Counterfactual Explanations Using Optimization with Constraint Learning

Ilker Birbil

UvA

Donato Maragno, Tabea Röber, Dick den Hertog, Rob Goedhart

Motivation

nature

Explore content · About the journal · Publish with us · Subscribe

nature › news › article

NEWS | 30 November 2020

'It will change everything'
DeepMind's AI makes giga
in solving protein structur

Google's deep-learning program for determining the 3D shapes

The New York Times

How A.I. Can Help Handle

Forbes

The Amazing Opportunities Of AI In The Future Of The Educational Metaverse

Rem Darbinyan Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

Apr 27, 2022, 08:00am EDT

The World Economic Forum

Why artificial intelligence is vital in t

Why artif
an ethica

1 week ge

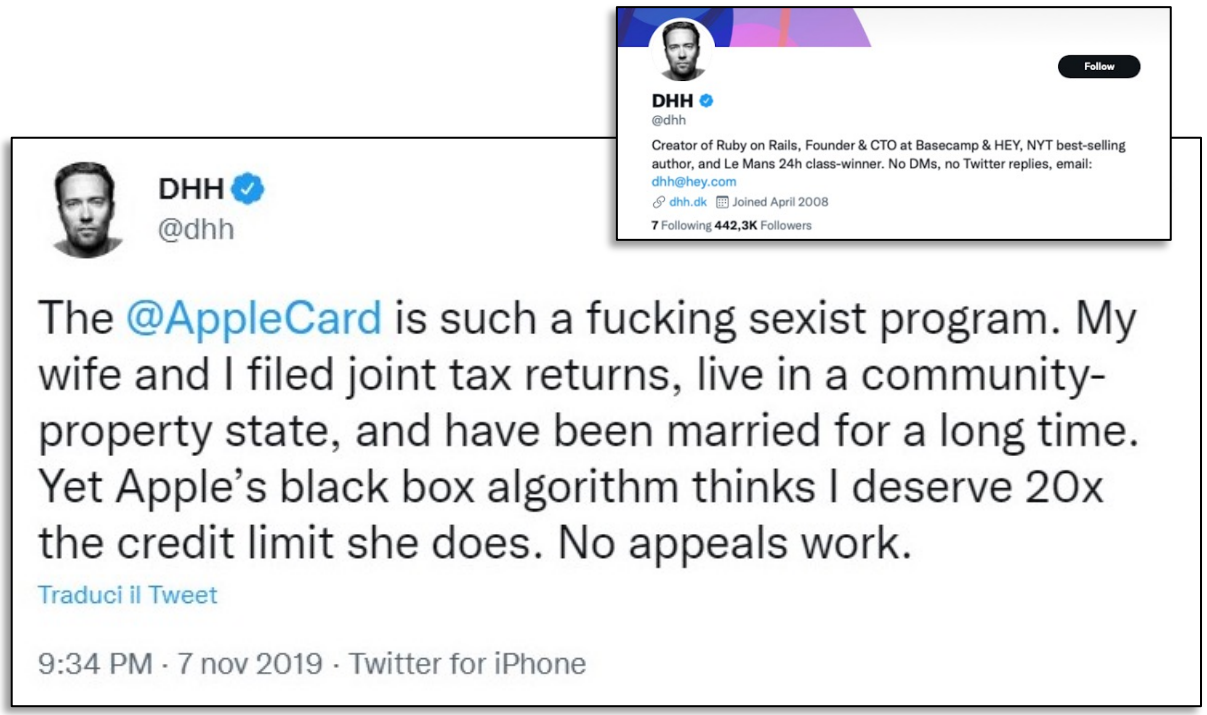
Waste less, sell more - how one start up is using AI to transform food retail

gist-level
tion of skin cancer

An artificial intelligence trained to classify images of skin lesions as benign lesions or malignant skin cancers achieves the accuracy of board-certified dermatologists.

In this work, we pretrain a deep neural network at general object recognition, then fine-tune it on a dataset of ~130,000 skin lesion images comprised of over 2000 diseases.

Motivation


DHH
@dhh
Follow
DHH
@dhh
Creator of Ruby on Rails, Founder & CTO at Basecamp & HEY, NYT best-selling author, and Le Mans 24h class-winner. No DMs, no Twitter replies, email: dhh@hey.com
dhh.dk Joined April 2008
7 Following 442,3K Followers
DHH
@dhh
The @AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work.
Traduci il Tweet
9:34 PM · 7 nov 2019 · Twitter for iPhone


DHH @dhh · 9 nov 2019
She spoke to two Apple reps. Both very nice, courteous people representing an utterly broken and reprehensible system. The first person was like "I don't know why, but I swear we're not discriminating, IT'S JUST THE ALGORITHM". I shit you not. "IT'S JUST THE ALGORITHM!".
64
696
4.553

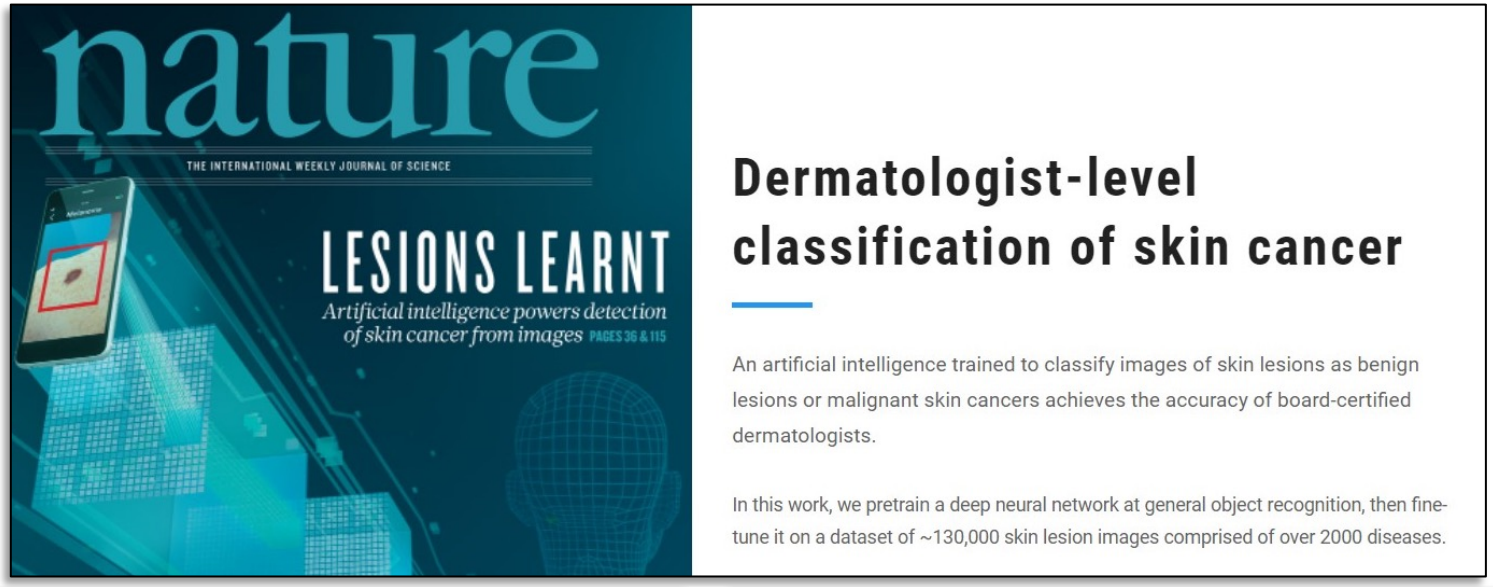

Steve Wozniak
@stevewoz
I'm a current Apple employee and founder of the company and the same thing happened to us (10x) despite not having any separate assets or accounts. Some say the blame is on Goldman Sachs but the way Apple is attached, they should share responsibility.
8:06 AM · Nov 10, 2019

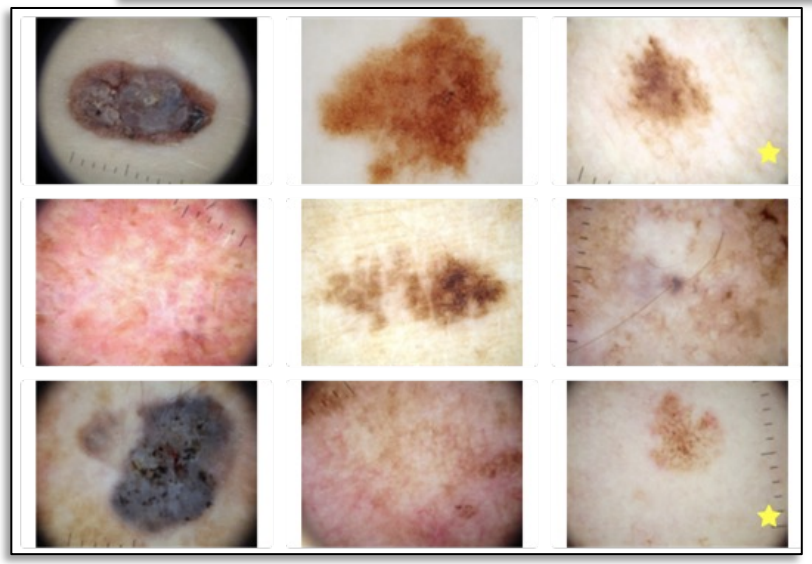

CNN BUSINESS
Apple co-founder Steve Wozniak says Apple Card discriminated against his wife
By Clare Duffy, CNN Business
Updated 1615 GMT (0015 HKT) November 11, 2019

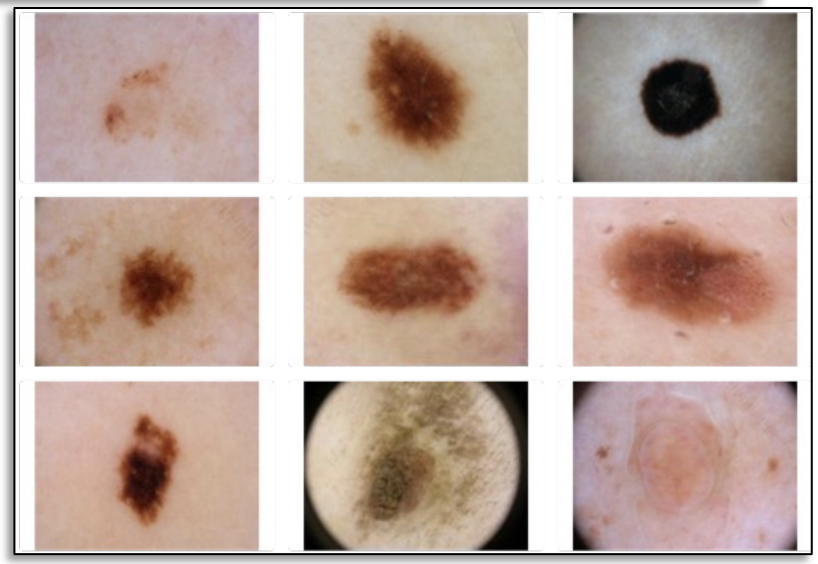
Motivation

Did the model learn the true pattern?
Is the model discriminating?

Test accuracy
In-lab vs. Real-life deployment







source

Motivation

> "An AI system is **explainable** if the task model is intrinsically interpretable (here the AI system is the task model) or if the non-interpretable task model is complemented with an interpretable and faithful explanation (here the AI system also contains a post-hoc explanation)."
>
> Markus et al. (2020)

Explainable Artificial Intelligence (XAI) Methods

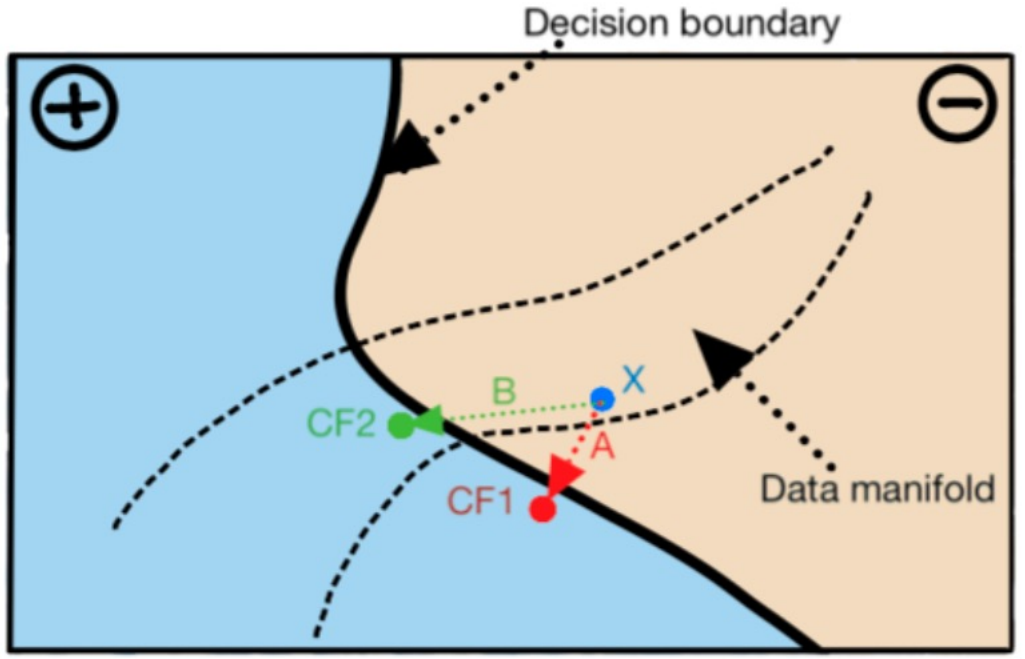
- Model-based explanations: linear/logistic regression, decision trees, k-nearest neighbours.
- Post-hoc explanations: instance level vs global level
  - **Model-based explanations**: other interpretable models are used to explain the uninterpretable model.
  - **Attribution-based explanations**: features importance methods.
  - **Example-based explanations** → Counterfactual Explanations (CE)

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



(Verma et al., 2020)

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

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

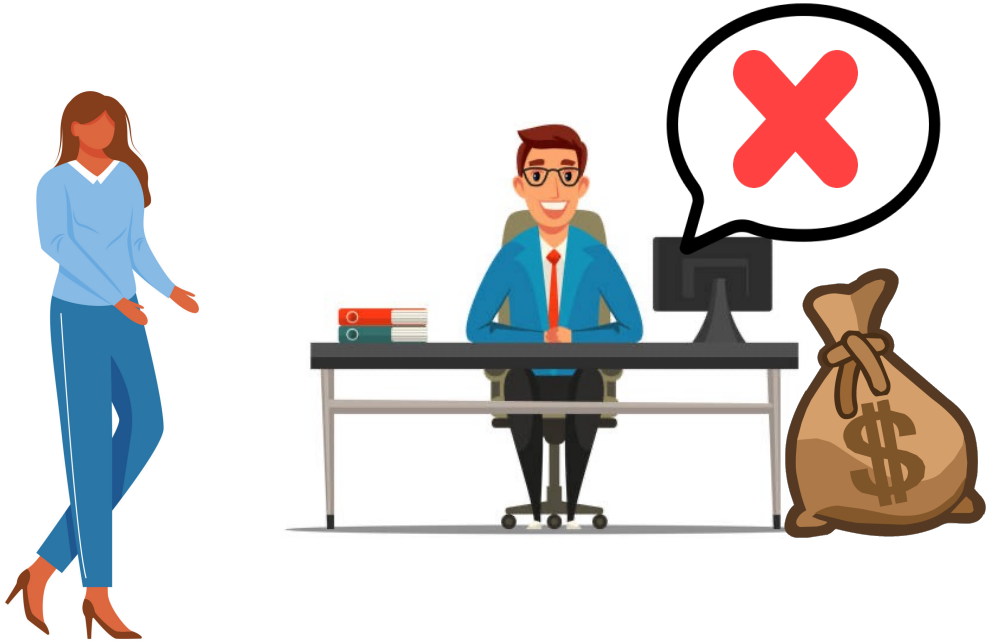
arXiv, 2020

Counterfactual Explanations

University of Amsterdam
Amsterdam Business School

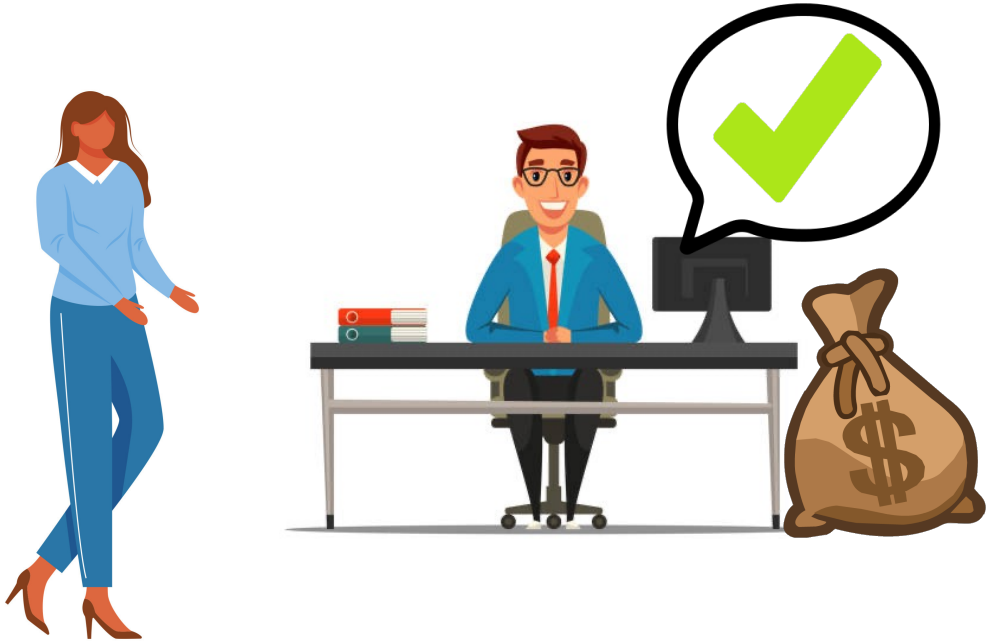
Factual Instance

Beatrice is 27yo
Full-time job: 45K $/y
Account balance: 50K $



Counterfactual

Beatrice is 27yo
Full-time job: **50K** $/y
Account balance: **60K** $

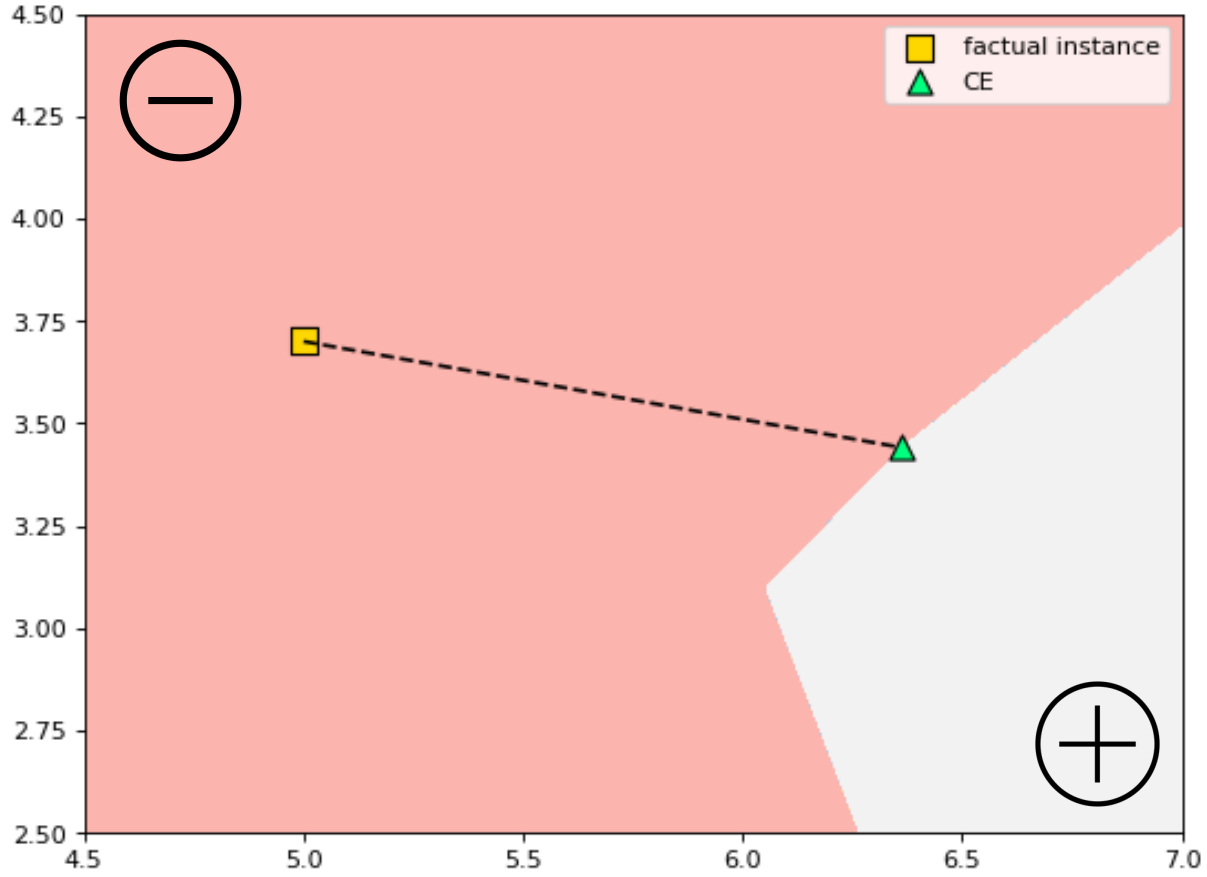


Counterfactual: set of features that should be changed in order to flip a model's prediction

Mathematical Model

$$\tilde{\boldsymbol{x}} = arg \min_{\boldsymbol{x} \in \mathbb{R}^n} d(\boldsymbol{x}, \hat{\boldsymbol{x}})$$
$$s.t. \quad h(\boldsymbol{x}) = \tilde{y}$$

$\boldsymbol{d}(.,.)$: distance function
$\boldsymbol{h}(.,.)$: trained model
$\hat{\boldsymbol{x}}$: factual instance
$\tilde{\boldsymbol{x}}$: counterfactual explanation
$\tilde{y}$: desired outcome



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	●	●	●	●	●	●	●

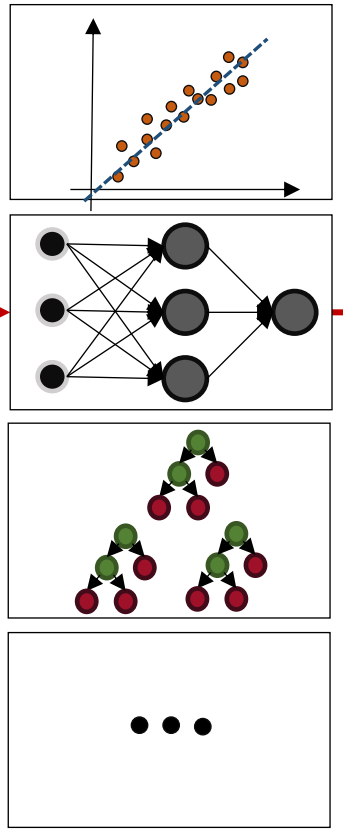
●: addressed; ◐: partially addressed; –: absent

“CE“

Optimization with Constraint Learning (OCL)

Maragno et al. (2021)

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



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

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

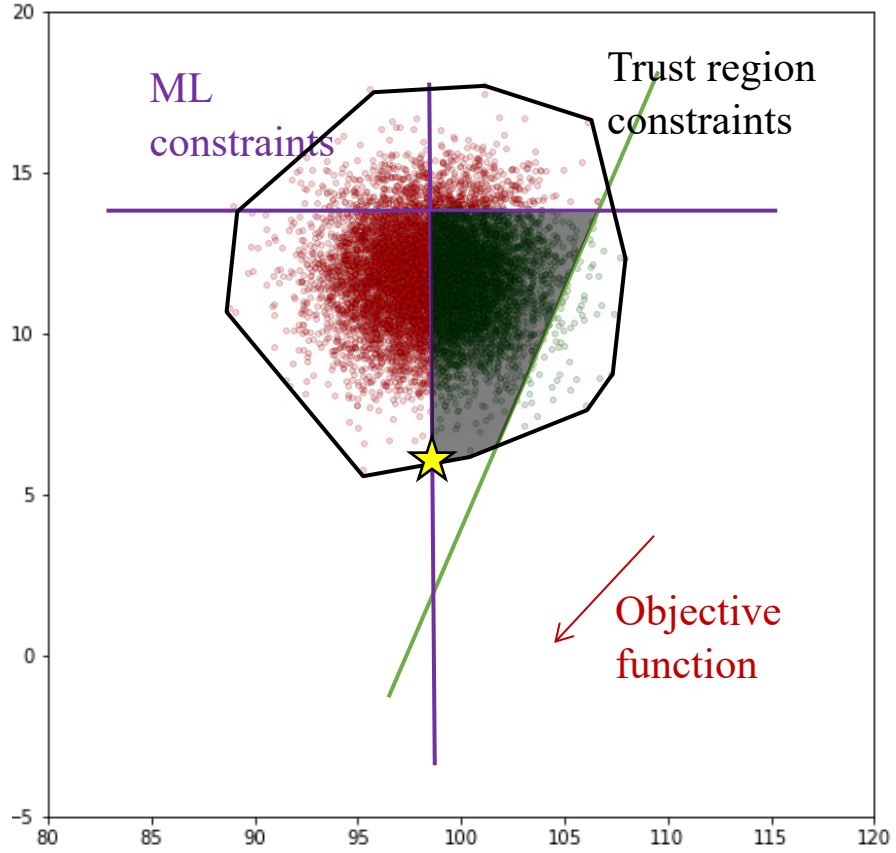
$$y = \hat{h}_D(x, w)$$

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

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

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

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

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

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

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

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 $h_{\mathcal{D}}(\boldsymbol{x}) = \tilde{y}$

Proximity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

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

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

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

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

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

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

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

Proximity

Validity

Causality

Data manifold closeness

Diversity

OCL Model

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

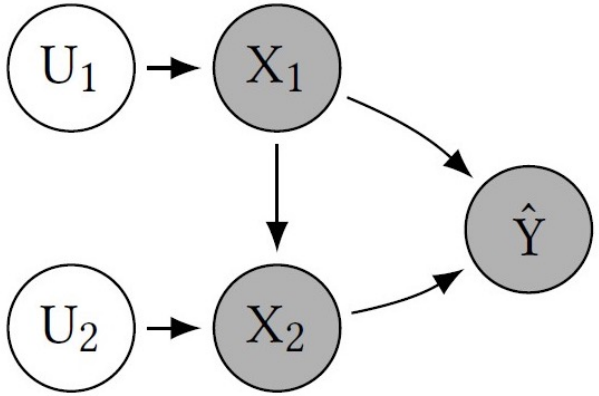
$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \boldsymbol{0}$$

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

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

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

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



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

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

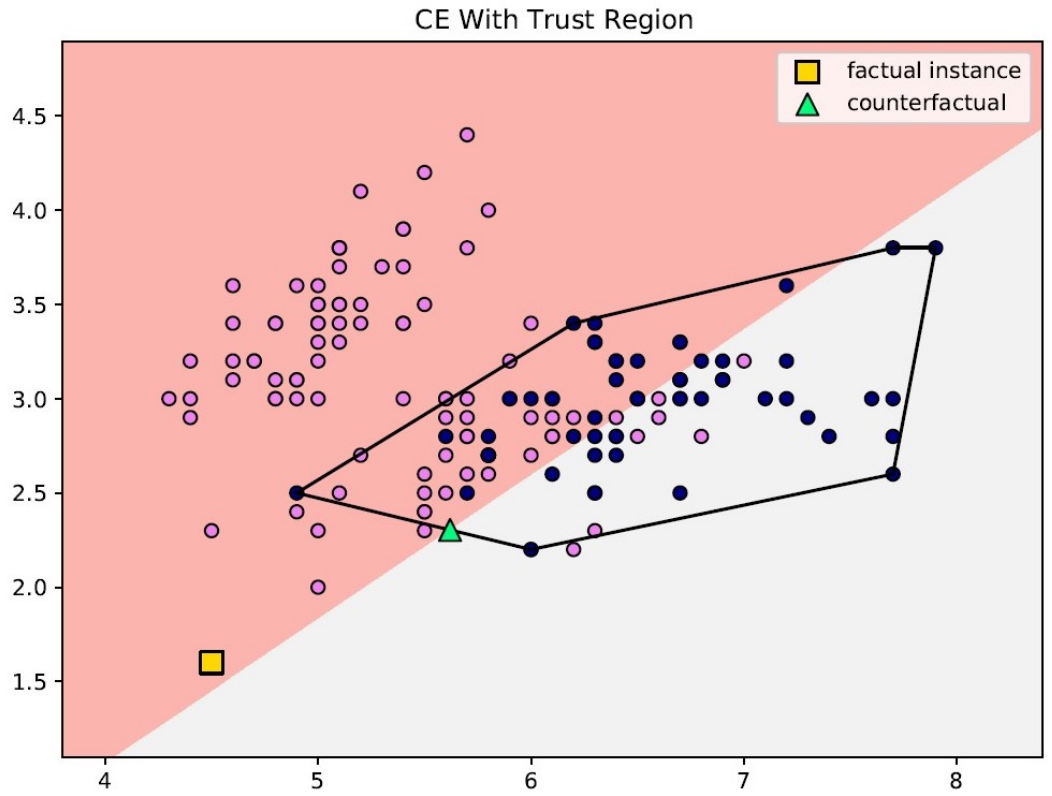
$$\text{s.t. } \boldsymbol{g}(\boldsymbol{x},\boldsymbol{w},\boldsymbol{y}) \leq \mathbf{0}$$

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

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

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

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



Data manifold closeness

Proximity

Validity

Coherence

Sparsity

tionability

Causality

Diversity

OCL Model

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

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

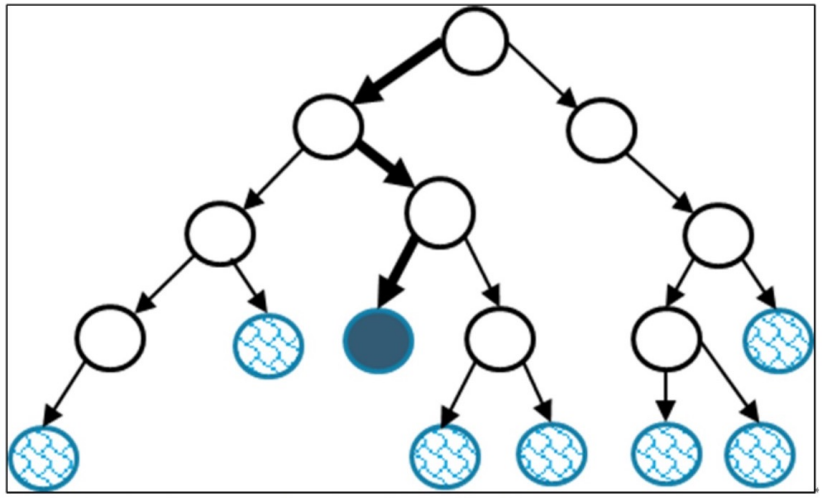
$$\boldsymbol{y} = \hat{\boldsymbol{h}}_{\mathcal{D}}(\boldsymbol{x},\boldsymbol{w})$$

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

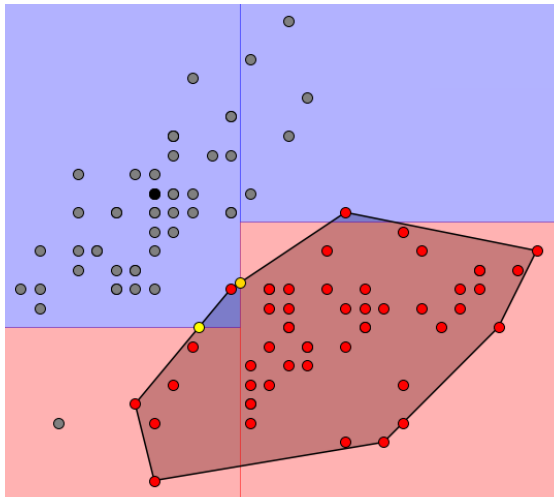
$$\sum_{i}^{N} \lambda_i = 1$$

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

Incumbent solutions



source



Diversity

Proximity

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Case Study

University of Amsterdam
Amsterdam Business School

OptiCL:

A Python Package for Optimization with Constraint Learning

Codes and Examples

https://github.com/hwiberg/OptiCL

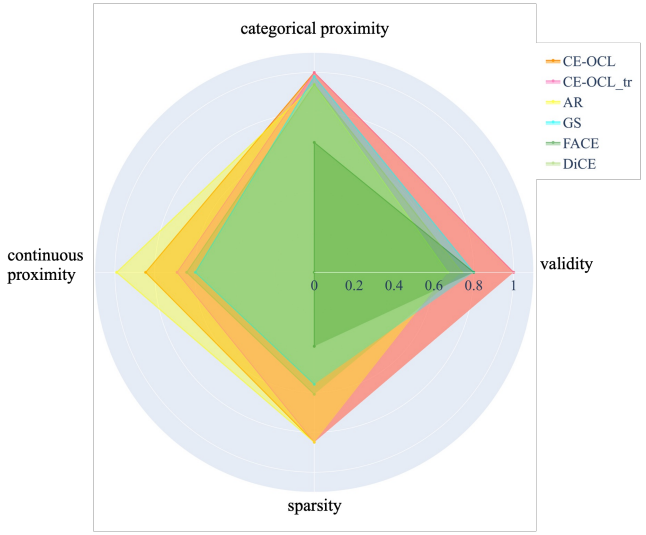
y: good or bad credit risk?

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

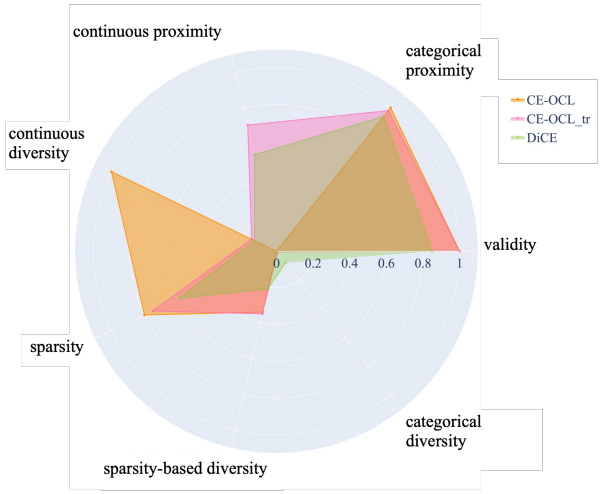
	F1	F2	F3	F4	F5	F6	F7	F8*	F9*	F
$\hat{x}$	24.0	1371.26	4.0	25.0	4.0	1.0	1.0	A	A	1
Part A: validity, proximity, coherence										
(a)	16.48	-57.75	3.88	26.71						
Part B: validity, proximity, coherence, sparsity										
(a)	7.12	–	–	–						
Part C: validity, proximity, coherence, sparsity, diversity										
(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	–	–						

Comparing Against Other Approaches

30 factual instances, one CE each



30 factual instances, three CEs each



arXiv > cs > arXiv:2209.10997

Counterfactual Explanations Using Optimization With Constraint Learning

Donato Maragno*
University of Amsterdam
d.maragno@uva.nl

Tabea E. Röber*
University of Amsterdam
t.e.rober@uva.nl

Ş. İlker Birbil
University of Amsterdam
s.i.birbil@uva.nl

https://github.com/tabearoeber/CE-OCL

Extensions

University of Amsterdam
Amsterdam Business School

- ϵ-convex hull for data manifold closeness
- Intervals for infinitely-many CEs
- User study via dedicated webpage
- Food for thought: Adversarial attacks