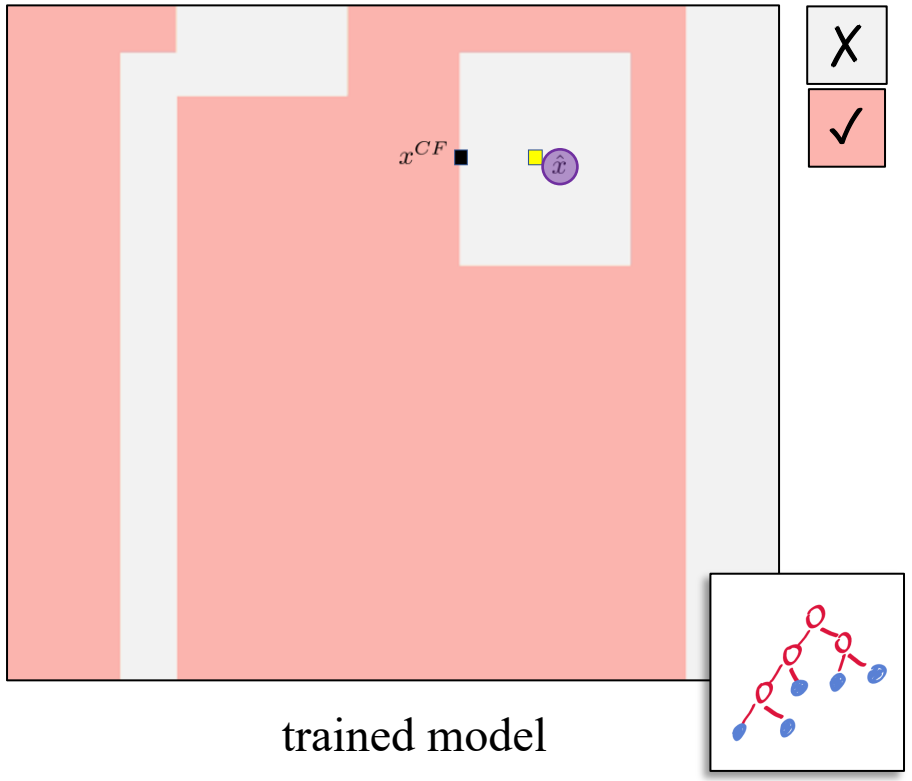
arXiv, 2020

Counterfactual Explanations

Elisabeth is 27 years old
Full-time job: 45K €/y
Account balance: 50K €

Counterfactual

Elisabeth is 27 years old
Full-time job: **50K** €/y
Account balance: **60K €**



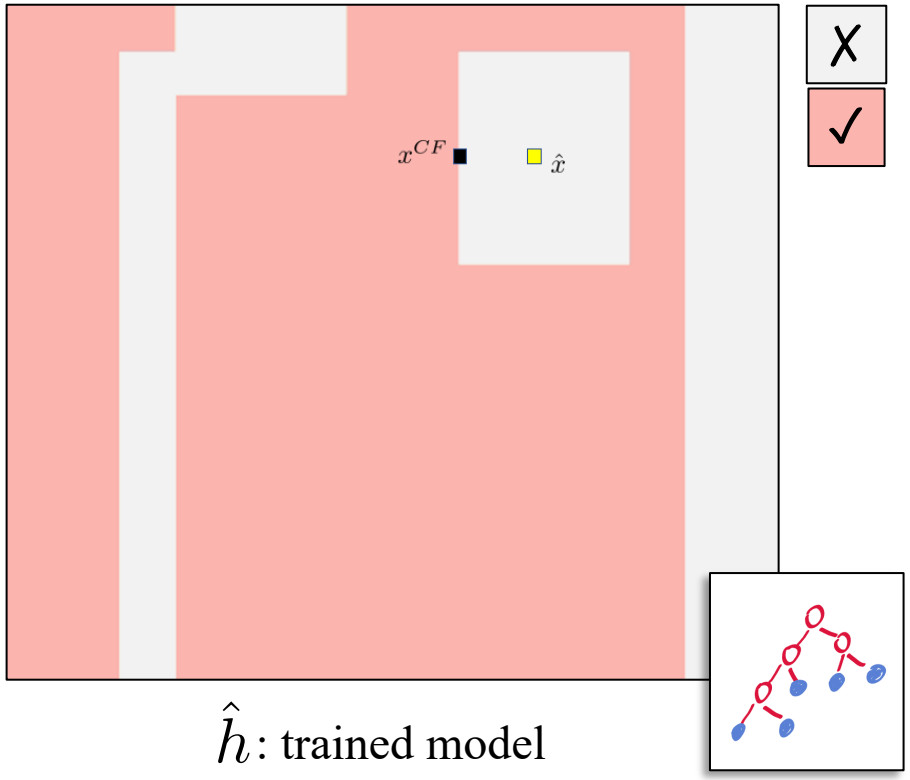
Counterfactual: set of features that should be changed to flip a model's prediction

Mathematical Model

$$\begin{aligned} \text{minimize} \quad & d(\hat{\boldsymbol{x}}, \boldsymbol{x}) \\ \text{subject to} \quad & \hat{h}(\boldsymbol{x}) \geq \tau, \\ & \boldsymbol{x} \in \mathcal{X} \end{aligned}$$

threshold to flip the decision

$$\boldsymbol{x}^{CF}$$



COUNTERFACTUAL EXPLANATIONS WITHOUT OPENING THE BLACK BOX: AUTOMATED DECISIONS AND THE GDPR

Sandra Wachter,* Brent Mittelstadt,** & Chris Russell***

Harvard Journal of Law & Technology, 2018

"Good" Counterfactual Explanations (CEs)

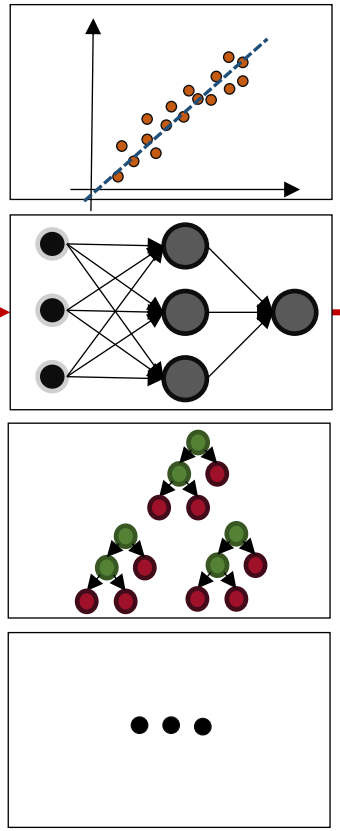
	Proximity	Sparsity	Coherence	Actionability	Data Manifold Closeness	Causality	Diversity
Russell [2019]	●	◐	●	-	-	-	●
Ustun et al. [2019]	●	●	●	●	-	-	-
Kanamori et al. [2020]	●	-	●	-	●	-	-
Mahajan et al. [2019]	●	-	●	◐	●	●	-
Karimi et al. [2021]	●	-	●	-	-	●	-
Kanamori et al. [2021]	●	●	●	●	-	●	●
Mothilal et al. [2020]	●	◐	●	●	-	◐	●
Karimi et al. [2020]	●	●	●	●	-	-	●
CE-OCL	●	●	●	●	●	●	●

●: addressed; ◐: partially addressed; –: absent

Optimization with Constraint Learning (OCL)

Maragno et al. (2021)

$$D = \{(\bar{x}_i, \bar{w}_i, \bar{y}_i)\}_{i=1}^{N}$$



$$\min_{x \in \mathbb{R}^n, y \in \mathbb{R}^k} \; f(x, w, y)$$

$$\text{s.t. } g(x, w, y) \leq 0$$

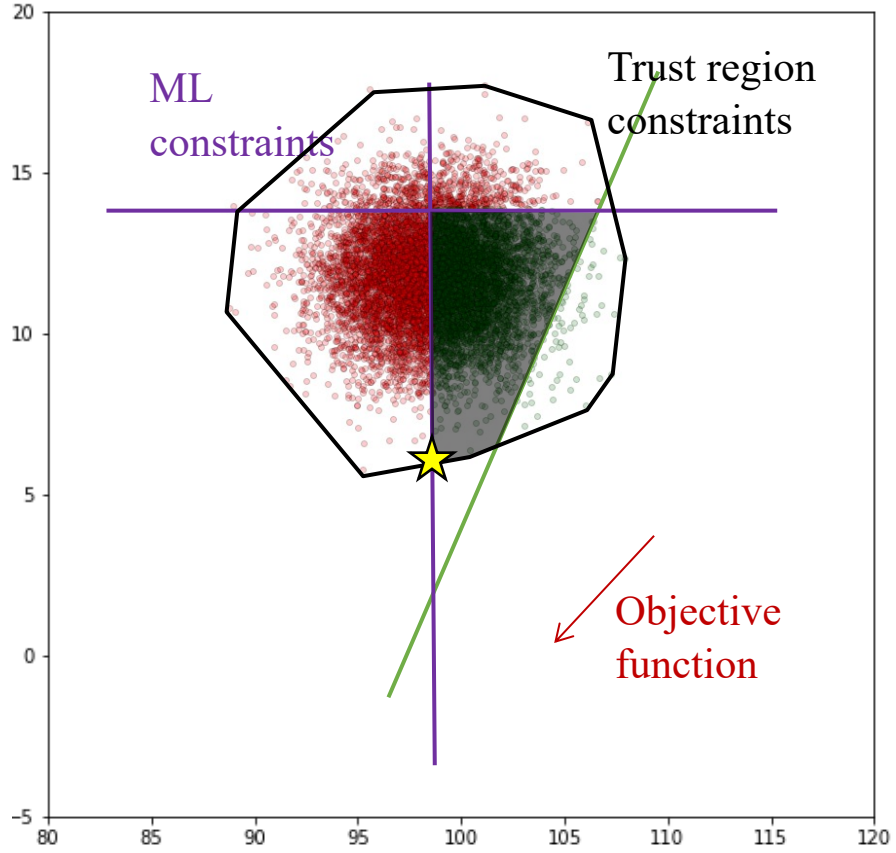
$$y = \hat{h}_D(x, w)$$

Trust region constraints

$$x = \sum_i^N \lambda_i \bar{x}_i$$

$$\sum_i^N \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$



OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \; f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1,\ldots,N$$

Proximity $\min_{\boldsymbol{x}\in\mathbb{R}^n} d(\boldsymbol{x},\hat{\boldsymbol{x}}) \rightarrow l_1, l_2, l_\infty - norms$

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \; f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \mathbf{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$

Validity $\hat{h}_{\mathcal{D}}(\boldsymbol{x}) \geq \tau$

Proximity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \; f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \mathbf{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_i^N \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_i^N \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1,\dots,N$$

Proximity

Validity

Coherence $\sum_{i\in\mathcal{C}_j} x_i = 1, \qquad j = 1,\dots,k$

Sparsity $\left|\left|\boldsymbol{x} - \hat{\boldsymbol{x}}\right|\right|_0 \leq K$

Actionability $x_i = \hat{x}_i, \qquad \forall i \in \mathcal{I}_{im}$

Causality

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \; f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

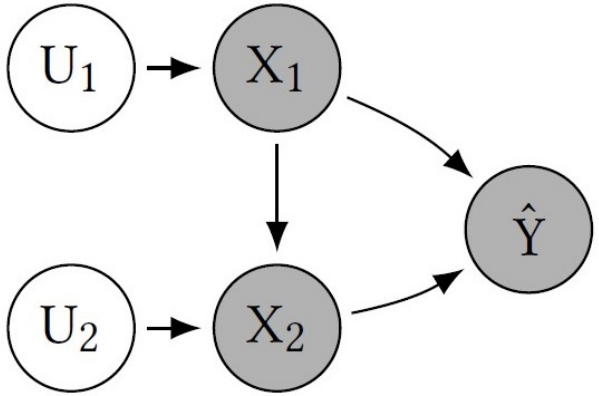
$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_i^N \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_i^N \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$



Causality $x_i = \hat{x}_i + c_i(\boldsymbol{p}_i) - c_i(\hat{\boldsymbol{p}}_i), \quad \forall i \in \mathcal{E}$

Karimi et al. (2020)

Proximity

Validity

Coherence

Sparsity

Actionability

Data manifold closeness

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n, \boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y})$$

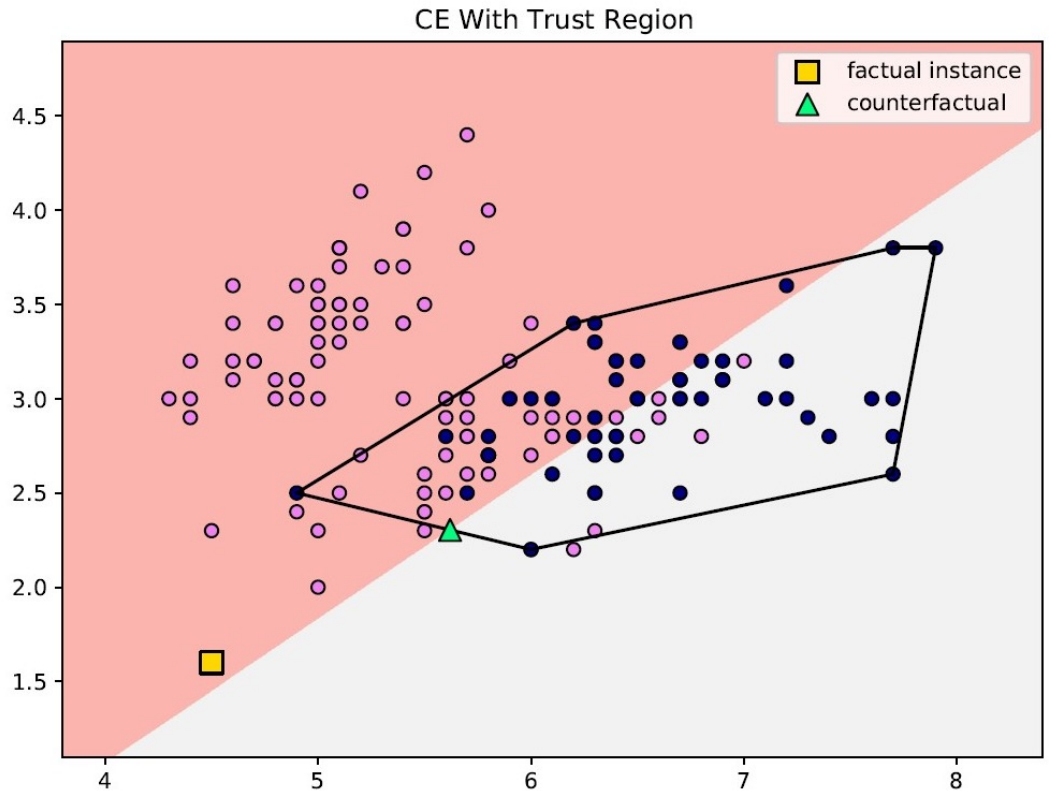
$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x}, \boldsymbol{w}, \boldsymbol{y}) \leq \mathbf{0}$$

$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x}, \boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1, \ldots, N$$



Data manifold closeness

Proximity

Validity

Coherence

Sparsity

tionability

Causality

Diversity

OCL Model

$$\min_{\boldsymbol{x}\in\mathbb{R}^n,\boldsymbol{y}\in\mathbb{R}^k} \quad f(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y})$$

$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \mathbf{0}$$

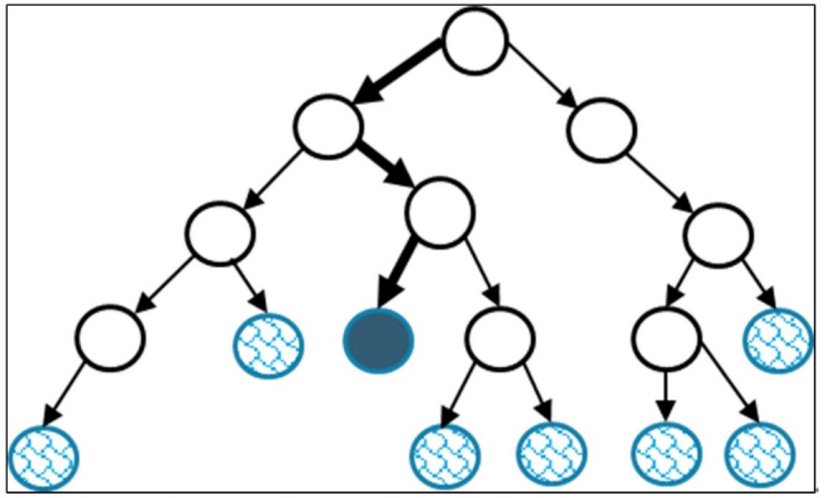
$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

$$\boldsymbol{x} = \sum_{i}^{N} \lambda_i \bar{\boldsymbol{x}}_i$$

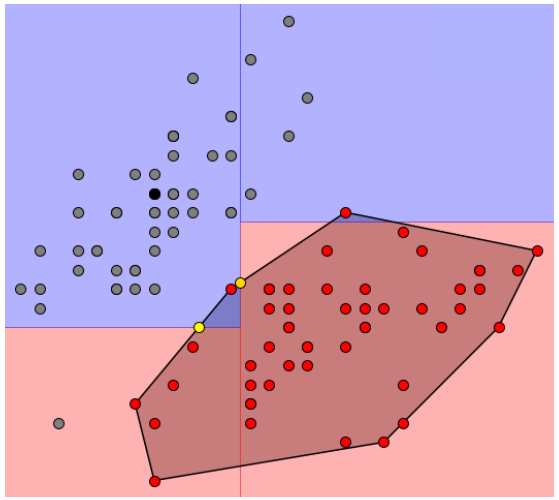
$$\sum_{i}^{N} \lambda_i = 1$$

$$\lambda_i \geq 0, \qquad i = 1,\ldots,N$$

Incumbent solutions



source



Diversity

Proximity

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Case Study

OptiCL:

A Python Package for Optimization with Constraint Learning

Codes and Examples

https://github.com/hwiberg/OptiCL

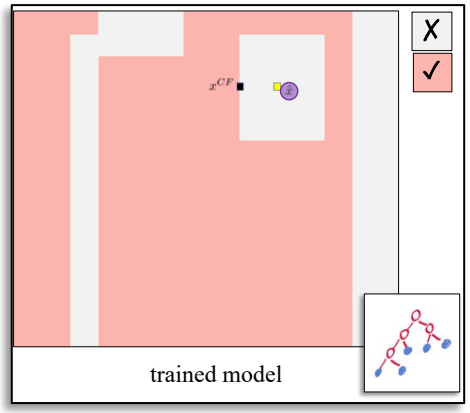
y: good or bad credit risk?

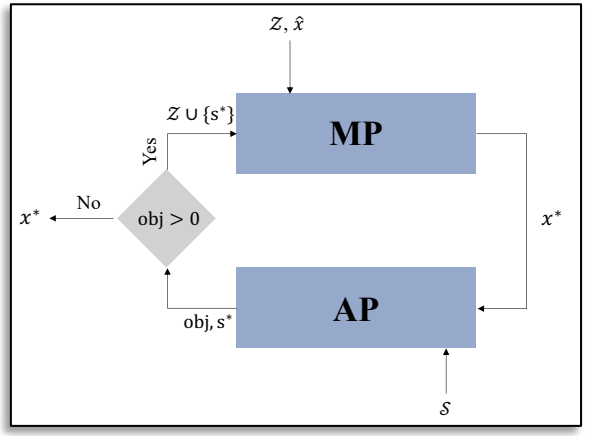
Label	Variable name
F1	duration
F2	credit_amount
F3	instalment_commitment
F4	age
F5	residence_since
F6	existing_credits

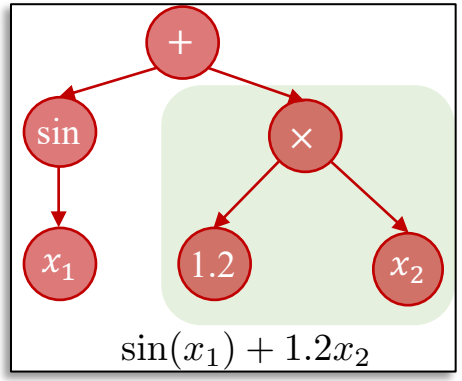
	F1	F2	F3	F4	F5	F6	F7	F8*	F9*	F
$\hat{x}$	24.0	1371.26	4.0	25.0	4.0	1.0	1.0	A	A	1
Part A: validity, proximity, coherence										
(a)	16.48	-57.75	3.88	26.71						
Part B: validity, proximity, coherence, sparsity										
(a)	7.12	–	–	–						
Part C: validity, proximity, coherence, sparsity, diversity										
(a)	7.12	–	–	–						
(b)	–	-3346.67	–	–						
(c)	–	–	–	–						
Part D: validity, proximity, coherence, sparsity, diversity, actionability										
(a)	7.12	–	–	–						
(b)	–	–	1.96	26.63						
(c)	–	–	–	75.52						
Part E: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness										
(a)	22.0	1283.52	–	–	–	–	–	B	–	4
(b)	–	1965.12	–	42.0	–	2.0	–	C	B	–
(c)	12.0	1893.04	–	29.0	–	–	–	–	–	–
Part F: validity, proximity, coherence, sparsity, diversity, actionability, data manifold closeness, causality										
(a)	–	–	–	–						
(b)	22.0	990.51	–	–						
(c)	26.83	1910.28	–	–						

Big Picture

- Counterfactual Explanations
  - Constraint Learning
  - Robust Optimization
- Symbolic Regression
  - Genetic Programming
  - Linear Optimization



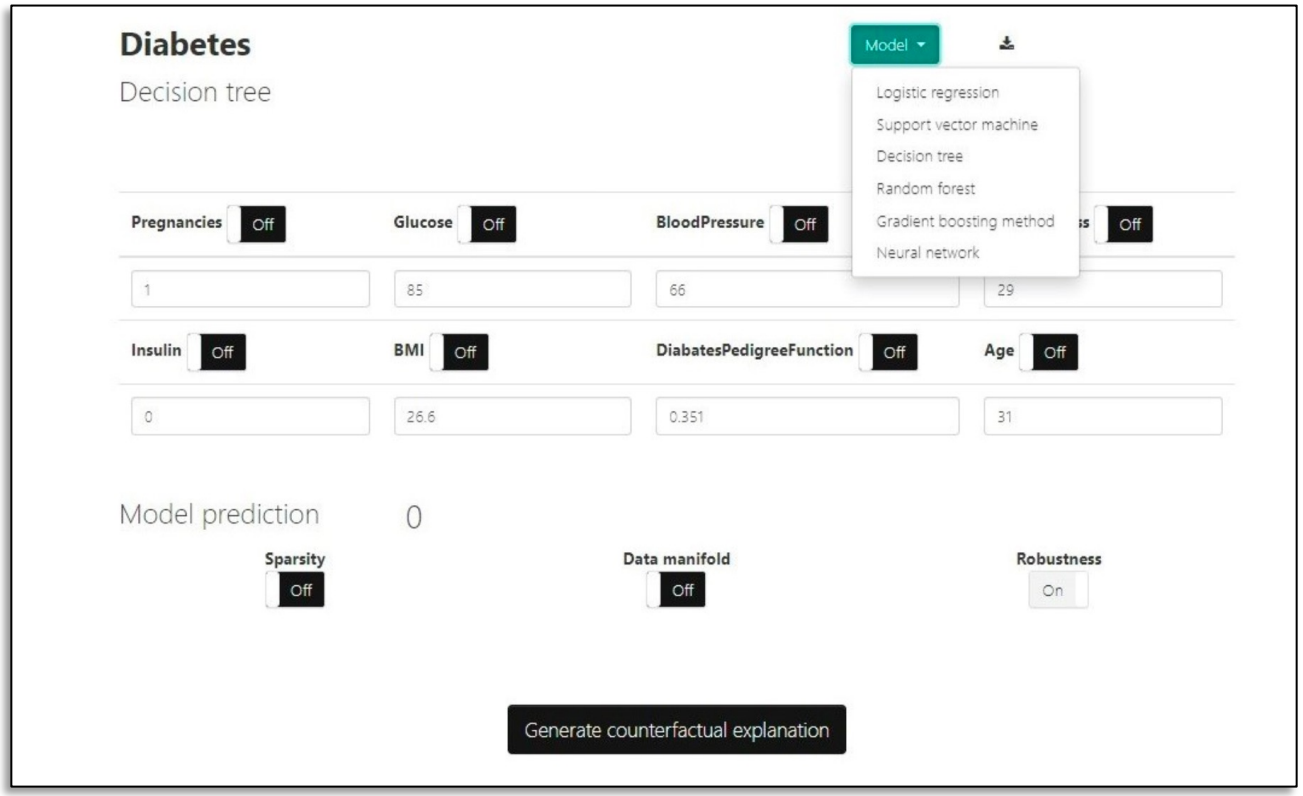


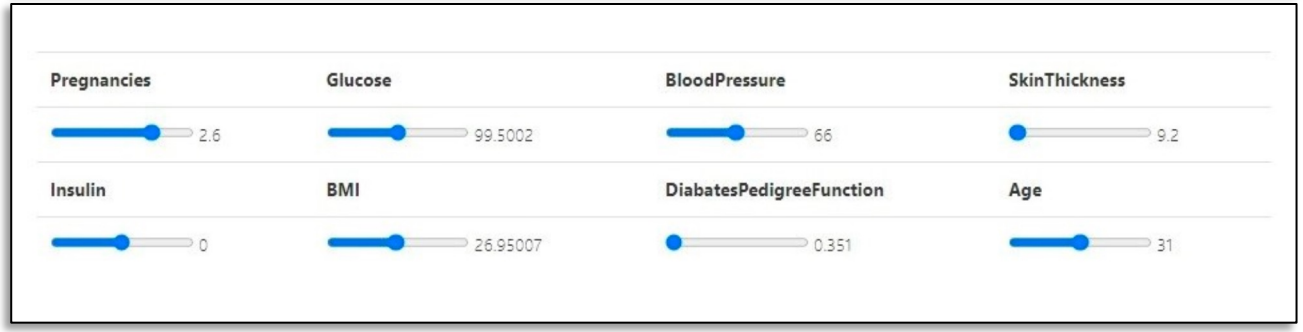


$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{n} v_i \\
\text{subject to} \quad & w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} - v_i \leq u_{i0} && i = 1, \ldots, n, \\
& w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} + v_i \geq u_{i0} && i = 1, \ldots, n, \\
& v_i \geq 0 && i = 1, \ldots, n, \\
& w_j \in \mathbb{R} && j = 0, \ldots, p.
\end{aligned}
$$

Robust Counterfactual Explanations

Let the user decide…





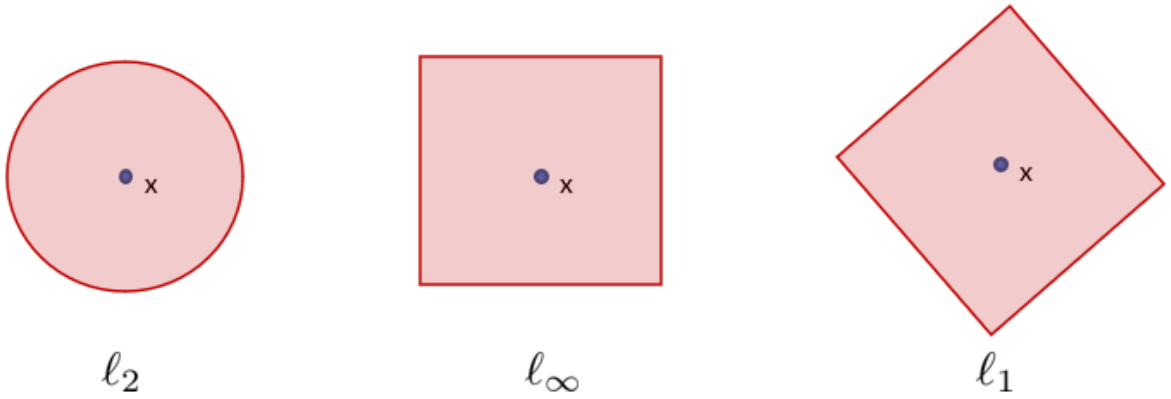
Robust Counterfactual Explanations

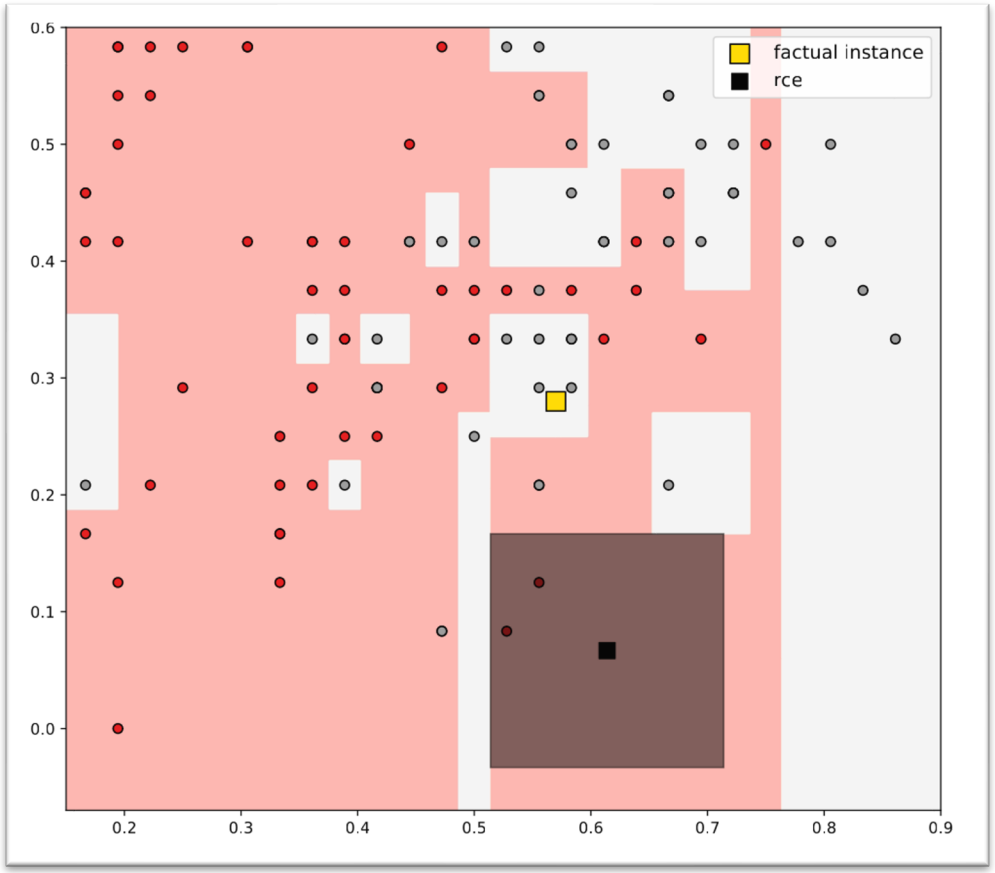
- Find counterfactual point x^{CE} such that all points in

$$x^{CE} + \mathcal{S}$$

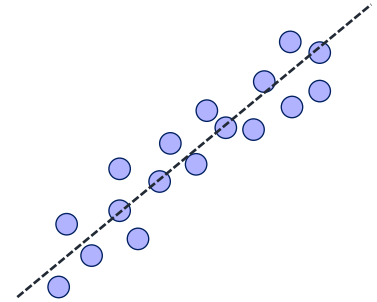
are counterfactual points.

- Uncertainty set $\mathcal{S} = \{s : \|s\| \leq \varepsilon\}$ for a small $\varepsilon > 0$
- E.g. ℓ_1,ℓ_2 or ℓ_∞-norm can be used



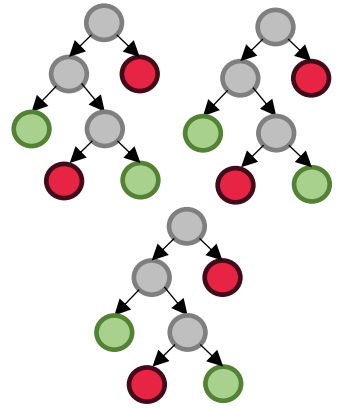


Robust Counterfactual Explanations



Goal: Derive a method which

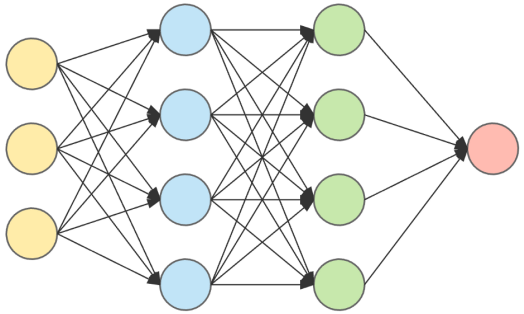
- guarantees full robustness
- is also applicable to decision trees and random forests



Optimization problem

Given a factual instance $\hat{x}$ a robust counterfactual explanation can be derived by solving

$$\begin{aligned} &\text{minimize} && d(\hat{\boldsymbol{x}}, \boldsymbol{x}) \\ &\text{subject to} && \hat{h}(\boldsymbol{x} + \boldsymbol{s}) \geq \tau, \quad \forall \boldsymbol{s} \in \mathcal{S} \\ & && \boldsymbol{x} \in \mathcal{X} \end{aligned} \tag{1}$$



Difficulties

- infinitely many constraints
- duality trick from robust optimization does not work for all types of h
- iterative approach needed: **adversarial approach**

Adversarial Approach

Master problem

The **master problem** (MP) is a relaxation of the problem (1) with a finite number of scenarios $\mathcal{Z} \subset \mathcal{S}$:

$$\begin{array}{ll} \text{minimize} & d(\hat{\boldsymbol{x}}, \boldsymbol{x}) \\ \text{subject to} & \hat{h}(\boldsymbol{x} + \boldsymbol{s}) \geq \tau, \quad \forall \boldsymbol{s} \in \mathcal{Z} \\ & \boldsymbol{x} \in \mathcal{X} \end{array}$$

Provides a lower bound for the optimal value.

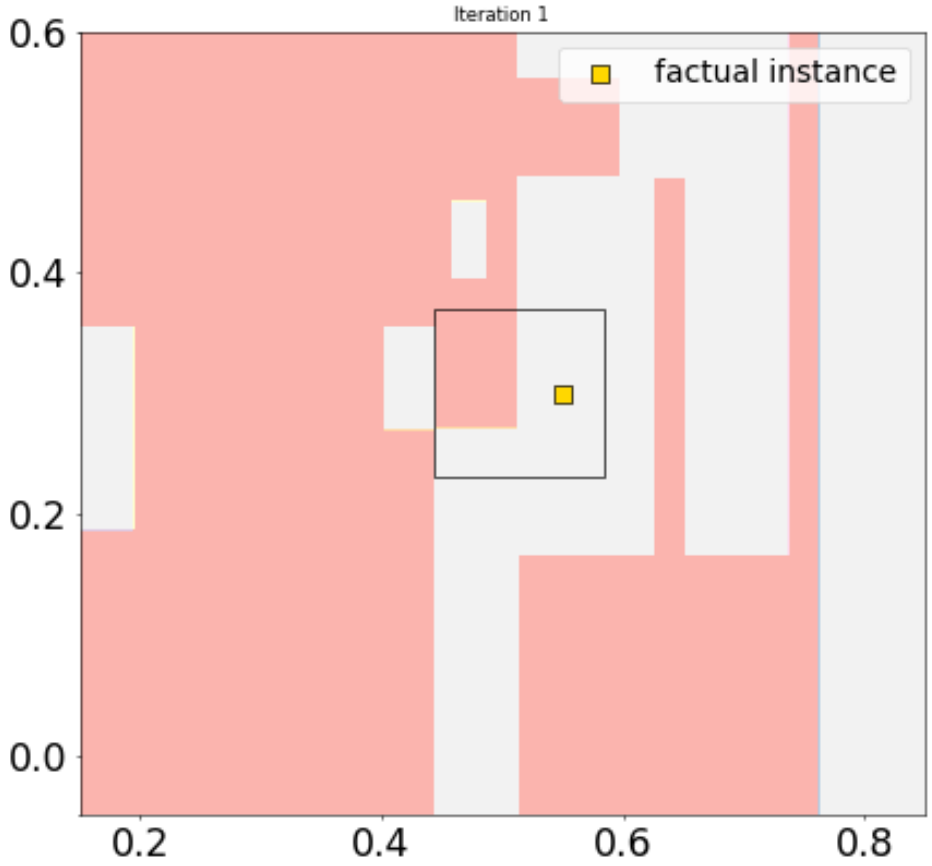
Adversarial problem

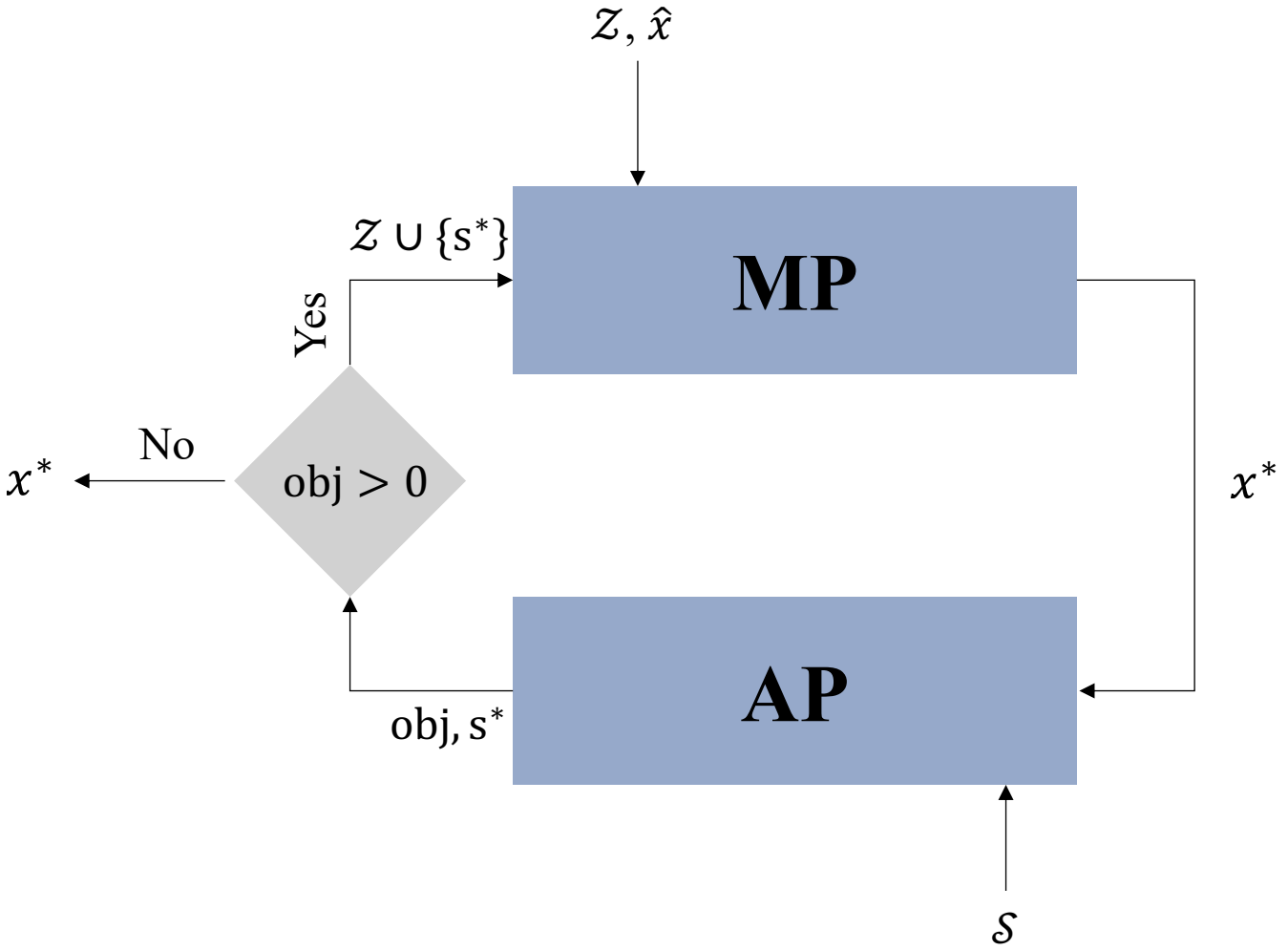
The **adversarial problem** (AP) finds a new scenario in $\mathcal{S}$ which maximally violates the constraints for current (MP) solution x^* :

$$\begin{array}{ll} \text{maximize} & \tau - \hat{h}(\boldsymbol{x}^* + \boldsymbol{s}) \\ \text{subject to} & \boldsymbol{s} \in \mathcal{S} \end{array}$$

Current solution x^* is cut-off if optimal (AP) objective value is > 0

Adversarial Approach

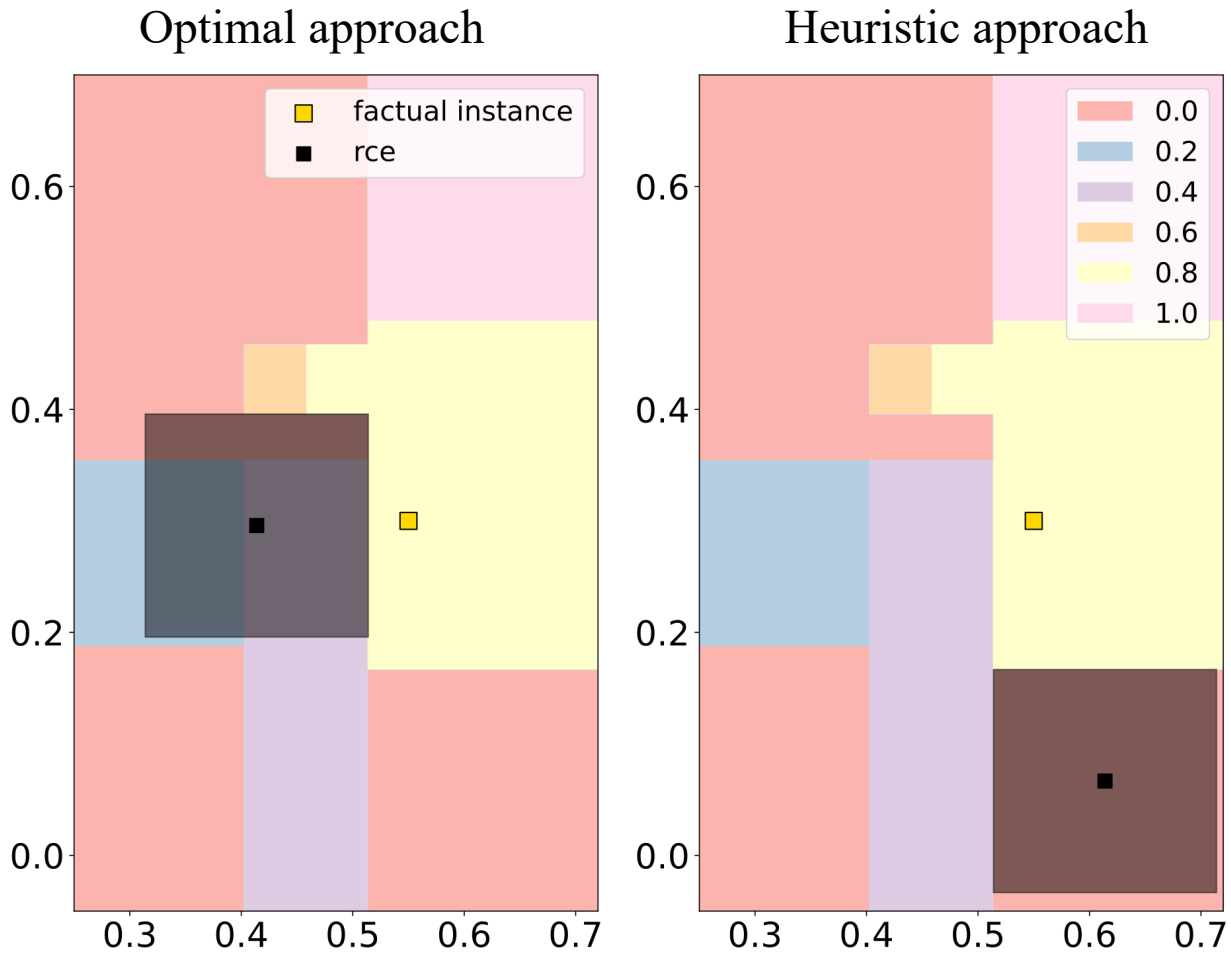




Decision Trees

Optimal approach: reformulate the DT as Lipschitz continuous function

Heuristic approach: restrict solution to be in one leaf only



Results

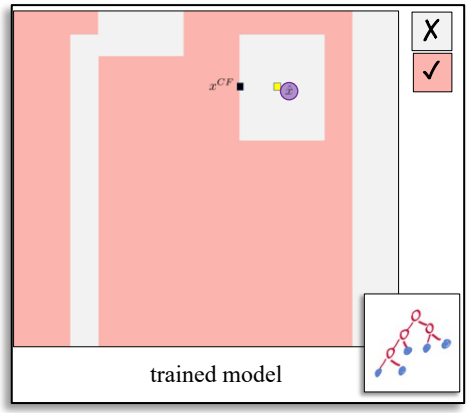
		BANKNOTE AUTHENTICATION 4 features			DIABETES 8 features			IONOSPHERE 34 features		
Model	**Specs**	**Comp. time (s)**	**# iterations**	**# early stops**	**Comp. time (s)**	**# iterations**	**# early stops**	**Comp. time (s)**	**# iterations**	**# early stops**
					$\rho = 0.01$					
Linear	ElasticNet	0.25 (0.01)	–	–	0.23 (0.01)	–	–	0.25 (0.01)	–	–
DT	max depth: 3	1.53 (0.04)	1.00 (0.00)	–	1.66 (0.07)	1.20 (0.09)	–	1.66 (0.07)	1.10 (0.07)	–
	max depth: 5	1.86 (0.09)	1.10 (0.07)	–	2.29 (0.11)	1.20 (0.09)	–	1.96 (0.08)	1.10 (0.07)	–
	max depth: 10	3.90 (0.88)	2.00 (0.58)	–	5.77 (0.48)	1.40 (0.13)	–	2.86 (0.16)	1.20 (0.09)	–
RF*	# est.: 5	3.50 (0.36)	1.75 (0.22)	–	5.89 (1.98)	2.60 (0.83)	–	3.79 (0.24)	1.70 (0.13)	–
	# est.: 10	6.74 (1.03)	2.60 (0.40)	–	7.20 (0.94)	2.35 (0.29)	–	8.76 (0.89)	2.85 (0.33)	–
	# est.: 20	21.21 (4.61)	4.55 (0.99)	–	33.55 (7.48)	6.15 (0.79)	–	22.33 (2.83)	4.40 (0.51)	–
	# est.: 50	115.79 (34.35)	7.80 (1.65)	–	110.24 (32.43)	6.47 (1.39)	3 ($\bar{\rho} = 0.007$)	137.26 (33.37)	8.20 (1.20)	–
	# est.: 100	214.38 (65.07)	6.44 (1.31)	2 ($\bar{\rho} = 0.009$)	274.09 (71.93)	8.87 (1.49)	5 ($\bar{\rho} = 0.004$)	285.62 (95.57)	8.27 (2.02)	9 ($\bar{\rho} = 0.004$)
GBM**	# est.: 5	2.70 (0.22)	1.20 (0.14)	–	2.76 (0.22)	1.85 (0.17)	–	2.37 (0.15)	1.60 (0.13)	–
	# est.: 10	3.20 (0.30)	1.45 (0.15)	–	2.72 (0.29)	1.50 (0.24)	–	4.35 (0.44)	2.75 (0.30)	–
	# est.: 20	5.94 (0.50)	2.60 (0.23)	–	4.25 (0.45)	2.15 (0.28)	–	9.01 (1.11)	3.85 (0.50)	–
	# est.: 50	18.38 (1.62)	4.05 (0.35)	–	24.60 (8.21)	5.85 (1.41)	–	81.33 (28.39)	8.90 (1.36)	–
	# est.: 100	87.11 (26.24)	7.28 (0.77)	2 ($\bar{\rho} = 0.006$)	164.32 (42.12)	11.63 (2.00)	1 ($\bar{\rho} = 0.004$)	137.98 (22.74)	10.33 (0.97)	2 ($\bar{\rho} = 0.007$)
MLP	(10,)	1.63 (0.04)	1.00 (0.00)	–	1.55 (0.06)	1.00 (0.00)	–	2.22 (0.19)	1.80 (0.21)	–
	(10, 10, 10)	2.98 (0.15)	1.15 (0.08)	–	2.40 (0.12)	1.15 (0.08)	–	13.01 (3.57)	2.30 (0.40)	–
	(50,)	2.60 (0.14)	1.00 (0.00)	–	2.09 (0.12)	1.05 (0.05)	–	5.75 (0.75)	1.20 (0.12)	–
	(100,)	3.53 (0.15)	1.00 (0.00)	–	4.36 (0.62)	1.10 (0.07)	–	61.31 (33.11)	1.50 (0.22)	10 ($\bar{\rho} = 0.000$)

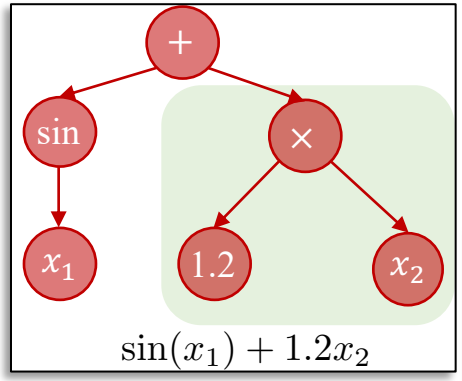
* max depth of each decision tree equal to 3.; ** max depth of each decision tree equal to 2.

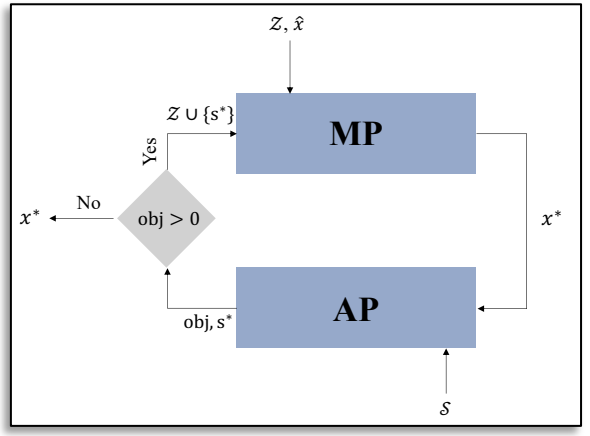
Table 1. Generation of robust CEs for 20 factual instances, using ℓ_∞-norm as uncertainty set.

Big Picture

- Counterfactual Explanations
  - Constraint Learning
  - Robust Optimization
- Symbolic Regression
  - Genetic Programming
  - Linear Optimization







$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{n} v_i \\
\text{subject to} \quad & w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} - v_i \leq u_{i0} && i = 1, \ldots, n, \\
& w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} + v_i \geq u_{i0} && i = 1, \ldots, n, \\
& v_i \geq 0 && i = 1, \ldots, n, \\
& w_j \in \mathbb{R} && j = 0, \ldots, p.
\end{aligned}
$$

Symbolic Regression

Symbolic Regression

- Very old ML technique to discover mathematical equation (symbolic expressions)
- Finds a mathematical equation that best fits the given dataset
- Important to understand complex relationships between input and target variables

x_1	x_2	y
65.13	8.05	4222.15
72.70	1.57	178.58
60.08	4.20	1058.34
91.44	2.44	543.58
48.78	3.36	550.43

$$y = x_1 x_2^2$$

$$\boxed{E = mc^2}$$

Applications: Science

Richard Feynman

"A core challenge for both physics and artificial intelligence (AI) is symbolic regression: finding a symbolic expression that matches data from an unknown function. […] In this spirit, we develop a recursive multidimensional symbolic regression algorithm that combines neural network fitting with a suite of physics-inspired techniques. We apply it to 100 equations from the Feynman Lectures on Physics, and it discovers all of them, while previous publicly available software cracks only 71; for a more difficult physics-based test set, we improve the state-of-the-art success rate from 15 to 90%."

SCIENCE ADVANCES | RESEARCH ARTICLE

COMPUTER SCIENCE

AI Feynman: A physics-inspired method for symbolic regression

Silviu-Marian Udrescu[1] and Max Tegmark[1,2]*

(link)

Table 6. Tested bonus equations. Goldstein 8.56 is for the special case where the vectors **p** and **A** are parallel.

Source	Equation	Solved	Solved by Eureqa	Methods used
Rutherford scattering	$A = \left(\frac{Z_1 Z_2 \alpha h c}{4E\sin^2(\frac{\theta}{2})}\right)^2$	Yes	No	da, bf-sqrt
Friedman equation	$H = \sqrt{\frac{8\pi G}{3}\rho - \frac{kc^2}{a^2}}$	Yes	No	da, bf-squared
Compton scattering	$U = \frac{E}{1+\frac{E}{mc^2}(1-\cos\theta)}$	Yes	No	da, bf
Radiated gravitational wave power	$P = -\frac{32}{5}\frac{G^4}{c^5}\frac{(m_1 m_2)^2(m_1+m_2)}{r^5}$	No	No	–
Relativistic aberration	$\theta_1 = \arccos\left(\frac{\cos\theta_2 - \frac{v}{c}}{1-\frac{v}{c}\cos\theta_2}\right)$	Yes	No	da, bf-cos
N-slit diffraction	$I = I_0\left[\frac{\sin(\alpha/2)}{\alpha/2}\frac{\sin(N\delta/2)}{\sin(\delta/2)}\right]^2$	Yes	No	da, sm, bf
Goldstein 3.16	$v = \sqrt{\frac{2}{m}\left(E - U - \frac{L^2}{2mr^2}\right)}$	Yes	No	da, bf-squared
Goldstein 3.55	…	Yes	No	da, sym-, bf

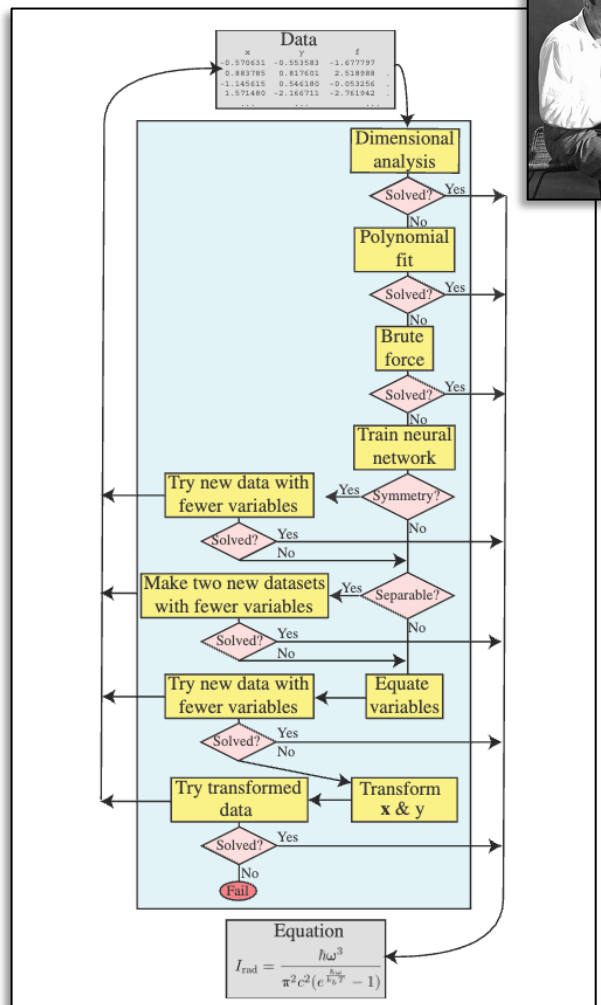


Fig. 1. Schematic illustration of our AI Feynman algorithm. It is iterative as described in the text, with four of the steps capable of generating new mystery datasets that get sent to fresh instantiations of the algorithm, which may or may not return a solution.

Applications: Fraud Detection

"Explainable AI has been gaining attention in recent years, with one area of research being Symbolic Regression (SR). SR aims to find analytical (concise, closed-form) expressions that describe functional dependencies in a data set. Since an expression can be understood simply by inspection, SR can be used to create a model that is transparent and explainable."

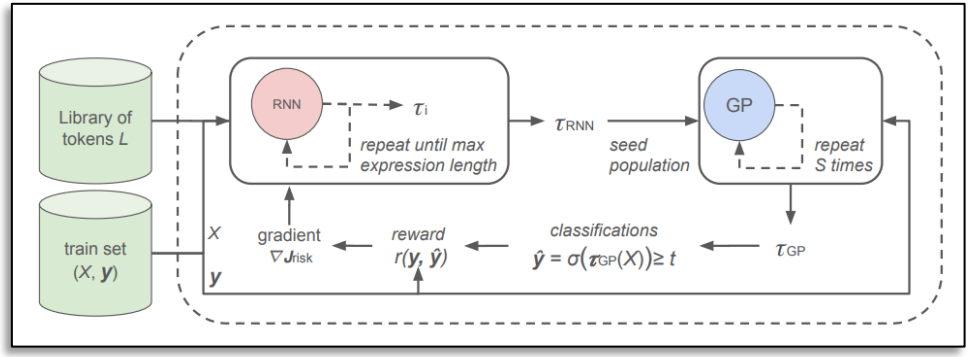
Explainable Fraud Detection with Deep Symbolic Classification

Samantha Visbeek
SamanthaVisbeek@hotmail.com
RiskQuest
Amsterdam, the Netherlands

Erman Acar*
e.acar@uva.nl
Universiteit van Amsterdam
Amsterdam, the Netherlands

Floris den Hengst*
f.den.hengst@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

(link)



D DERIVATION OF DECISION RULE

The expression that resulted in the highest performance was:

$$f = \sqrt{\text{externalDest} + \text{type\_cash-out}} \cdot (\text{amount} - \text{maxDest7} + \text{type\_transfer}), \quad (10)$$

where we have three Boolean features that either have value 0 or 1: *externalDest*, *type_cash-out* and *type_transfer*. The other features *amount* and *maxDest7* are numerical and positive. The decision rule is defined as:

$$\hat{y} = 1(\text{fraud}),$$
$$\text{if } \sigma(f) > 0.7,$$

as this expression was found by training DSC on a threshold $t = 0.7$. Rewriting the sigmoid $\sigma(f) = (1 + e^{-f})^{-1}$ gives us:

$$\hat{y} = 1(\text{fraud}),$$
$$\text{if } f > 0.85.$$

Solution Approaches

But…

Published in Transactions on Machine Learning Research (10/2022)

Symbolic Regression is NP-hard

Marco Virgolin
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands

Solon P. Pissis
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
Vrije Universiteit, Amsterdam, the Netherlands

(link)

"**Abstract.** Symbolic regression (SR) is the task of learning a model of data in the form of a mathematical expression. By their nature, SR models have the potential to be accurate and human-interpretable at the same time. Unfortunately, finding such models, *i.e.*, performing SR, appears to be a computationally intensive task. Historically, SR has been tackled with heuristics such as greedy or genetic algorithms and, while some works have hinted at the possible hardness of SR, no proof has yet been given that SR is, in fact, NP-hard. This begs the question: Is there an exact polynomial-time algorithm to compute SR models? We provide evidence suggesting that the answer is probably negative by showing that SR is NP-hard."

Solution Approaches

Dataset: $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$$

function space

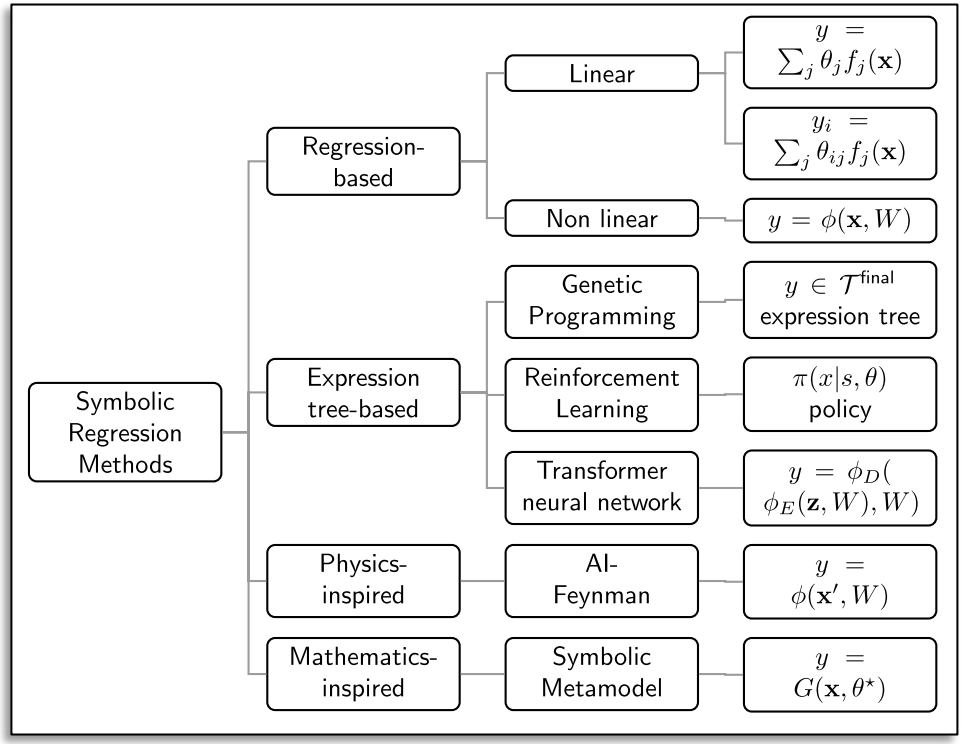
loss function

Example:

$$\{+, -, \times, x_1, x_2, \mathbb{R}\}$$

$\mathcal{F}$: All polynomials of arbitrary degree in x_1 and x_2

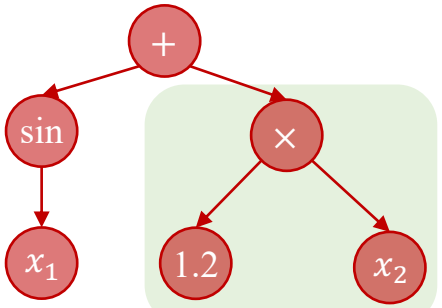
$$\ell(f(\boldsymbol{x}_i), y_i) = \big(f(\boldsymbol{x}_i) - y_i\big)^2$$



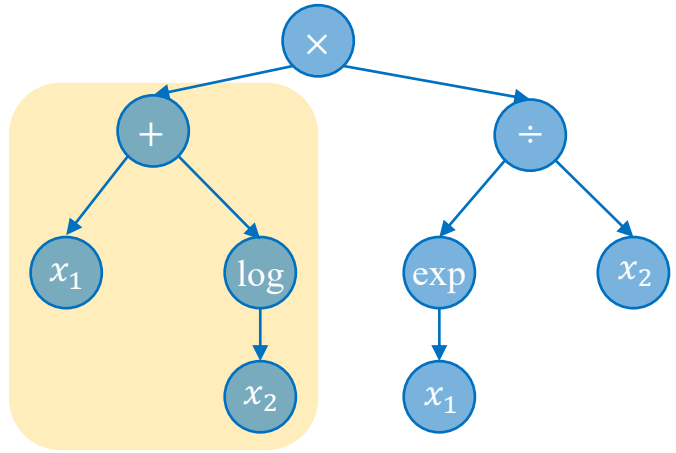
Makke, N. & Chawla, S. "Interpretable scientific discovery with symbolic regression: a review," Artificial Intelligence Review, 57:2, 2024.

Solution Approaches: Genetic Programming

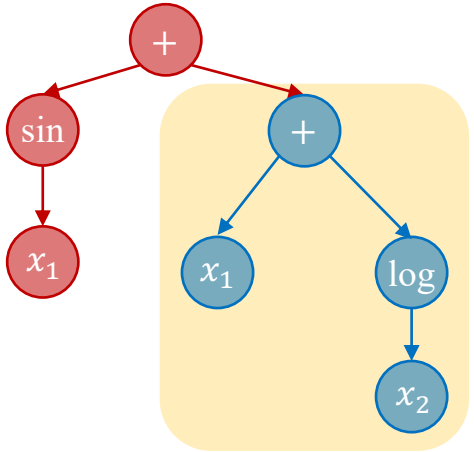
- A function can be represented in tree structure
- New trees are generated using operations inspired by Darwinian evolution:
  - Crossover
  - Mutation
  - Selection
  - Replication
- Search space is very complex



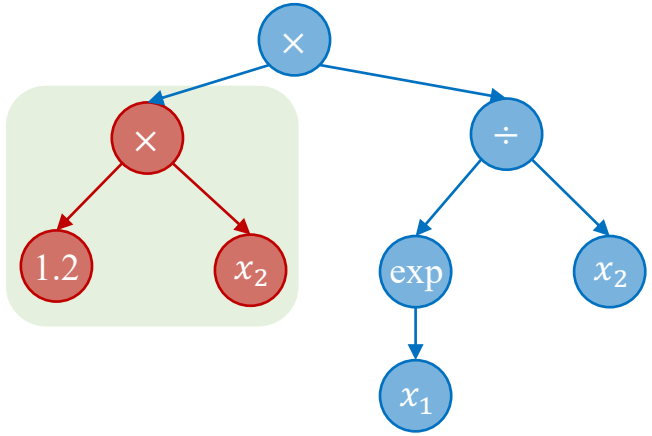
$\sin(x_1) + 1.2x_2$



$(x_1 + \log(x_2))(\exp(x_1)/x_2)$



$\sin(x_1) + x_1 + \log(x_2)$



$1.2x_2(\exp(x_1)/x_2)$

Koza, J.R. "Genetic Programming: On the Programming of Computers by Means of Natural Selection," MIT Press, Cambridge, MA, USA, 1992.

Solution Approaches: Linear Optimization

Dataset: $\{(\boldsymbol{x}_i, y_i) : i = 1, \ldots, n\}$

monomial $$f(\boldsymbol{x}) = h \cdot x_1^{w_1} \cdot x_2^{w_2} \ldots x_p^{w_p}$$

Example: $$f(\boldsymbol{x}) = \frac{1}{4\pi} x_1^1 x_2^1 x_3^{-1} x_4^{-2}$$

$$\boxed{f(q_1, q_2, \varepsilon, r) = \frac{q_1 q_2}{4\pi\varepsilon r^2}}$$

$$\log(f(\boldsymbol{x}_i)) = \underbrace{w_0}_{\log(h)} + \sum_{j=1}^{p} w_j \underbrace{u_{ij}}_{\log(x_{ij})}$$

Solution Approaches: Linear Optimization

$$f^* = \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(f(\boldsymbol{x}_i), y_i)$$

$$\ell(f(\boldsymbol{x}_i), y_i) = |\log(f(\boldsymbol{x}_i)) - \log(y_i)| = |w_0 + \sum_{j=1}^{p} w_j u_{ij} - \underbrace{u_{i0}}_{\log(y_i)}|$$

$$\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{n} v_i & & \\
\text{subject to} \quad & w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} - v_i \leq u_{i0} & & i = 1, \ldots, n, \\
& w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} + v_i \geq u_{i0} & & i = 1, \ldots, n, \\
& v_i \geq 0 & & i = 1, \ldots, n, \\
& w_j \in \mathbb{R} & & j = 0, \ldots, p.
\end{aligned}$$

Solution Approaches: Linear Optimization

Eqn. label	Eqn.	Running time [s]			Max. rel. noise level		
		LP	MINLP	AIF	LP	MINLP	AIF
I.12.2	$\frac{q_1 q_2}{4\pi\epsilon r^2}$	0.622	1	5975	10^{-3}		10^{-2}
I.12.4	$\frac{q_1}{4\pi\epsilon r^2}$	0.554	1	12	10^{-2}		10^{-2}
I.24.13	$\frac{q}{C}$	0.614	1	10	10^{-2}		10^{-2}
I.32.5	$\frac{q^2 a^2}{6\pi\epsilon c^3}$	0.606		13	10^{-3}		10^{-2}
II.11.20	$\frac{n_\rho r_d^2 E_f}{3k_b T}$	0.641	9	18	10^{-3}		10^{-3}

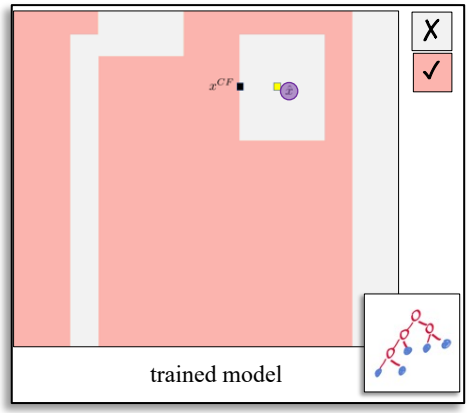
Work in progress…

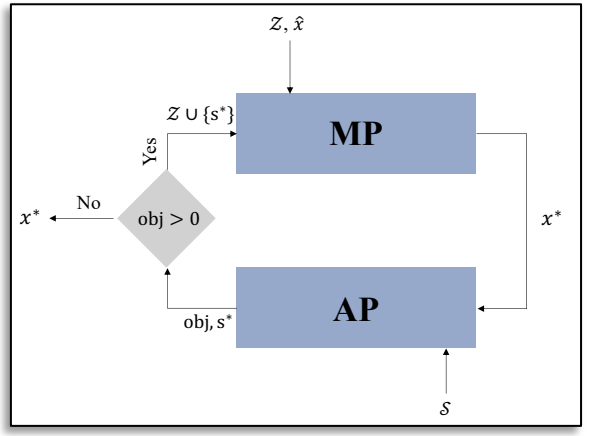
Austel, V., Cornelio, C., Dash, S., Goncalves, J., Horesh, L. & Jospehson, T. "Symbolic regression using mixed-integer nonlinear optimization," arXiv:2006.06813, 2020.

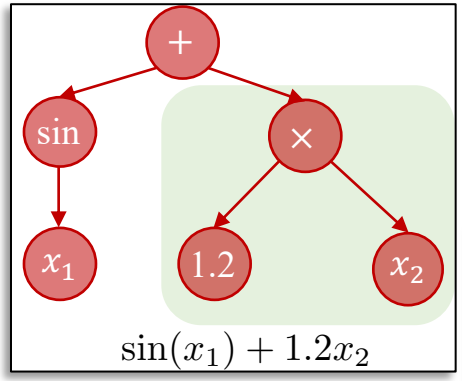
Udrescu, S-M. &Tegmark, M. "AI Feynman: A physics-inspired method for symbolic regression," Sci. Adv. 6, eaay 2631, 2020.

Big Picture

- Counterfactual Explanations
  - Constraint Learning
  - Robust Optimization
- Symbolic Regression
  - Genetic Programming
  - Linear Optimization

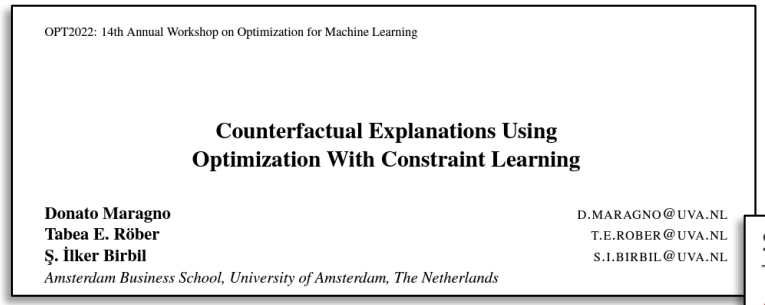






$$
\begin{aligned}
\text{minimize} \quad & \textstyle\sum_{i=1}^{n} v_i \\
\text{subject to} \quad & w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} - v_i \leq u_{i0} && i = 1, \ldots, n, \\
& w_0 + \textstyle\sum_{j=1}^{p} w_j u_{ij} + v_i \geq u_{i0} && i = 1, \ldots, n, \\
& v_i \geq 0 && i = 1, \ldots, n, \\
& w_j \in \mathbb{R} && j = 0, \ldots, p.
\end{aligned}
$$

Reading Material



OPT2022: 14th Annual Workshop on Optimization for Machine Learning

Counterfactual Explanations Using Optimization With Constraint Learning

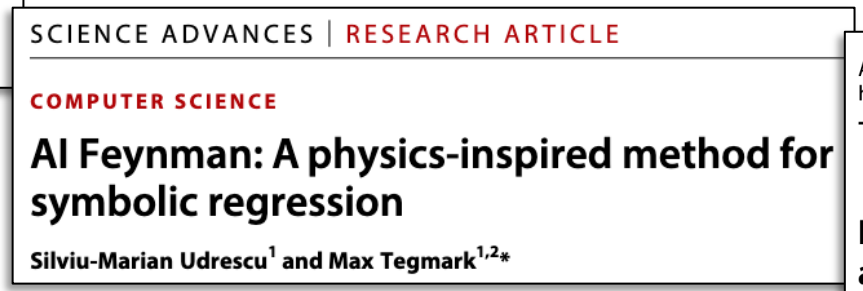
Donato Maragno D.MARAGNO@UVA.NL
Tabea E. Röber T.E.ROBER@UVA.NL
Ş. İlker Birbil S.I.BIRBIL@UVA.NL
Amsterdam Business School, University of Amsterdam, The Netherlands

Explainable Fraud Detection with Deep Symbolic Classification

Samantha Visbeek
SamanthaVisbeek@hotmail.com
RiskQuest
Amsterdam, the Netherlands

Erman Acar*
e.acar@uva.nl
Universiteit van Amsterdam
Amsterdam, the Netherlands

Floris den Hengst*
f.den.hengst@vu.nl
Vrije Universiteit Amsterdam
Amsterdam, the Netherlands

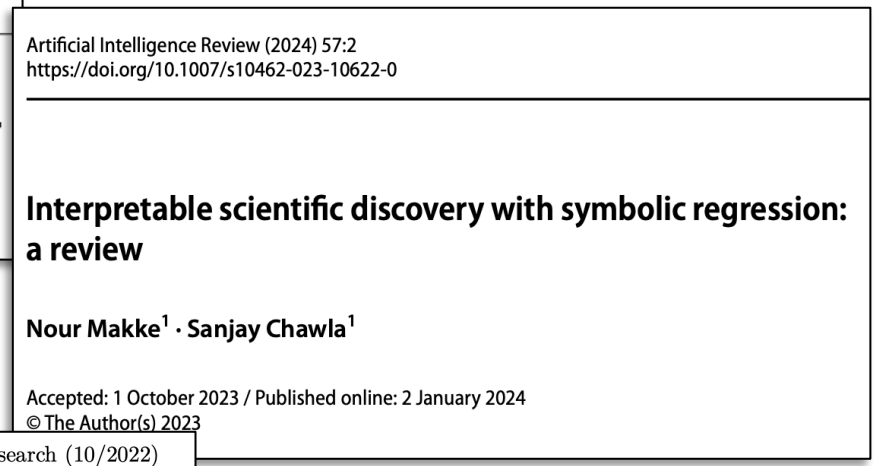


SCIENCE ADVANCES | RESEARCH ARTICLE

COMPUTER SCIENCE

AI Feynman: A physics-inspired method for symbolic regression

Silviu-Marian Udrescu[1] and Max Tegmark[1,2]*

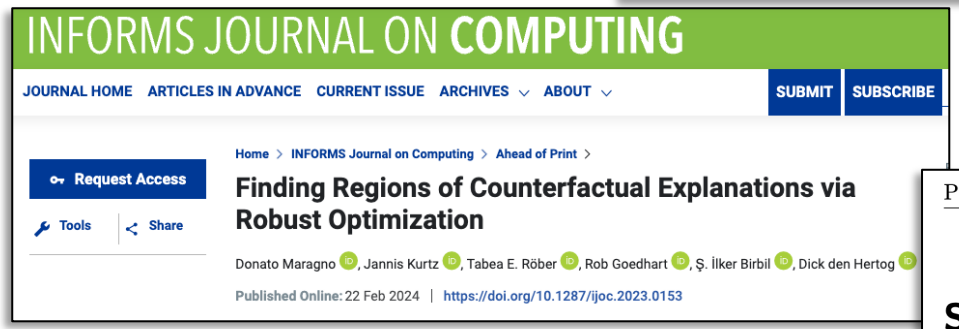


Artificial Intelligence Review (2024) 57:2
https://doi.org/10.1007/s10462-023-10622-0

Interpretable scientific discovery with symbolic regression: a review

Nour Makke[1] · Sanjay Chawla[1]

Accepted: 1 October 2023 / Published online: 2 January 2024



INFORMS JOURNAL ON **COMPUTING**

JOURNAL HOME | ARTICLES IN ADVANCE | CURRENT ISSUE | ARCHIVES | ABOUT | SUBMIT | SUBSCRIBE

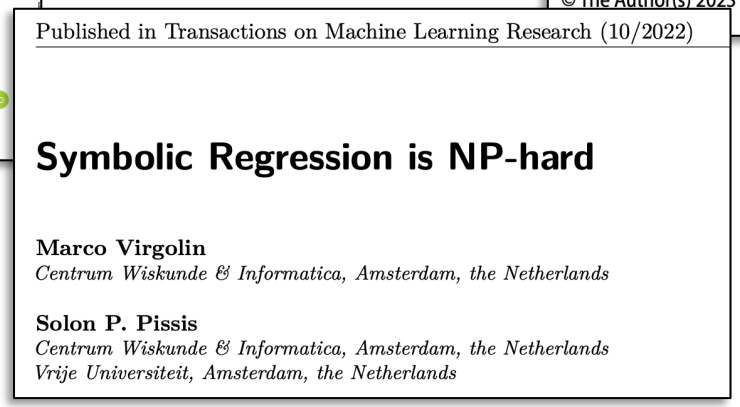
Request Access | Tools | Share

Home > INFORMS Journal on Computing > Ahead of Print >

Finding Regions of Counterfactual Explanations via Robust Optimization

Donato Maragno, Jannis Kurtz, Tabea E. Röber, Rob Goedhart, Ş. İlker Birbil, Dick den Hertog

Published Online: 22 Feb 2024 | https://doi.org/10.1287/ijoc.2023.0153



Published in Transactions on Machine Learning Research (10/2022)

Symbolic Regression is NP-hard

Marco Virgolin
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands

Solon P. Pissis
Centrum Wiskunde & Informatica, Amsterdam, the Netherlands
Vrije Universiteit, Amsterdam, the Netherlands