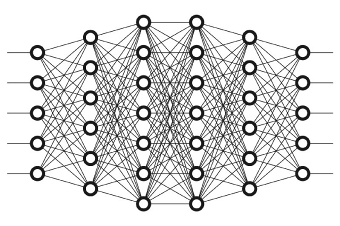
Local Explanations via Necessity and Sufficiency

Unifying Theory and Practice

David S. Watson, Limor Gultchin, Ankur Taly, Luciano Floridi

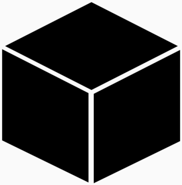
The Problem

Supervised learning algorithms are increasingly used in a variety of high-stakes domains, from credit scoring to medical diagnosis.



The Problem

However, many such methods are *opaque*, in that humans cannot understand the reasoning behind particular predictions.

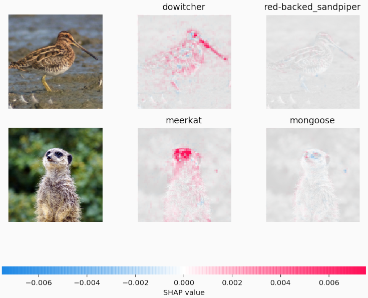


The Problem

uai2021

The last few years have seen an explosion of post-hoc model-agnostic tools for explainable artificial intelligence (XAI), e.g.

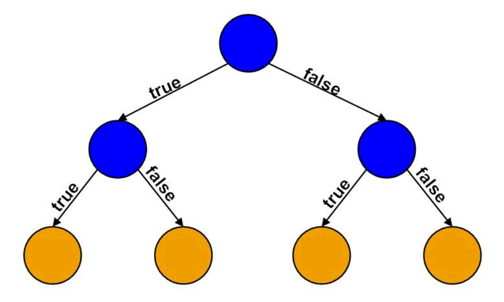
- **feature attributions** [12, 7, 17]



The Problem

The last few years have seen an explosion of post-hoc model-agnostic tools for explainable artificial intelligence (XAI), e.g.

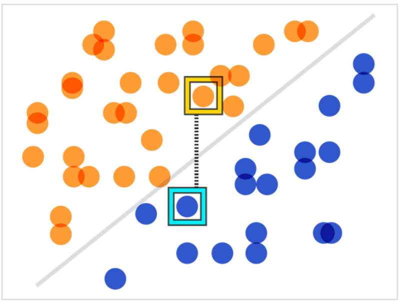
- **feature attributions**,
- **rule lists** [13, 6, 16]



The Problem

The last few years have seen an explosion of post-hoc model-agnostic tools for explainable artificial intelligence (XAI), e.g.

- **feature attributions**,
- **rule lists**, and
- **counterfactuals** [19, 3, 20].



The Problem

uai2021

Inconsistency

These tools are *mutually inconsistent* [8, 11, 2] and *often unreliable* [14, 5, 15].

The Problem

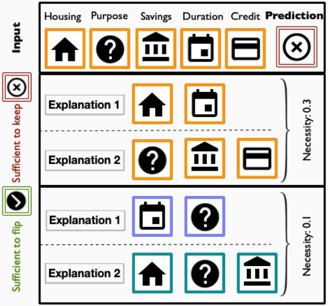
Inconsistency

These tools are *mutually inconsistent* [8, 11, 2] and *often unreliable* [14, 5, 15].

Lack of theory

Despite the proliferation of XAI methods, a dearth of theory persists.

The Solution



Necessity and *sufficiency* are the building blocks of all successful explanations, and therefore deserve a privileged position in the theory and practice of XAI.

The Solution

Our contributions:

The Solution

Our contributions:

Unifying Framework

A formal framework for XAI that unifies existing approaches.

The Solution

uai2021

Our contributions:

Unifying Framework

A formal framework for XAI that unifies existing approaches.

Explanatory Measures

Novel definitions of necessity and sufficiency that quantify the explanatory value of feature subsets.

The Solution

Our contributions:

Unifying Framework

A formal framework for XAI that unifies existing approaches.

Explanatory Measures

Novel definitions of necessity and sufficiency that quantify the explanatory value of feature subsets.

Method: LENS

An optimal algorithm for computing explanatory factors.

Logic

Necessity and sufficiency are expressed in logic via material implication.

If x is logically sufficient for y, then $LS_1(x, y) : x \rightarrow y$.

If x is logically necessary for y, then $LN_1(x, y) : x' \rightarrow y'$.

Necessity and sufficiency are expressed in logic via material implication.

If x is logically sufficient for y, then $LS_1(x, y) : x \rightarrow y$.

If x is logically necessary for y, then $LN_1(x, y) : x' \rightarrow y'$.

By law of contraposition, both formulae can be rewritten:

If x is logically sufficient for y, then $LS_2(x, y) : y' \rightarrow x'$.

If x is logically necessary for y, then $LN_2(x, y) : y \rightarrow x$.

Probability

uai2021

Necessity and sufficiency are expressed in probability via conditioning.

The probability that x is sufficient for y is $PS(x, y) : P(y \mid x)$.

The probability that x is necessary for y is $PN(x, y) : P(y' \mid x')$.

Probability

Necessity and sufficiency are expressed in probability via conditioning.

The probability that x is sufficient for y is $PS(x, y) : P(y \mid x)$.

The probability that x is necessary for y is $PN(x, y) : P(y' \mid x')$.

There is no probabilistic law of contraposition!

What about $P(x' \mid y')$ and $P(x \mid y)$?

Necessity and sufficiency are expressed in causality via counterfactuals [10].

The probability that x is causally sufficient for y is $CS(x, y) : P(y_x \mid x', y')$.

The probability that x is causally necessary for y is $CN(x, y) : P(y'_{x'} \mid x, y)$.

Framework

An explanatory *basis* $\mathcal{B} = \langle f, P, \mathcal{C}, \preceq \rangle$ is a tuple containing:

Framework

uai2021

An explanatory *basis* $\mathcal{B} = \langle f, P, \mathcal{C}, \preceq \rangle$ is a tuple containing:

- **Target function** $f : \mathcal{X} \mapsto \{0, 1\}$.

Framework

An explanatory *basis* $\mathcal{B} = \langle f, P, \mathcal{C}, \preceq \rangle$ is a tuple containing:

- **Target function** $f : \mathcal{X} \mapsto \{0, 1\}$.
- **Distribution** P on contexts $\boldsymbol{z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{W}$.

An explanatory *basis* $\mathcal{B} = \langle f, P, \mathcal{C}, \preceq \rangle$ is a tuple containing:

- **Target function** $f : \mathcal{X} \mapsto \{0, 1\}$.
- **Distribution** P on contexts $\boldsymbol{z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{W}$.
- **Factors** $\mathcal{C}$, a finite set in which each $c : \mathcal{Z} \mapsto \{0, 1\}$.

Framework

uai2021

An explanatory *basis* $\mathcal{B} = \langle f, P, \mathcal{C}, \preceq \rangle$ is a tuple containing:

- **Target function** $f : \mathcal{X} \mapsto \{0, 1\}$.
- **Distribution** P on contexts $\boldsymbol{z} \in \mathcal{Z} = \mathcal{X} \times \mathcal{W}$.
- **Factors** $\mathcal{C}$, a finite set in which each $c : \mathcal{Z} \mapsto \{0, 1\}$.
- **Partial ordering** $\preceq$ encodes preferences over $\mathcal{C}$.

Explanatory Measures

uai2021

Probability of Sufficiency

$$PS(c, y) := P(f(\mathbf{z}) = y \mid c(\mathbf{z}) = c)$$

Probability of Necessity

$$PN(c, y) := P(c(\mathbf{z}) = c \mid f(\mathbf{z}) = y)$$

Probability of Sufficiency

$$PS(c, y) := P(f(\mathbf{z}) = y \mid c(\mathbf{z}) = c)$$

Probability of Necessity

$$PN(c, y) := P(c(\mathbf{z}) = c \mid f(\mathbf{z}) = y)$$

Claim: The converse formulation (a) is more expressive than the inverse alternative, and (b) accords better with intuition.

Confusion matrix of labels (rows) and factors (columns), with accompanying definitions of the four fundamental explanatory probabilities.

	$c(\boldsymbol{z})$	
$f(\boldsymbol{z})$	c	c'
y	q_{11}	q_{10}
y'	q_{01}	q_{00}

$$PS(c,y) = q_{11}/(q_{11}+q_{01})$$
$$PN(c,y) = q_{11}/(q_{11}+q_{10})$$
$$PS(c',y') = q_{00}/(q_{10}+q_{00})$$
$$PN(c',y') = q_{00}/(q_{01}+q_{00})$$

Explanatory Measures

Confusion matrix of labels (rows) and factors (columns), with accompanying definitions of the four fundamental explanatory probabilities.

	$c(\boldsymbol{z})$	
$f(\boldsymbol{z})$	c	c'
y	q_{11}	q_{10}
y'	q_{01}	q_{00}

$$
\begin{aligned}
PS(c, y) &= q_{11}/(q_{11} + q_{01}) \\
PN(c, y) &= q_{11}/(q_{11} + q_{10}) \\
PS(c', y') &= q_{00}/(q_{10} + q_{00}) \\
PN(c', y') &= q_{00}/(q_{01} + q_{00})
\end{aligned}
$$

These are akin to common measures used to evaluate machine learning classifiers: precision, recall, negative predictive value, specificity.

Confusion matrix of labels (rows) and factors (columns), with accompanying definitions of the four fundamental explanatory probabilities.

	$c(\boldsymbol{z})$	
$f(\boldsymbol{z})$	c	c'
y	q_{11}	q_{10}
y'	q_{01}	q_{00}

$$PS(c, y) = q_{11}/(q_{11} + q_{01})$$
$$PN(c, y) = q_{11}/(q_{11} + q_{10})$$
$$PS(c', y') = q_{00}/(q_{10} + q_{00})$$
$$PN(c', y') = q_{00}/(q_{01} + q_{00})$$

These are akin to common measures used to evaluate machine learning classifiers: precision, recall, negative predictive value, specificity.

Note that $PN(c, y) = PS(c', y') \leftrightarrow q_{11} = q_{00}$.

Explanatory Measures

Example contingency table of loan application outcome by education level.

	BA	No BA	Total
Approved	5	10	15
Denied	45	40	85
Total	50	50	100

To what extent is college education necessary for loan approval?

Example contingency table of loan application outcome by education level.

	BA	No BA	Total
Approved	5	10	15
Denied	45	40	85
Total	50	50	100

To what extent is college education necessary for loan approval?

If we take necessity to be the *converse* of sufficiency, we have:

$$P(\text{BA} \mid \text{Approved}) = 5/(5+10) = 1/3.$$

Explanatory Measures

Example contingency table of loan application outcome by education level.

	BA	No BA	Total
Approved	5	10	15
Denied	45	40	85
Total	50	50	100

To what extent is college education necessary for loan approval?

If we take necessity to be the *converse* of sufficiency, we have:

$$P(\text{BA} \mid \text{Approved}) = 5/(5+10) = 1/3.$$

If we take necessity to be the *inverse* of sufficiency, we have:

$$P(\text{Denied} \mid \text{No BA}) = 40/(40+10) = 4/5.$$

Explanatory Measures

Example contingency table of loan application outcome by education level.

	BA	No BA	Total
Approved	5	10	15
Denied	45	40	85
Total	50	50	100

To what extent is college education necessary for loan approval?

If we take necessity to be the *converse* of sufficiency, we have:

$$P(\text{BA} \mid \text{Approved}) = 5/(5+10) = 1/3.$$

If we take necessity to be the *inverse* of sufficiency, we have:

$$P(\text{Denied} \mid \text{No BA}) = 40/(40+10) = 4/5.$$

Lacking a BA may be sufficient for loan denial, but having a BA is not necessary for loan approval!

Unification

Feature attributions

Shapley values are a popular feature attribution method [7, 17, 1].

$$f(\mathbf{x}) = \sum_{j=0}^{d} \phi_v(j, \mathbf{x})$$

Feature attributions

Shapley values are a popular feature attribution method [7, 17, 1].

$$f(\mathbf{x}) = \sum_{j=0}^{d} \phi_v(j, \mathbf{x})$$

They are defined with respect to a characteristic function.

$$v(S, \mathbf{x}) = \mathbb{E}\big[f(\mathbf{x}) \mid \mathbf{X}_S = \mathbf{x}_S\big]$$

Feature attributions

Shapley values are a popular feature attribution method [7, 17, 1].

$$f(\mathbf{x}) = \sum_{j=0}^{d} \phi_v(j, \mathbf{x})$$

They are defined with respect to a characteristic function.

$$v(S, \mathbf{x}) = \mathbb{E}\big[f(\mathbf{x}) \mid \mathbf{X}_S = \mathbf{x}_S\big]$$

Each $\phi_v(j, \mathbf{x})$ represents a weighted average of j's marginal contribution to subsets that exclude it.

$$\phi_v(j, \mathbf{x}) = \sum_{S \subseteq [d] \backslash \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \big[v(S \cup \{j\}, \mathbf{x}) - v(S, \mathbf{x})\big]$$

Feature attributions

Shapley values are a popular feature attribution method [7, 17, 1].

$$f(\mathbf{x}) = \sum_{j=0}^{d} \phi_v(j, \mathbf{x})$$

They are defined with respect to a characteristic function.

$$v(S, \mathbf{x}) = \mathbb{E}\big[f(\mathbf{x}) \mid \mathbf{X}_S = \mathbf{x}_S\big]$$

Each $\phi_v(j, \mathbf{x})$ represents a weighted average of j's marginal contribution to subsets that exclude it.

$$\phi_v(j, \mathbf{x}) = \sum_{S \subseteq [d] \setminus \{j\}} \frac{|S|!\,(d - |S| - 1)!}{d!} \big[v(S \cup \{j\}, \mathbf{x}) - v(S, \mathbf{x})\big]$$

Proposition 1. Let $c_S(\boldsymbol{x}) = c$ iff $\boldsymbol{x} \sim \delta(\boldsymbol{x}_S)\, p(\boldsymbol{x}_{\overline{S}} \mid \boldsymbol{x}_S)$. Then $v(S, \mathbf{x}) = PS(c_S, y)$.

Rule lists

Anchors [13] learn a set of Boolean conditions A such that $A(\boldsymbol{x}) = 1$ and

$$prec(A) := P_{\mathcal{D}_{(\boldsymbol{x}_i|A)}}(f(\boldsymbol{x}) = f(\boldsymbol{x}_i)) \geq \tau.$$

For fixed τ, the goal is to maximize *coverage*: $\mathbb{E}[A(\boldsymbol{x}_i) = 1]$, i.e. the proportion of datapoints to which the anchor applies.

Unification

uai2021

Rule lists

Anchors [13] learn a set of Boolean conditions A such that $A(\boldsymbol{x}) = 1$ and

$$prec(A) := P_{\mathcal{D}_{(\boldsymbol{x}_i | A)}}(f(\boldsymbol{x}) = f(\boldsymbol{x}_i)) \geq \tau.$$

For fixed τ, the goal is to maximize *coverage*: $\mathbb{E}[A(\boldsymbol{x}_i) = 1]$, i.e. the proportion of datapoints to which the anchor applies.

Proposition 2. Let $c_A(\boldsymbol{z}) = c$ iff $A(\boldsymbol{x}) = 1$. Then $prec(A) = PS(c_A, y)$.

Unification

Counterfactuals

The counterfactual recourse objective [4] is simply to find the highest ranked factor in the partial ordering that exceeds a sufficiency threshold.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}_i \in \mathrm{CF}(\boldsymbol{x})}{\operatorname{argmin}} \; cost(\boldsymbol{x}, \boldsymbol{x}_i).$$

Unification

Counterfactuals

The counterfactual recourse objective [4] is simply to find the highest ranked factor in the partial ordering that exceeds a sufficiency threshold.

$$\boldsymbol{x}^* = \underset{\boldsymbol{x}_i \in \mathrm{CF}(\boldsymbol{x})}{\mathrm{argmin}} \; cost(\boldsymbol{x}, \boldsymbol{x}_i).$$

Proposition 3. Let *cost* be a function representing $\preceq$, and let c be some factor spanning reference values. Then the counterfactual recourse objective is:

$$c^* = \underset{c \in \mathcal{C}}{\mathrm{argmin}} \; cost(c) \; \text{ s.t. } PS(c, y') \geq \tau.$$

Probabilities of Causation

Pearl [18, 10] defines probabilities of causation over counterfactual domains to quantify the extent to which an effect is sensitive to its cause—turning on in its presence and off in its absence.

Probabilities of Causation

Pearl [18, 10] defines probabilities of causation over counterfactual domains to quantify the extent to which an effect is sensitive to its cause—turning on in its presence and off in its absence.

Proposition 4. Let $X, Y \in \{0,1\}^2$. We have two counterfactual distributions: $\mathcal{I} := P(y_x \mid x', y')$ and $\mathcal{R} := P(y'_{x'} \mid x, y)$ and a uniform mixture over the two, $P(y) = 0.5\mathcal{I} + 0.5\mathcal{R}$. Let auxiliary variable W tag each sample with a label indicating whether it comes from the input or reference distribution. Define $c(\boldsymbol{z}) = w$. Then we have $CS(x, y) = PS(c, y)$ and $CN(x, y) = PS(c', y')$.

LENS

Local Explanations via Necessity and Sufficiency (LENS) computes minimally sufficient factors with respect to a given basis $\mathcal{B}$ and threshold τ.

Local Explanations via Necessity and Sufficiency (LENS) computes minimally sufficient factors with respect to a given basis $\mathcal{B}$ and threshold τ.

Theorem 1

With oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$, LENS is sound and complete. That is, for any C returned by LENS and all $c \in \mathcal{C}$, c is τ-minimal iff $c \in C$.

Local Explanations via Necessity and Sufficiency (LENS) computes minimally sufficient factors with respect to a given basis $\mathcal{B}$ and threshold τ.

Theorem 1

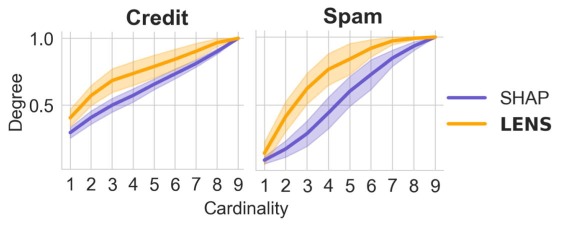
With oracle estimates $PS(c, y)$ for all $c \in \mathcal{C}$, LENS is sound and complete. That is, for any C returned by LENS and all $c \in \mathcal{C}$, c is τ-minimal iff $c \in C$.

Theorem 2

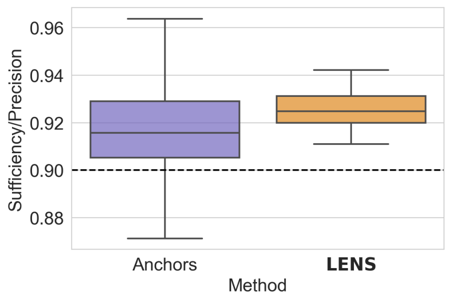
With sample estimates $\hat{PS}(c, y)$ for all $c \in \mathcal{C}$, LENS is uniformly most powerful. That is, LENS identifies the most τ-minimal factors of any method with fixed type I error α.

Experiments

uai2021

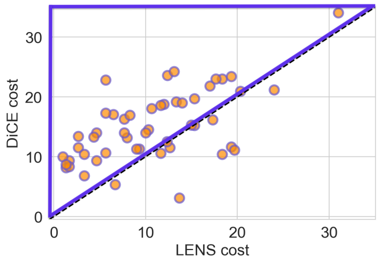


LENS provides more informative explanations than SHAP [7] for any fixed degree of sparsity.



Anchors [13] satisfy a PAC bound, which means some explanations may be less than τ-sufficient. Factors output by LENS, however, are guaranteed to meet the τ-minimality criterion.

Experiments



LENS produces lower-cost counterfactuals than DiCE [9] on average.

Conclusion

Theoretical contribution

Our formal framework clarifies the relationship between various XAI methods, as well as their connections to fundamental quantities from logic, probability, and causality.

Conclusion

Theoretical contribution

Our formal framework clarifies the relationship between various XAI methods, as well as their connections to fundamental quantities from logic, probability, and causality.

Algorithmic contribution

LENS is an optimal procedure for computing minimally sufficient factors.

Conclusion

Theoretical contribution

Our formal framework clarifies the relationship between various XAI methods, as well as their connections to fundamental quantities from logic, probability, and causality.

Algorithmic contribution

LENS is an optimal procedure for computing minimally sufficient factors.

Limitations

LENS prioritizes completeness over efficiency. Future work will explore more scalable approximations, as well as model-specific variants.

References I

uai2021

[1] H. Chen, J. D. Janizek, S. Lundberg, and S.-I. Lee. True to the Model or True to the Data? *arXiv* preprint, 2006.16234, 2020.

[2] C. Fernández-Loría, F. Provost, and X. Han. Explaining data-driven decisions made by AI systems: The counterfactual approach. *arXiv* preprint, 2001.07417, 2020.

[3] A.-H. Karimi, G. Barthe, B. Balle, and I. Valera. Model-agnostic counterfactual explanations for consequential decisions. In *AISTATS*, pages 895–905, 2020.

[4] A.-H. Karimi, G. Barthe, B. Schölkopf, and I. Valera. A survey of algorithmic recourse: Definitions, formulations, solutions, and prospects. *arXiv* preprint, 2010.04050, 2020.

[5] I. Kumar, S. Venkatasubramanian, C. Scheidegger, and S. Friedler. Problems with Shapley-value-based explanations as feature importance measures. In *ICML*, pages 5491–5500, 2020.

[6] H. Lakkaraju, E. Kamar, R. Caruana, and J. Leskovec. Faithful and customizable explanations of black box models. In *AIES*, pages 131–138, 2019.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *NeurIPS*, pages 4765–4774. 2017.

[8] R. K. Mothilal, D. Mahajan, C. Tan, and A. Sharma. Towards unifying feature attribution and counterfactual explanations: Different means to the same end. *arXiv* preprint, 2011.04917, 2020.

[9] R. K. Mothilal, A. Sharma, and C. Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *FAT**, pages 607–617, 2020.

[10] J. Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York, 2000.

[11] Y. Ramon, D. Martens, F. Provost, and T. Evgeniou. A comparison of instance-level counterfactual explanation algorithms for behavioral and textual data: SEDC, LIME-C and SHAP-C. *Adv. Data Anal. Classif.*, 2020.

[12] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of Aay classifier. In *SIGKDD*, pages 1135–1144, New York, NY, USA, 2016.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin. Anchors: High-precision model-agnostic explanations. In *AAAI*, pages 1527–1535, 2018.

[14] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.*, 1(5):206–215, 2019.

[15] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. Fooling LIME and SHAP: Adversarial attacks on post-hoc explanation methods. In *AIES*, pages 180–186, 2020.

[16] K. Sokol and P. Flach. LIMEtree: Interactively customisable explanations based on local surrogate multi-output regression trees. *arXiv* preprint, 2005.01427, 2020.

[17] M. Sundararajan and A. Najmi. The many Shapley values for model explanation. In *ACM*, New York, 2019.

[18] J. Tian and J. Pearl. Probabilities of causation: Bounds and identification. *Ann. Math. Artif. Intell.*, 28(1-4):287–313, 2000.

[19] S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard J. Law Technol.*, 31(2):841–887, 2018.

[20] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Trans. Vis. Comput. Graph.*, 26(1):56–65, 2020.

Thanks!

uai2021

Comments? Questions? Get in touch!

david.watson@kcl.ac.uk

https://dswatson.github.io