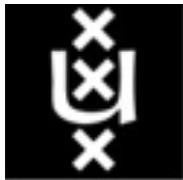


Counterfactual Explanations Using Optimization with Constraint Learning

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 gigascale breakthrough in solving protein structures

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

The New York Times

How A.I. Can Help Handle

The World Economic Forum

Why artificial intelligence is vital in the future of work

Why artificial intelligence is vital in the future of work

an ethical framework for AI

1 week ago

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

High-level detection of skin cancer

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



Artificial intelligence powers detection of skin cancer from images

PAGES 36 & 115

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](#)

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

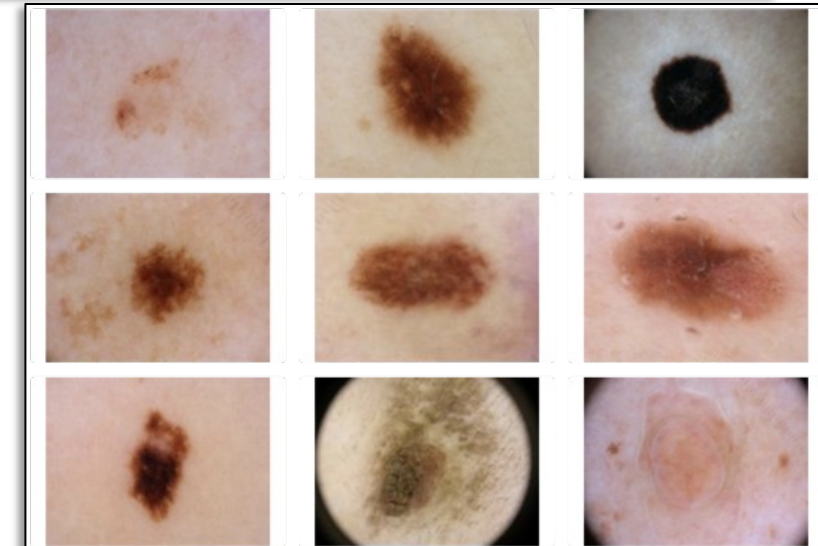
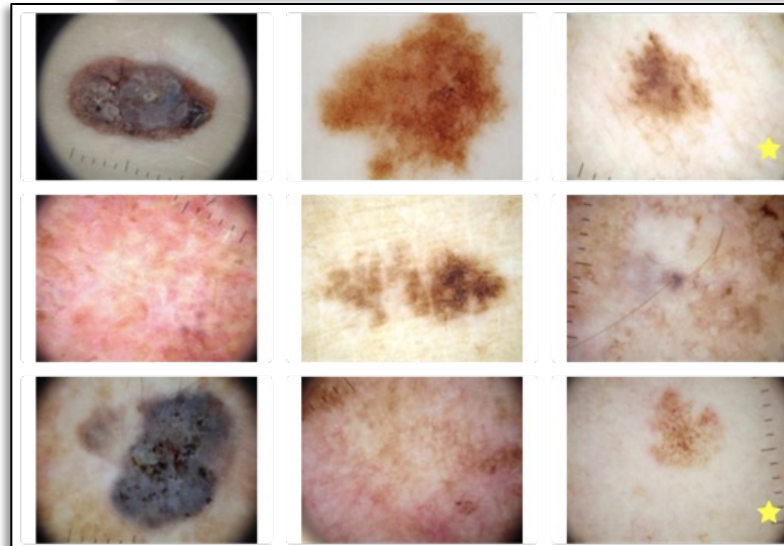
In-lab vs. Real-life deployment



Dermatologist-level classification 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.



“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^{***}

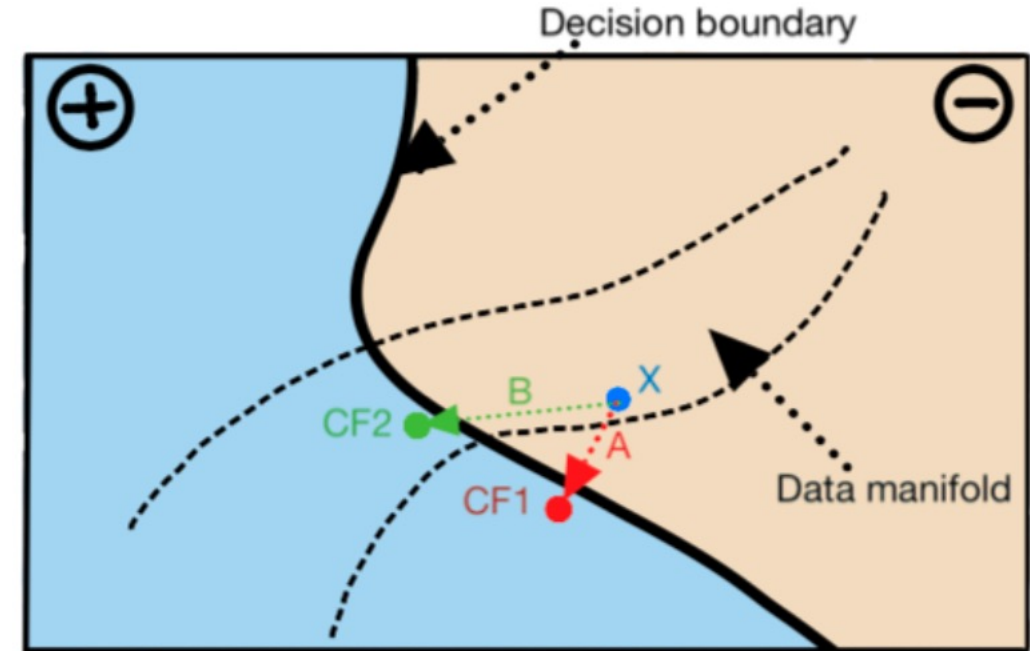
Harvard Journal of Law & Technology, 2018

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

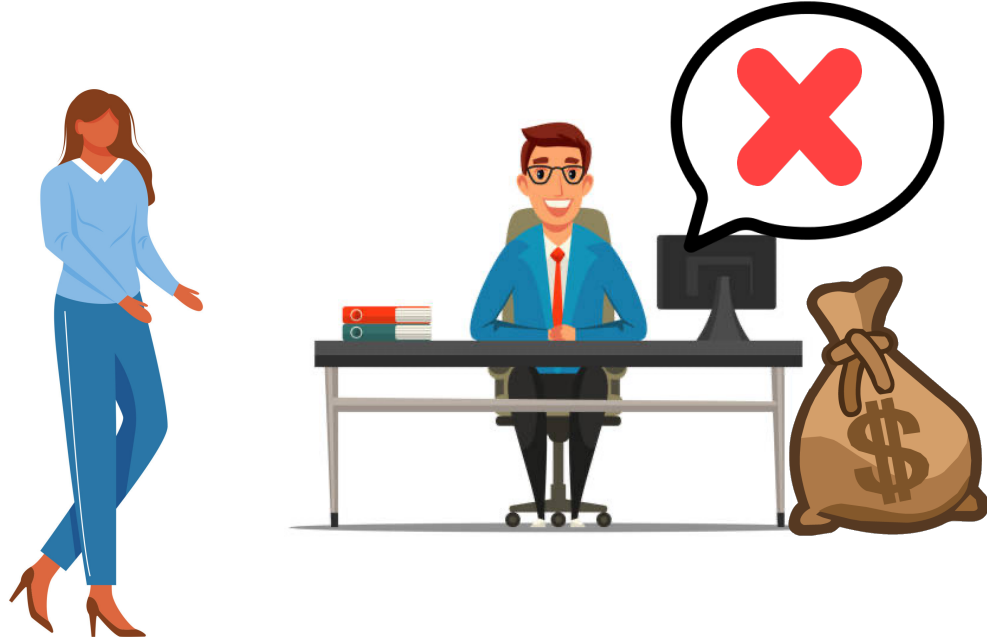
Keegan Hines
Arthur AI
keegan@arthur.ai

arXiv, 2020

Counterfactual Explanations

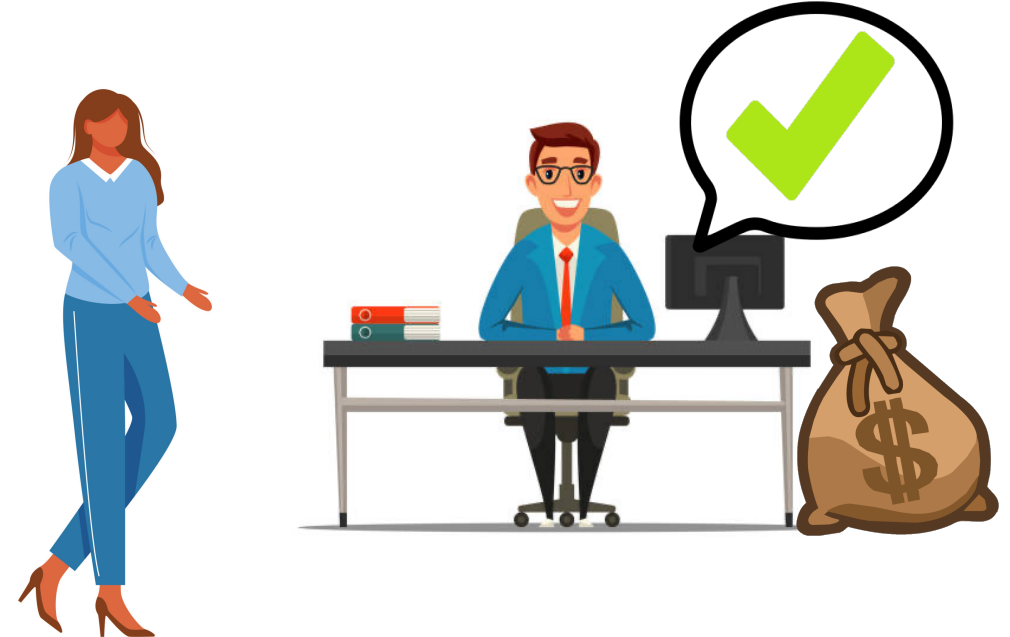
Factual Instance

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

$$\tilde{x} = \arg \min_{x \in \mathbb{R}^n} d(x, \hat{x})$$

$$s.t. \quad h(x) = \tilde{y}$$

$d(.,.)$: distance function

$h(.,.)$: **trained** model

\hat{x} : factual instance

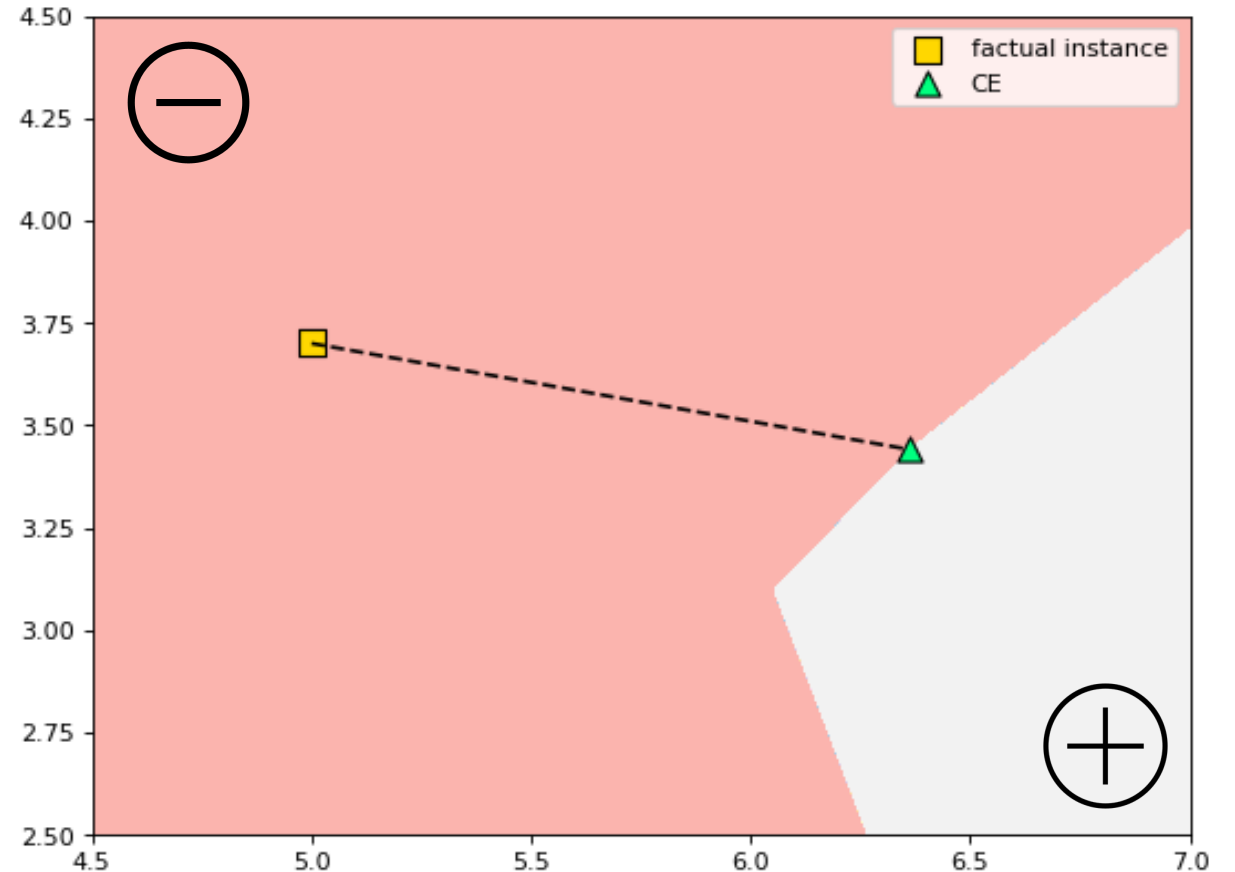
\tilde{x} : counterfactual explanation

\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)

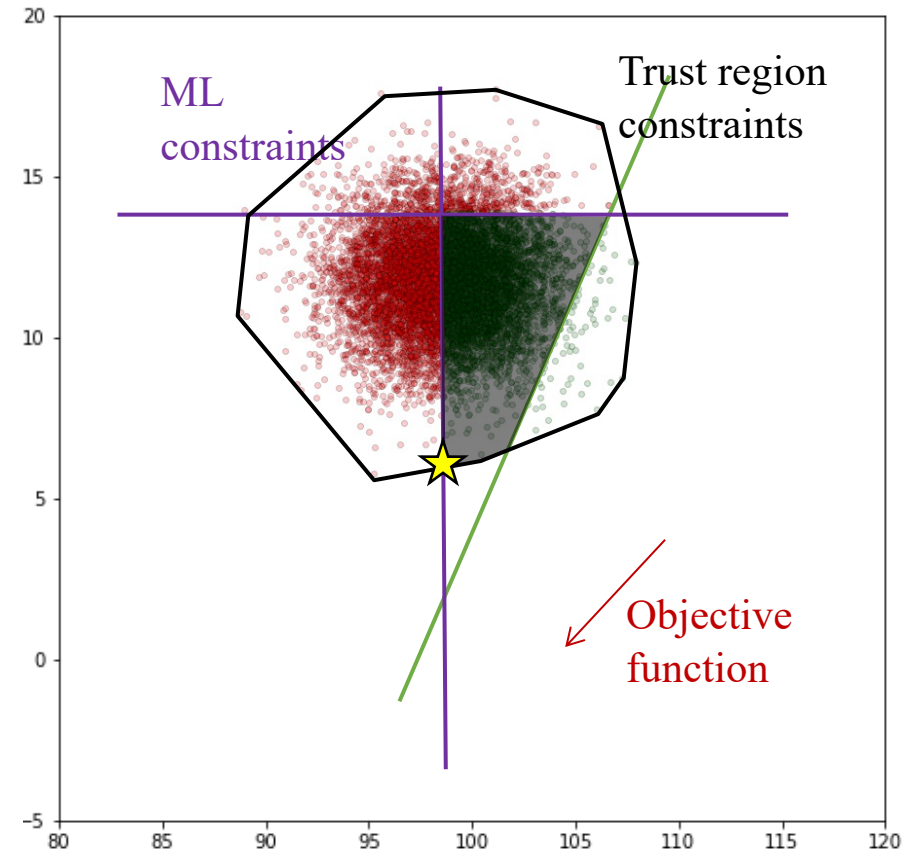
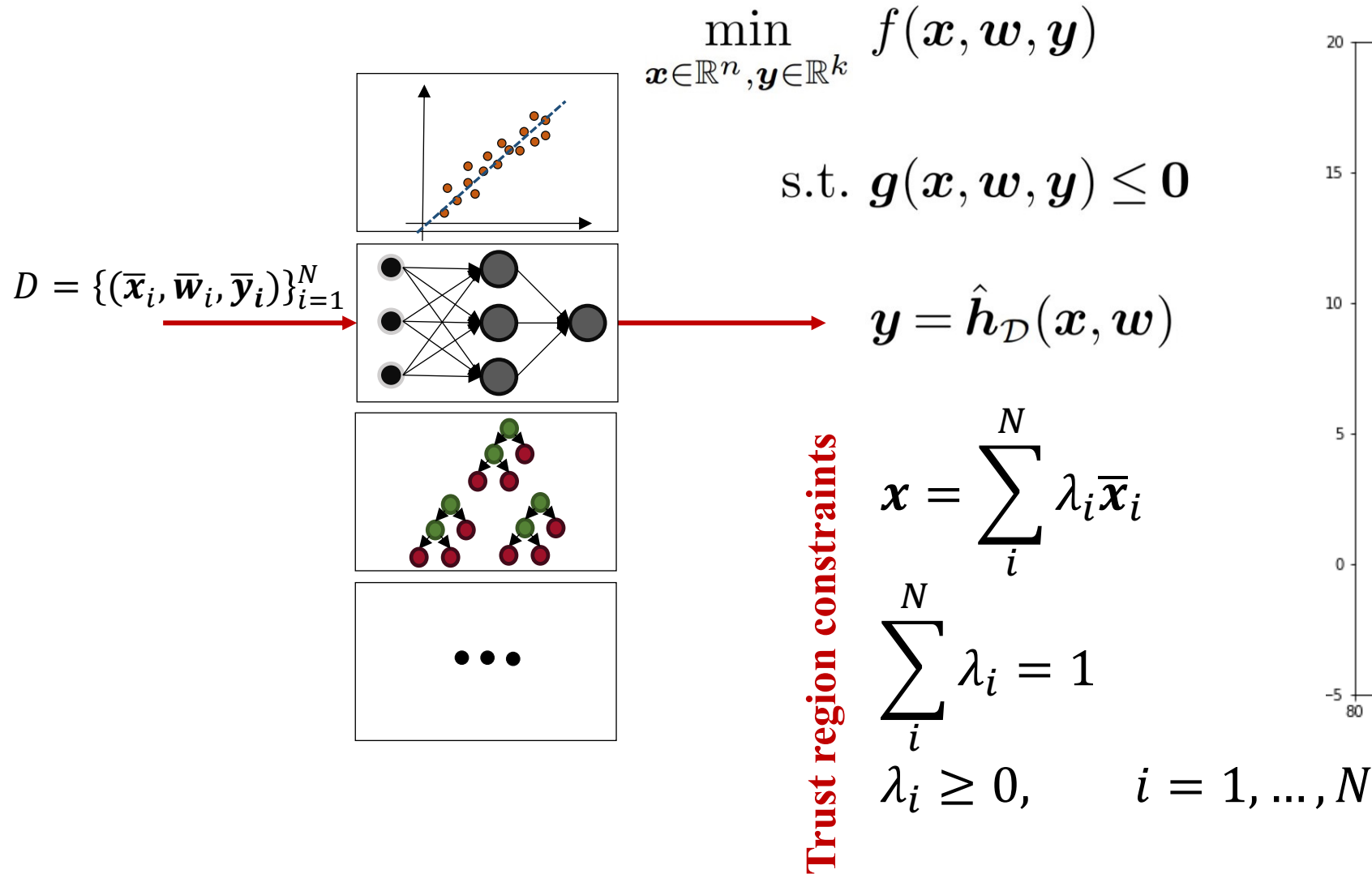
	Proximity	Sparsity	Coherence	Actionability	Data Manifold Closeness	Causality	Diversity
Russell [2019]	●	◐	●	-	-	-	●
Ustun et al. [2019]	●	●	●	●	-	-	-
Kanamori et al. [2020]	●	-	●	-	●	-	-
Mahajan et al. [2019]	●	-	●	◐	●	●	-
Karimi et al. [2021]	●	-	●	-	-	●	-
Kanamori et al. [2021]	●	●	●	●	-	●	●
Mothilal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

●: addressed; ◐: partially addressed; -: absent

“CE-OCL (OCL) is the most principled and effective method for generating CE’s”

Optimization with Constraint Learning (OCL)

Maragno et al. (2021)



OCL Model

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

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

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

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

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

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

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

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

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

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

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

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

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

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

Validity

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

Proximity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Diversity

OCL Model

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

Proximity

Validity

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

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

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

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

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

Coherence

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

Sparsity

$$\|\mathbf{x} - \hat{\mathbf{x}}\|_0 \leq K$$

Actionability

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

Causality

Data manifold closeness

Diversity

OCL Model

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

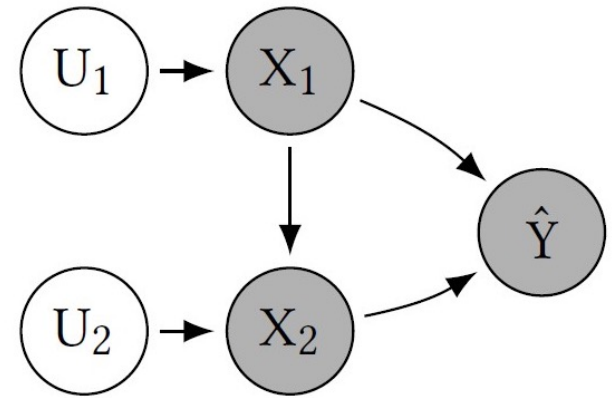
$$\text{s.t. } \mathbf{g}(\mathbf{x}, \mathbf{w}, \mathbf{y}) \leq \mathbf{0}$$

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

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

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

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



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

[Karimi et al. \(2020\)](#)

Data manifold closeness

Proximity

Validity

Coherence

Sparsity

Actionability

Diversity

OCL Model

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

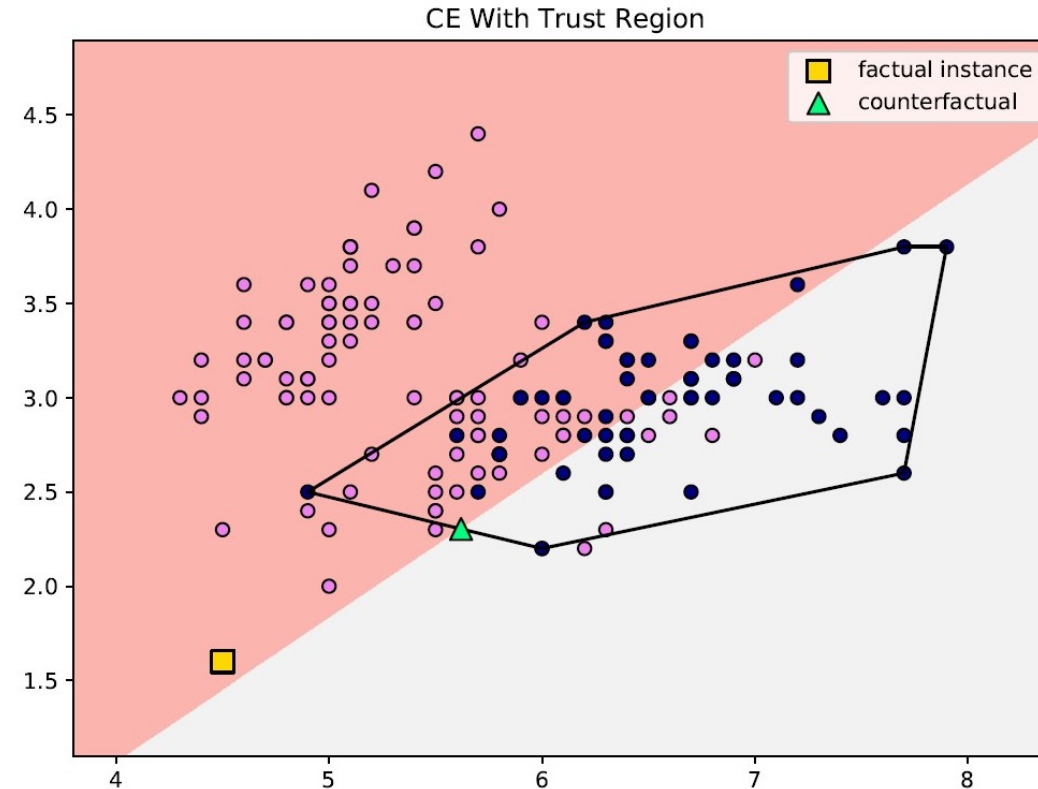
$$\text{s.t. } \mathbf{g}(\mathbf{x}, \mathbf{w}, \mathbf{y}) \leq \mathbf{0}$$

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

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

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

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



Data manifold closeness

Proximity

Validity

Coherence

Sparsity

tionability

Causality

Diversity

OCL Model

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

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

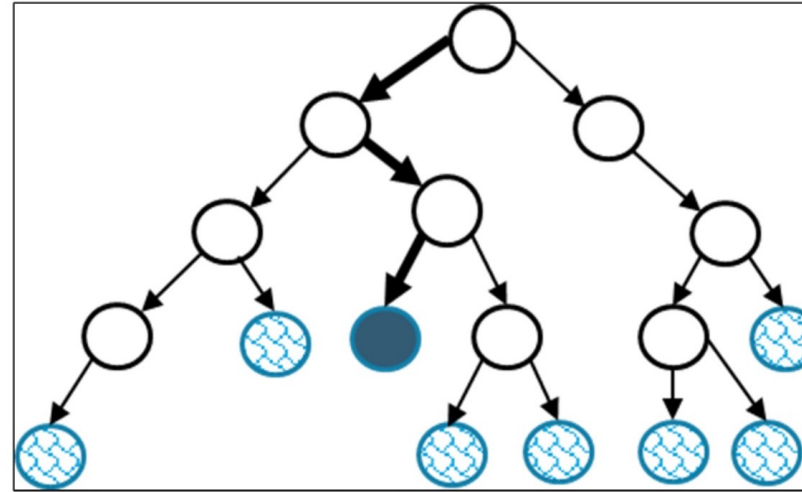
$$\mathbf{y} = \hat{\mathbf{h}}_{\mathcal{D}}(\mathbf{x}, \mathbf{w})$$

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

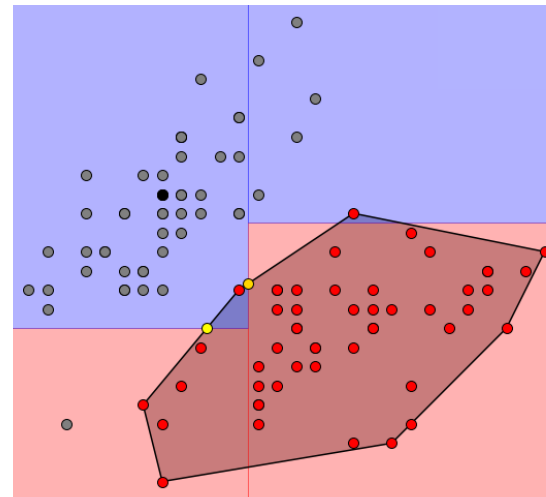
$$\sum_i^N \lambda_i = 1$$

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

Incumbent solutions



source



Diversity

Proximity

Validity

Coherence

Sparsity

Actionability

Causality

Data manifold closeness

Case Study

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*	F10*
\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	-	-						

OptiCL:

A Python Package for
Optimization with Constraint
Learning

Codes and Examples

<https://github.com/hwiberg/OptiCL>

Next Steps: Enlarged Trust Region

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

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

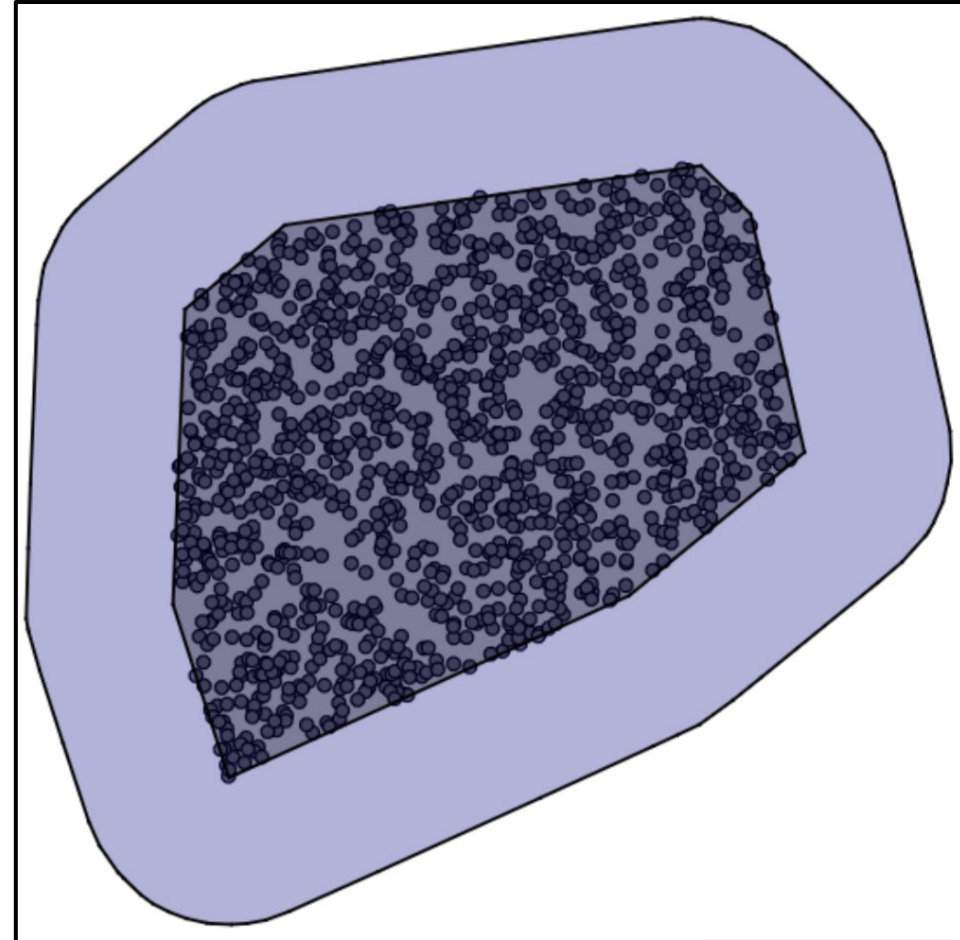
$$\lambda_i \geq 0, \quad i = 1, \dots, N$$

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

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

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

$$\|\mathbf{s}\|_p \leq \epsilon$$



Next Steps: Intervals of CEs

$$\tilde{x} = \arg \min_{x \in \mathbb{R}^n} d(x, \hat{x})$$

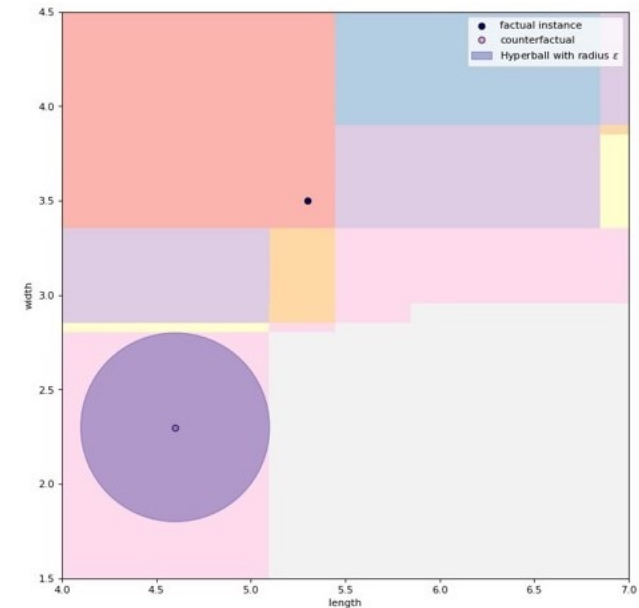
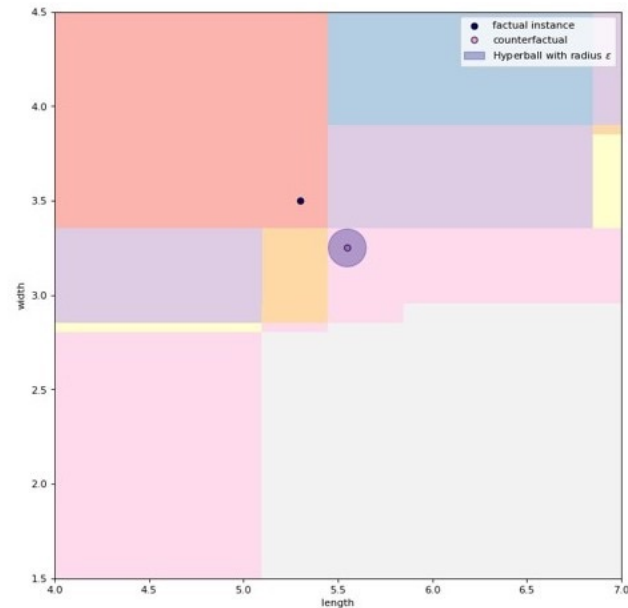
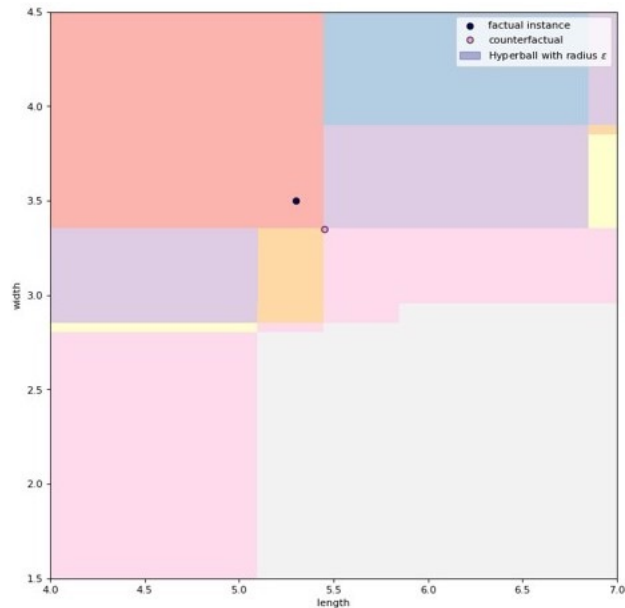
$$s.t. h(x) = \tilde{y}$$

$$\tilde{x} = \arg \min_{x \in \mathbb{R}^n} d(x, \hat{x})$$

$$s.t. h(x + s) = \tilde{y}, s \in S$$

“Uncertainty Set”

Example: $S = \{s \in \mathbb{R}^n : \|s\|_p \leq \epsilon\}$



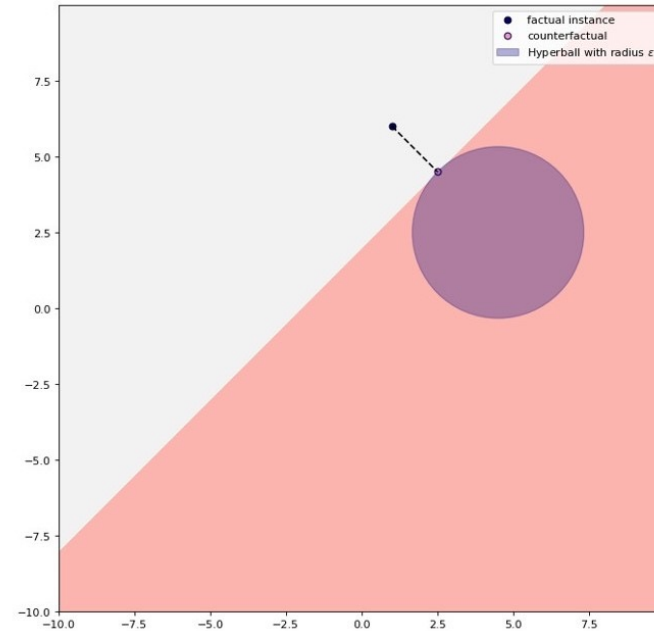
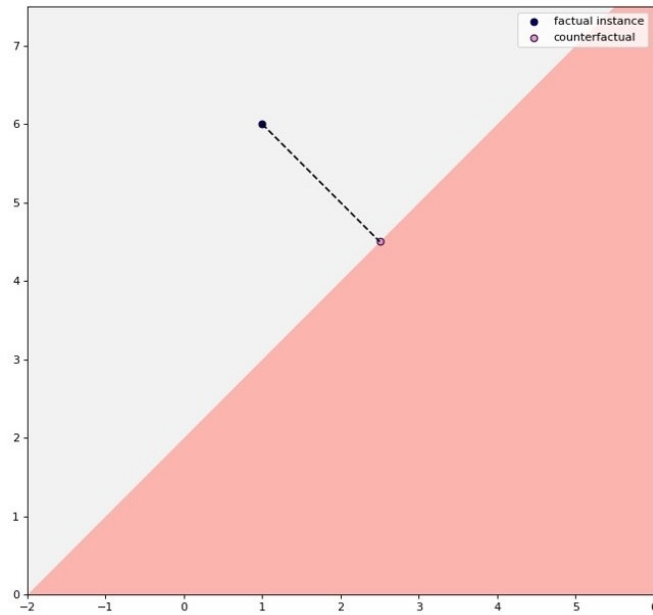
Next Steps: Intervals of CEs

$$h(\mathbf{x}) = \tilde{y}$$

$$h(\tilde{\mathbf{x}} + \mathbf{w}) = \tilde{y}, \quad \mathbf{w} \in W$$

$$\max_{\mathbf{w}} g(W)$$

$$s. t. \quad h(\tilde{\mathbf{x}} + \mathbf{w}) = \tilde{y}, \quad \mathbf{w} \in W, \\ \tilde{\mathbf{x}} \in W$$





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



Apple Card ✓
@AppleCard

Well, if your wife would have had a 20+ years relationship with our bank, and would have been regarded as Premium customer at some point in time, she would also receive a 20x credit limit.

Conclusion

- A general framework for CEs
- ϵ -convex hull for data manifold closeness
- Intervals for infinitely-many CEs

- Additional case studies
- Comparison against other methods
- User study via dedicated webpage

- Food for thought: Adversarial attacks

The screenshot shows the CE-OCL web application interface for a Diabetes decision tree model. The interface includes a navigation bar with links for Home, Demo, Questionnaire, Documentation, About, FAQ, and Contact us. The main content area is titled "Diabetes" and "Decision tree". It features a grid of input fields for various features, each with a toggle switch for "Off". The input values are: Pregnancies (1), Glucose (85), BloodPressure (66), SkinThickness (29), Insulin (0), BMI (26.6), DiabetesPedigreeFunction (0.351), and Age (31). Below the input fields, the "Model prediction" is shown as 0. There are three toggle switches: "Sparsity" (On), "Data manifold" (Off), and "Robustness" (0.6). A "Generate counterfactual explanation" button is located below the model prediction. At the bottom, there is a grid of sliders for each feature, showing the current value and a range of possible values: Pregnancies (13.70002), Glucose (273.9002), BloodPressure (66), SkinThickness (29), Insulin (0), BMI (26.6), DiabetesPedigreeFunction (1.7477), and Age (31).