Explainability AI: Counterfactual/Actionable Explanations

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

Summer School on Data Science
February 4th, 2020 - Rio de Janeiro - Brazil

Why explain?

A) Explanations to UNDERSTAND decisions

B) Explanations to CONTEST decisions

C) Explanations to ALTER FUTURE decisions

Material based on the course CS282BR: Topics in Machine Learning Interpretability and Explainability at Harvard. Lectured by Hima Lakkaraju and Ike Lage. `canvas.harvard.edu/courses/68154`

GDPR

GDPR

What? “The General Data Protection Regulation (GDPR) codifies and unifies the data privacy laws across all the EU member countries.”

Who? “The GDPR is applicable to any citizen of the European Union and, most importantly, for any company doing business with a citizen of the EU.”

Why care? “the penalties laid out for violations are significant. Enterprises found to be in violation of the provisions of the GDPR can be fined up to 4turnover or 20 Million Euros, whichever is greater.”

When? “Enforcement of the GDPR went into effect May 25, 2018.”

GDPR - Main provisions

Informed Consent (intelligible, clear, easy to withdraw)

Rights:

- Breach notifications
- Right to access and information
- Right of erasure, rectification
- Data portability
- Contest automated decisions

Principles:

- Data minimization
- Security

GDPR - Individual's Rights

The GDPR establishes the following rights for individuals: The right to be informed, access, rectification, erasure, restrict processing, data portability, object.

Rights in relation to automated decision making and profiling - right to explanation.

Counterfactual explanations

What is a counterfactual explanation?

Readable explanations - "If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5, you would have a score of 0.51."

Flipset explanations:

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	\$3,250	⟶	\$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Figura 1: Example of set of changes for a original sample, on the left, leading to a new state, on the right, that achieves the desired outcome.

Counterfactual explanation to solve GDPR [Wachter et al., 2017]

Their two main arguments:

1. The GDPR does not require “opening the black box”.
2. Counterfactual explanations fulfill (and go beyond) the requirements of the GDPR.

Why use counterfactual explanation?

The usual approach to explanation: Focuses primarily on an explanation of the internal structure of the algorithms and how it led to the decisions.

Counterfactual approach to explanation: Describes dependency on the external facts that led to the decision.

Mathematical definition

Let's suppose a learning machine $f(\theta, \mathbf{x})$:

- $f(\bullet)$ - is the decision function.
- θ - is the parameter vector, already adjusted to a dataset.
- $\mathbf{x}$ - is a sample.

a conterfactual explanation consists in a synthetic sample $\mathbf{x}'$ that achieves a desired outcome y' in similarity $f(\theta, \mathbf{x}') \approx y'$ or constraint $f(\theta, \mathbf{x}') \geq y'$.

Important property: reduce the cost $c(\bullet)$ of changing an instance. So, $\min c(\mathbf{x}, \mathbf{x}')$.

Creating counterfactual explanations

How to achieve a counterfactual explanation? [Wachter et al., 2017]

We want to find a new outcome $f(\theta, \mathbf{x}')$ as close as possible to y', then: $\min(f(\theta, \mathbf{x}') \geq y')^2$.

We want find minimal change on sample, then: $\min d(\mathbf{x}', \mathbf{x})$.

$$\min_{\mathbf{x}'} \max_{\lambda} \lambda(f(\theta, \mathbf{x}') \geq y')^2 + d(\mathbf{x}', \mathbf{x}) \tag{1}$$

$d(\mathbf{x}', \mathbf{x})$ can be:

- $\sum_k (x_k - {x'}_k)^2$.
- $\sum_k \frac{(x_k - {x'}_k)^2}{\sigma_k}$.
- $\sum_k \frac{|x_k - {x'}_k|}{MAD_k}$, MAD_k is median deviation of the median – equivalent to standard deviation.

LSAT dataset

				Unnormalised L2					
	Original Data			Counterfactuals			Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	39.0	0.3	1.5	38.4	0
0.54	3.7	48.0	0	3.5	47.9	0.9	-1.6	45.9	0
-0.77	3.3	28.0	1	3.5	28.1	-0.3	5.3	28.9	0
-0.83	2.4	28.5	1	2.6	28.6	-0.4	4.8	29.4	0
-0.57	2.7	18.3	0	2.9	18.4	-1.0	8.4	20.6	0

				Normalised L2					
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.0	37.0	0.2	3.0	34.0	0
0.54	3.7	48.0	0	3.5	39.5	0.4	3.5	33.1	0
-0.77	3.3	28.0	1	3.5	39.8	0.4	3.4	33.4	0
-0.83	2.4	28.5	1	2.7	37.4	0.2	2.6	35.7	0
-0.57	2.7	18.3	0	2.8	28.1	-0.4	2.9	34.1	0

				Normalised L1					
	Original Data			Counterfactuals Continuous			Counterfactual Hybrid		
score	GPA	LSAT	Race	GPA	LSAT	Race	GPA	LSAT	Race
0.17	3.1	39.0	0	3.1	35.0	0.1	3.1	34.0	0
0.54	3.7	48.0	0	3.7	33.5	0.0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	34.4	0.1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	39.3	0.2	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	35.8	0.1	2.7	34.9	0

Figura 2: Representation of three experiments showing possible counterfactual explanations to the LSAT dataset.

Readable explanations - LSAT

1. If your LSAT was 34.0, you would have an average predicted score.
2. If your LSAT was 32.4, you would have an average predicted score.
3. If your LSAT was 33.5, and you were 'white', you would have an average predicted score.
4. If your LSAT was 35.8, and you were 'white', you would have an average predicted score.
5. If your LSAT was 34.9, you would have an average predicted score.

Readable explanations - Pima diabetes

1. If your 2-Hour serum insulin level was 154.3, you would have a score of 0.51.
2. If your 2-Hour serum insulin level was 169.5, you would have a score of 0.51.
3. If your Plasma glucose concentration was 158.3 and your 2-Hour serum insulin level was 160.5, you would have a score of 0.51.

Other approaches

Other approaches - Single changes [Krause et al., 2016]

Patient: 3530 Truth: 1 Original: 0.42753

Decreasing Risk:

Feature	Current	Suggested Change
bmi (count) vital (bmi)	0	1 (0.08021)
eGFR lab	59.18887	59.59549 (0.23110)
bmi vital (bmi)	28.27873	28.23937 (0.27954)
eGFR (count) lab	0	1 (0.28705)
Calcifediol (Vit D) (25-O…	0	1 (0.31857)

Increasing Risk:

Feature	Current	Suggested Change
BUN (count) lab	0	1 (0.77246)
Peripheral Vascular Dis…	0	1 (0.68666)
Uric Acid (count) lab	0	1 (0.64202)
Calcium lab	9.37486	9.38749 (0.59175)
Carbon Dioxide lab	26.56109	27.35469 (0.59025)

Figura 3: Impact of changing single features on the risk of developing diabetes.

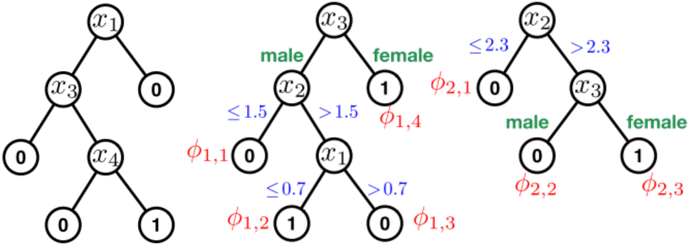
Other approaches - Inverse classification

Searches in the datasets the set of features (also interpreted as actions) that achieves the desired class.

Use of the Gini index to find a set of features that describe samples with high Gini index, enough support (number of samples) and have the desired class as dominant [Aggarwal et al., 2010].

Use of greedy changes (changes that increase the probability of the desired class) using KNN as classifier [Yang et al., 2012].

Other approaches - inspecting trees



(a) Single tree **(b)** Additive trees

Figura 4: Representation of trees.

Other approaches - inspecting trees

- Greedy algorithms [Yang et al., 2003, Yang et al., 2007],
- Mixed linear-integer formulation of the swaps between leaves of the trees [Cui et al., 2015],
- A^*-like search [Lu et al., 2017, Lv et al., 2018].

Counterfactual explanations in linear classification

Actionable recourse in linear classification [Ustun et al., 2019]

$$
\begin{aligned}
\min_{\mathbf{a}} \quad & cost(\mathbf{a}; \mathbf{x}) \\
s.t. \quad & f(\mathbf{x} + \mathbf{a}) = 1 \\
& a \in A(\mathbf{x}).
\end{aligned} \tag{2}
$$

$A(\mathbf{x})$ is the set of possible actions of $\mathbf{x}$,

$cost(\mathbf{a}; \mathbf{x})$ have to increase with the increase of $\mathbf{a}$.

Linear-integer model [Ustun et al., 2019]

$$
\begin{aligned}
\min_{\mathbf{a}} \quad & \sum_j \sum_k c_{jk} v_{jk} \\
s.t. \quad & \sum_j w_j a_j + \sum_j w_j x_j \geq 0 \\
& a_j = \sum_k a_{jk} v_{jk}, \ \forall j \\
& 1 = u_j + \sum_k v_{jk}, \ \forall j \\
& a_j \in \mathbb{R}, \ \forall j \\
& u_j \in \{0, 1\}, \ \forall j \\
& v_{jk} \in \{0, 1\}, \ \forall i, \ \forall j
\end{aligned}
\tag{3}
$$

Example of an explanation [Ustun et al., 2019]

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Figura 5: Example of set of changes for a original sample, on the left, leading to a new state, on the right, that achieves the desired outcome.

Results [Ustun et al., 2019]

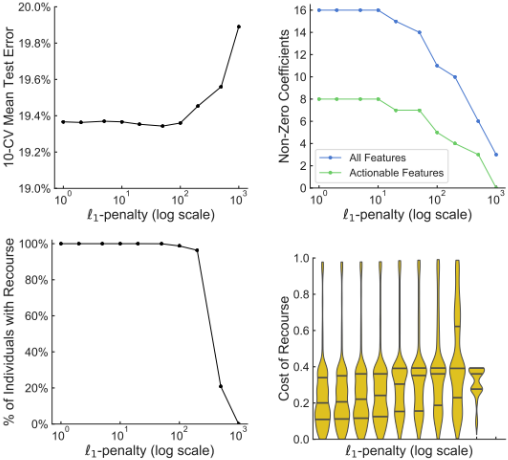


Figura 6: Impact of increasing the l_1-penalty of the test error, on the number of non-zero coefficients in the model, on the number of individuals with recourse and in the cost of the recourse.

Counterfactual Explanations Limitations [Rudin, 2019]

- Counterfactual explanations show the easiest change to the user.
- But we don't know, for sure, if this explanation is, in fact, easy.
- The function that depicts the cost to the user is hard to design. And it is hard to help the user to design a personalized cost function.

"For that reason, it is unclear that counterfactual explanations would suffice for high stakes decisions." [Rudin, 2019]

Ok, but if we enumerate a diverse set of explanations?

Enumerating possible explanations [Ustun et al., 2019]

Feature Subset	Current Values		Required Values
MostRecentPaymentAmount	\$0	⟶	\$790
MostRecentPaymentAmount	\$0	⟶	\$515
MonthsWithZeroBalanceOverLast6Months	1	⟶	2
MonthsWithZeroBalanceOverLast6Months	1	⟶	4
MostRecentPaymentAmount	\$0	⟶	\$775
MonthsWithLowSpendingOverLast6Months	6	⟶	5
MostRecentPaymentAmount	\$0	⟶	\$500
MonthsWithLowSpendingOverLast6Months	6	⟶	5
MonthsWithZeroBalanceOverLast6Months	1	⟶	2

Figura 7: Example of sets of feature changes that change the outcome.

Multi-objective optimization

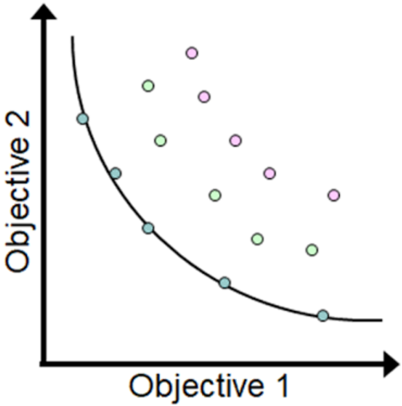


Figura 8: Representation of the objetive space for a multi-objetive optimization problem.

Enumerating multi-objective options

Here we have two objectives, number of changes and cost of the change.

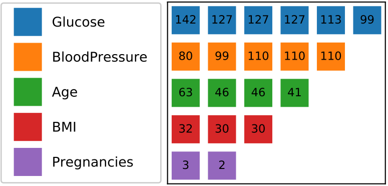


Figura 9: Representation enumerated actions using multi-objective optimization.

Enumerating multi-objective options

Here we considered the intensity of the change of every action as the objective.

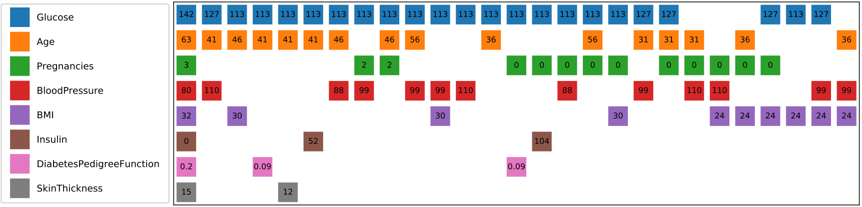


Figura 10: Representation enumerated actions using multi-objective optimization.

Explainability AI: Counterfactual/Actionable Explanations

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

Summer School on Data Science
February 4th, 2020 - Rio de Janeiro - Brazil

References i

Aggarwal, C. C., Chen, C., and Han, J. (2010).
The inverse classification problem.
Journal of Computer Science and Technology, 25(3):458–468.

Cui, Z., Chen, W., He, Y., and Chen, Y. (2015).
Optimal action extraction for random forests and boosted trees.
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2015-Augus:179–188.

Krause, J., Perer, A., and Ng, K. (2016).
Interacting with predictions: Visual inspection of black-box machine learning models.
Conference on Human Factors in Computing Systems - Proceedings, pages 5686–5697.

Lu, Q., Cui, Z., Chen, Y., and Chen, X. (2017).
Extracting optimal actionable plans from additive tree models.
Frontiers of Computer Science, 11(1):160–173.

Lv, Q., Chen, Y., Li, Z., Cui, Z., Chen, L., Zhang, X., and Shen, H. (2018).
Achieving data-driven actionability by combining learning and planning.
Frontiers of Computer Science, 12(5):939–949.

Rudin, C. (2019).
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature Machine Intelligence, 1(5):206–215.

Ustun, B., Spangher, A., and Liu, Y. (2019).
Actionable recourse in linear classification.
FAT 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 10–19.

Wachter, S., Mittelstadt, B., and Russell, C. (2017).
Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR.
SSRN Electronic Journal, pages 1–52.

Yang, C., Street, W. N., and Robinson, J. G. (2012).
10-year CVD risk prediction and minimization via inverse classification.
IHI'12 - Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium, pages 603–609.

Yang, Q., Yin, J., Ling, C., and Pan, R. (2007).
Extracting actionable knowledge from decision trees.
IEEE Transactions on Knowledge and Data Engineering, 19(1):43–55.

Yang, Q., Yin, J., Ling, C. X., and Chen, T. (2003).
Postprocessing decision trees to extract actionable knowledge.
Proceedings - IEEE International Conference on Data Mining, ICDM, pages 685–688.