Framework for Evaluating Faithfulness of Local Explanations

Sanjoy Dasgupta [1] **Nave Frost** [2] **Michal Moshkovitz** [2]

Abstract

We study the faithfulness of an explanation system to the underlying prediction model. We show that this can be captured by two properties, *consistency* and *sufficiency*, and introduce quantitative measures of the *extent* to which these hold. Interestingly, these measures depend on the test-time data distribution. For a variety of existing explanation systems, such as anchors, we analytically study these quantities. We also provide estimators and sample complexity bounds for empirically determining the faithfulness of black-box explanation systems. Finally, we experimentally validate the new properties and estimators.

1. Introduction

Machine learning is an integral part of many human-facing computer systems and is increasingly a key component of decisions that have profound effects on people's lives. There are many dangers that come with this. For instance, statistical models can easily be error-prone in regions of the input space that are not well-reflected in training data but that end up arising in practice. Or they can be excessively complicated in ways that impact their generalization ability. Or they might implicitly make their decisions based on criteria that would not considered acceptable by society. For all these reasons, and many others, it is crucial to have models that are understandable or can *explain* their predictions to humans (Kim & Doshi-Velez, 2021).

Explanations of a classification system can take many forms, but should accurately reflect the classifier's inner workings. Perhaps the best scenario is where the model itself is inherently understandable by humans. This is arguably true of decision trees, for instance. If the tree is small, then it can be fathomed in its entirety: a *global explanation* of every prediction the model makes. If the tree is large, it can be hard to understand as a whole, but as long as it has modest depth, any individual prediction can be *locally explained* using the features on the corresponding root-to-leaf path.

A common situation is where the predictive model is not inherently understandable, either at a global or local level, and so a separate *post-hoc explanation* is needed. These are typically local, in the sense that they explain a specific prediction and perhaps also explain what the model does on other nearby instances. Over the past few years, many strategies for post-hoc explanation have emerged, such as LIME (Ribeiro et al., 2016), Anchors (Ribeiro et al., 2018), and SHAP (Lundberg & Lee, 2017).

Explanation systems need to satisfy two broad criteria: the explanations should (i) make sense to a human user and (ii) be an accurate reflection of the actual predictive model. The first of these is hard to pin down because it is inextricably linked to vagaries of human cognition: is a linear model "understandable", for instance? Further research is needed to better characterize what (i) might mean. This paper focuses on criterion (ii): gauging the *faithfulness* of explanations to the underlying predictive model, or put differently, the *internal coherence* of the overall explanation system.

1.1. Contributions

We focus on classification problems and on explanation systems that consist of two components:

- A prediction function (the classifier) $f : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is the instance space and $\mathcal{Y}$ is the label space.
- An explanation function $e : \mathcal{X} \to \mathcal{E}$, where $\mathcal{E}$ is the space of explanations, or properties.

The explanation function explains the prediction $f(x)$ by pointing out some relevant property of the input. These properties can be quite general. Consider, for instance, a decision tree. Its prediction $f(x)$ on a point x can be explained by the features on the root-to-leaf path for x; the explanation $e(x)$ is the conjunction of these features, e.g. "$(x_2 > 0.5) \wedge (x_4 = \text{true}) \wedge (x_{10} < -1)$". Thus the set $\mathcal{E}$ has a conjunction for each leaf of the tree.

Or consider a classifier that takes an image x of a landscape

[1]University of California San Diego. [2]Tel-Aviv University. Correspondence to: Sanjoy Dasgupta <dasgupta@eng.ucsd.edu>, Nave Frost <navefrost@mail.tau.ac.il>, Michal Moshkovitz <moshkovitz5@mail.tau.ac.il>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

and returns its biome, e.g., `rainforest`. One way this predictor $f(x)$ might operate is by identifying telltale flora or fauna in the image. For instance, if the image contains a zebra then its biome must be savannah: $f(x) = \texttt{savannah}$ and $e(x) =$ "contains a zebra". Although such explanations are based on nontrivial attributes of the input, they are comprehensible to humans and are within the scope of our setup.

For an explanation system to be internally coherent, it should satisfy two properties:

- **Consistency:** Roughly, two instances x, x' that get the same explanation should also have the same prediction.

  For instance, if two different images are assigned the same explanation, $e(x) = e(x') =$ "contains a zebra", then their assigned labels should also be the same.

- **Sufficiency:** If x is assigned an explanation $e(x) = \pi$ that also holds for another instance x' (even if $e(x') \neq \pi$), then x' should have the same label as x.

  For instance, if an image x is assigned explanation $e(x) =$ "contains a zebra" and label $f(x) = \texttt{savannah}$, then a different image x' that also happens to contain a zebra should get the same label, even if it is assigned a different explanation, e.g. $e(x') =$ "contains a baobab tree".

These properties are desirable but might not hold in all cases. We introduce quantitative measures of the *extent* to which they hold.

With these measures in hand, we study a variety of established explanation systems: decision trees, Anchors, highlighted text, LIME, SHAP, gradient-based method, k-nearest neighbors, and counterfactuals. We show how they map into our framework and study their faithfulness. For instance, we prove that SHAP has perfect consistency, while LIME does not. We also have results at a higher level of abstraction. We formalize a natural sub-category of explanation systems that we call *explicitly scoped rules*, that includes decision trees, anchors, and highlighted text. These have a common structure that permits their faithfulness to be studied in generality.

Another important use of these quantitative measures is to *empirically* characterize the faithfulness of black-box explanation systems whose internals might not be known. We give statistical estimators for doing so and characterize their sample complexity. Along the way, we formalize what property a black-box explanation system should possess in order for its faithfulness to be easily verifiable. Roughly, this corresponds to a particular type of *compression* achieved by the explanations. Indeed, we show (Claim 2) that absent any such compression, verification is not possible.

An interesting aspect of our measures is that the extent of faithfulness of an explanation system depends on the data distribution to which it will be applied, and thus might not be known at training time[1]. Thus faithfulness may need to be assessed anew for each new setting in which the system will be used. In general, there is a tradeoff between simplicity of explanations and fidelity to the predictor. When explaining an animal recognizer, for instance, it might be reasonable to ignore special cases like marsupials if the system is used in North America, but not if it is used in Australia.

Summary of contributions:

- Framework for evaluating the faithfulness of black-box explanation systems
- Analysis of popular explanation methods
- Estimators for faithfulness, with rates of convergence
- Ease of estimation depends upon a notion of compression achieved by the explanations
- Empirical evaluation of these measures and estimators
- Highlighting fundamental properties of the faithfulness measure such as data dependence

1.2. Related Work

There are many types of explanation (Lipton, 2018; Molnar, 2019). At a high level, we can separate them into two groups. In *intrinsic explanations*, the prediction models themselves are simple and self-explanatory, such as decision trees (Quinlan, 1986), decision lists (Rivest, 1987), and risk scores (Ustun & Rudin, 2019). *Post-hoc explanations* are applied to existing predictors and come in many varieties, as described throughout the paper.

The importance of evaluating explanation methods has been discussed in the literature (Leavitt & Morcos, 2020; Zhou et al., 2021; Kim & Doshi-Velez, 2021; Pruthi et al., 2022). There are various attempts to measure different aspects of an explanation: usefulness to humans (Jesus et al., 2021; Mohseni et al., 2018; Poursabzi-Sangdeh et al., 2021); complexity (Poppi et al., 2021); difficulty of answering queries (Barceló et al., 2020); and robustness (Alvarez-Melis & Jaakkola, 2018; Agarwal et al., 2022). In this paper, we measure faithfulness to the model. Earlier work has looked at global measures of this type (Wolf et al., 2019) and measures that are specialized to neural networks (Poppi et al., 2021; Tomsett et al., 2020; Yeh et al., 2019; Ancona et al., 2017), feature importance (Amparore et al., 2021; Carmichael &

[1] This was observed for surrogate explanations, e.g., (Lakkaraju et al., 2020), however, we observe that faithfulness being data-depended is a general phenomenon applicable to any local explanation system.

Scheirer, 2021; Sundararajan et al., 2017; Velmurugan et al., 2021; Bhatt et al., 2020), rule-based explanations (Margot & Luta, 2020), surrogate explanation (Ribeiro et al., 2016), or highlighted text (Chen et al., 2018; Wang et al., 2020a; Yoon et al., 2018). In contrast to these works which are dedicated to a single type of explanation system, this paper suggests a *general* framework for faithfulness evaluation, applicable to any black-box explanation system.

2. Framework

As described in the introduction, we think of an explanation system as consisting of a *prediction function* (classifier) $f : \mathcal{X} \to \mathcal{Y}$ and an *explanation function* $e : \mathcal{X} \to \mathcal{E}$. (If $e(\cdot)$ is randomized, we can focus on one random seed.) The local explanation for model f at instance x is some relevant property of x, denoted $e(x)$. The selected property should ideally be enough, on its own, to predict label $f(x)$. This general intuition has appeared in many places in the literature. Here we break it into two components—*consistency* and *sufficiency*—and provide precise measures of each.[2].

2.1. Consistency

For any explanation $\pi \in \mathcal{E}$, consider the set of instances that are assigned this explanation:

$$C_\pi = \{x \in \mathcal{X} : e(x) = \pi\}.$$

If π is a good explanation, then we would hope that these instances all have the same predicted label. This is *consistency*: instances that are assigned the same explanation should also be assigned the same prediction.

For some explanation systems, this may not hold all the time. We would like to quantify the extent to which it holds. We start by introducing a measure of the homogeneity of predictions in C_π. In order to do this, we need a distribution μ over instances $\mathcal{X}$. This can be thought of as the distribution of instances that arise in practice.

Definition 1 (local consistency)**.** *The consistency of explainer e for model f at instance x, with respect to distribution μ, is defined as*

$$m^c(x) = \Pr_{x' \in_\mu C_\pi}(f(x') = f(x))$$

where $\pi = e(x)$ and the notation $x' \in_\mu C_\pi$ means "x' is drawn from distribution μ restricted to the set C_π."

Global consistency. We have so far quantified consistency at a specific instance. It is also of interest to measure the consistency of the *entire* model.

[2]The terms consistency and sufficiency have been previously used in the context of explainability (Hase et al., 2021; Fel et al., 2022) but with different meaning than this paper.

Definition 2 (global consistency)**.** *The global consistency of explanation system (f, e), with respect to distribution μ over $\mathcal{X}$, is*

$$m^c = \mathop{\mathbb{E}}_{x \in_\mu \mathcal{X}}[m^c(x)].$$

Relation to decoding. The definition of global consistency implicitly defines a decoder d from explanations π to labels y. Recall that $C_\pi \subseteq \mathcal{X}$ is the set of all instances that get assigned explanation π. These instances might not all have the same predicted label, but we can look at the distribution over labels,

$$\Pr(y|\pi) = \Pr_{x \in_\mu C_\pi}(f(x) = y).$$

With this in place, there are two natural ways to define the decoder: (i) the (randomized) *Gibbs decoder* that, given explanation π, returns a label y with probability $\Pr(y|\pi)$, and (ii) the optimal *deterministic decoder* that returns the label y that maximizes $\Pr(y|\pi)$. We can denote the resulting decoding error, $\Pr_x(f(x) \neq d(e(x)))$, by E_G for the Gibbs decoder and E_O for the deterministic decoder. Standard manipulations show that these two errors are very similar:

Claim 1. $E_O \leq E_G \leq 2E_O$.

Our notion of consistency is exactly the accuracy of the Gibbs decoder: $m^c = 1 - E_G$.

2.2. Sufficiency

A complementary requirement from an explainer is that if a property π is used to justify the prediction at instance x, then any other instance x' with property π should also be classified the same way. Moreover, this should hold even if the supplied explanation, $e(x')$, is different from π. In the earlier biome example, if the explanation "contains a zebra" is ever used to justify a prediction of `savannah`, then any picture with a zebra in it should get the same prediction, even if assigned some other property as explanation.

To start with, we say that explanations $\mathcal{E}$ are *intelligible* if for any instance $x \in \mathcal{X}$ and property $\pi \in \mathcal{E}$, it is possible to assess whether π applies to x. If so, we define this as a relation $A(x, \pi)$. Ideally, the relation would not only be well-defined but would also be checkable by humans. Note that the relation depends solely on the instance and not on the true or predicted label. We define the set of instances C_x that share the same property as x's explanation by

$$C_x = \{x' \in \mathcal{X} : A(x', e(x))\}.$$

As with the consistency measure, each instance can have a different level of sufficiency; it is not a binary value. To define this measure, we use a probability distribution over C_x. The sufficiency measure tests the homogeneity of predictions made on C_x.

Definition 3 (local sufficiency)**.** *The local sufficiency of explainer e for model f at instance x, with respect to distribution μ, is defined as*

$$m^s(x) = \Pr_{x' \in_\mu C_x}(f(x') = f(x)).$$

Recall that the notation $x' \in_\mu C_x$ means "x' is drawn from distribution μ restricted to the set C_x".

Consistency and sufficiency are complementary measures. For any given instance x, $m^c(x)$ can be larger, smaller, or equal to $m^s(x)$. Similarly to global consistency, we define a sufficiency measure for the entire model.

Definition 4 (Global sufficiency)**.** *The global sufficiency of explanation system (f, e), with respect to distribution μ on $\mathcal{X}$, is equal to $m^s = \mathbb{E}_{x \in_\mu \mathcal{X}}[m^s(x)]$.*

3. Analysis of common explanation systems

In this section we review some popular explanation methods and assess the extent to which they achieve consistency and sufficiency. We divide these methods into three sub-categories: explicitly scoped rules, feature importance scores, and example-based explanations.

3.1. Explicitly scoped rules

In "scoped rules", each explanation is an explicit region of the instance space, e.g., "$(x_2 > 0.5) \wedge (x_4 = \text{true}) \wedge (x_{10} < -1)$". This type of explanation includes decision trees, anchors, and highlighted text, which we elaborate on next.

3.1.1. Decision trees

Suppose the instance space is some $\mathcal{X} \subseteq \mathbb{R}^d$. When a decision tree is used to "explain" a classifier $f : \mathcal{X} \to \mathcal{Y}$, the tree is fit to f's predictions (Dasgupta et al., 2020; Hu et al., 2019; Moshkovitz et al., 2021). The explanation of an instance x is the conjunction of the features along the path from T's root to the leaf in which x lies. Thus the explanations, $\mathcal{E}$, are in one-to-one correspondence with the leaves of the tree.

In this case, an explanation π applies to an instance x if and only if x falls in π's leaf. Therefore, the relation $A(x, \pi)$ is intelligible (well-defined) and easy for a human to assess. Moreover, consistency is equal to sufficiency, and they measure the accuracy of the tree in capturing f:

$$\begin{aligned} m^c = m^s &= \Pr_{X, X' \sim \mu}(f(X) = f(X') | e(X) = e(X')) \\ &= \sum_{\text{leaves } \pi} \mu(C_\pi) \Pr(f(X) = f(X') | X, X' \in C_\pi) \\ &= 1 - \sum_\pi \mu(C_\pi)(\text{Gini-index of } f \text{ in } C_\pi) \end{aligned}$$

where C_π is the subset of $\mathcal{X}$ that ends up in leaf π.

3.1.2. Anchors

Pick any data space $\mathcal{X} \subseteq \mathbb{R}^d$ and prediction function $f : \mathcal{X} \to \mathcal{Y}$. An *anchor* explanation (Ribeiro et al., 2018) for an instance $x \in \mathcal{X}$ is an explicitly-specified hyperrectangle $H_x \subset \mathbb{R}^d$ that contains x and that is meant to correspond, roughly, to a region around x that is similarly labeled.

The quality of an anchor is typically formalized using the notion of *precision*, which is the probability, over the distribution μ, that a random instance in H_x has label $f(x)$, that is, $\Pr_{x' \in_\mu H_x}[f(x') = f(x)]$.

In this case, the space of explanations is the set of all anchor-hyperrectangles, $\mathcal{E} = \{H_x : x \in \mathcal{X}\}$. It is easy to check whether an anchor applies to an instance: the relation $A(x, H) \equiv (x \in H)$ is well-defined. Moreover, our notion of local sufficiency is exactly the precision of anchors and global sufficiency is exactly the average precision of anchors:

$$\begin{aligned} m^s &= \Pr_{X, X' \sim \mu}(f(X') = f(X) | A(X', e(X))) \\ &= \mathop{\mathbb{E}}_{X \sim \mu}[\text{precision}(H_X)]. \end{aligned}$$

If anchors are chosen to be discrete—that is, the same hyperrectangles are used many times—then our notion of consistency gauges the uniformity of prediction over all instances for which a particular anchor is specified:

$$m^c = \Pr_{X, X' \sim \mu}(f(X') = f(X) | H_{X'} = H_X).$$

There is no immediate relation between this and sufficiency or precision.

3.1.3. Highlighted text

The goal in *highlighted text* explanations is to pick out the features—for instance, words in text—that are most important for a model's prediction (Jacovi & Goldberg, 2021). For an instance $x \in \mathbb{R}^d$, the explanation can be thought of as a subset of features $S \subseteq [d]$, and the values (e.g., text) of these features, $x_S \in \mathbb{R}^{|S|}$.

These explanations are anchors at the level of generality of Section 3.1.2. Thus the same observations apply here.

3.1.4. A unified framework for explicitly scoped rules

The last three examples—decision trees, anchors, and highlighted text—have a common structure that is appealingly simple and may also hold for many future explanation systems. To formalize it, we say an explanation system (f, e) has *explicitly scoped rules* if each explanation π is a description of a region $S_\pi \subseteq \mathcal{X}$ of the instance space. For a given point x, the explanation $\pi = e(x)$ has the property

that $x \in S_\pi$. The terminology "explicitly scoped" means that subset S_π is specified in a form where it is easy to check whether a specific point lies in it or not. Thus the set of explanations is $\mathcal{E} = \{e(x) : x \in \mathcal{X}\}$ and the relation $A(x, \pi) \equiv (x \in S_\pi)$ is well-defined (intelligible). This is the key property of explicitly-scoped rules.

We can generalize the notion of precision to any region of space (not just hyperrectangles) and as in the case of anchors, sufficiency will then correspond to average precision.

For the explanation systems we will cover next, intelligibility—determining whether an explanation applies to a given instance—is more tricky.

3.2. Feature importance methods

Feature importance methods aim to give a precise indication of which features of an input x are most relevant to the prediction $f(x)$. This often takes the form of a local linear model g_x (sometimes on a simplified instance space) that approximates f in the vicinity of x. However, the scope of this g_x—the region over which the approximation is accurate—is sometimes unspecified, in which case it is unclear when a particular g_x can be thought of as being applicable to some other point x'. Because of the ambiguity in the intelligibility of these explanations, we will focus on consistency in what follows.

3.2.1. LIME

LIME (Ribeiro et al., 2016) provides an explanation of $f(x)$ by (1) using an *interpretable representation* $\psi : \mathcal{X} \to \mathcal{X}'$, e.g. the presence or absence of individual words in a document, and (2) approximating f near x with a simple model $g_x : \mathcal{X}' \to \mathcal{Y}$. Typically, g_x is a linear classifier.

LIME does not exhibit perfect consistency, e.g., points x with the same interpretable representation get assigned the same g_x, while their predicted labels may vary. Another example is depicted in Appendix B.

3.2.2. SHAP

SHAP (Lundberg & Lee, 2017) is similar in spirit to LIME. It uses a Boolean feature space $\mathcal{X}'$ and its explanations are linear functions $g_x : \mathcal{X}' \to \mathcal{Y}$. But this time the choice of g_x is inspired by Shapley values (Shapley, 1953) from game theory, and is chosen to satisfy four axioms for fair distribution of gains: efficiency, symmetry, linearity, and null player. In particular, the coefficients of g_x are guaranteed to sum to $f(x) - \phi_0$, where ϕ_0 is constant for all x.

This last property guarantees that if two examples have the same explanation, then their label must be the same, thus ensuring perfect consistency.

3.2.3. GRADIENT-BASED METHOD

Gradient-based explanations are popular for neural nets (Agarwal et al., 2021; Ancona et al., 2017; Shrikumar et al., 2016; Simonyan et al., 2013; Smilkov et al., 2017). The explanation is the gradient of the network with respect to the instance, the intuition being that features with highest gradient values have the most influence on the model's output.

The gradient alone determines a function only up to an additive constant. This offset must also be provided to complete the explanation; otherwise there is imperfect consistency. The lack of decodability was empirically observed in several previous works (Adebayo et al., 2018; Anders et al., 2020; Kim & Doshi-Velez, 2021; Nie et al., 2018; Wang et al., 2020b).

3.3. Example-based explanations

An example-based explainer justifies the prediction on an instance x by returning instances related to x. Explanations of this type include *nearest neighbors* and *counterfactuals*.

3.3.1. NEAREST NEIGHBORS

Let's focus on 1-nearest neighbor for concreteness. For a given prediction function $f : \mathcal{X} \to \mathcal{Y}$, a nearest neighbor explanation system maintains a set of prototypical instances $\mathcal{P} \subseteq \mathcal{X}$ and justifies the prediction on instance x by returning a prototype $p \in \mathcal{P}$ close to x (with respect to an underlying distance function d on $\mathcal{X}$). Thus the space of explanations is $\mathcal{E} = \mathcal{P}$.

Our consistency measure then checks the extent to which points x, x' that get mapped to the same prototype $p \in \mathcal{P}$ also get the same prediction under f.

For sufficiency, we also need to define the relation $A(x, p)$: when do we consider prototype p to be "applicable to" instance x? Here are two options.

1. When p is the nearest neighbor of x in $\mathcal{P}$.
2. When $d(x, p) \leq \tau$ for some threshold $\tau > 0$.

The first option strictly follows the nearest neighbor rule, but leads to problems with verifiability; for instance, it is not easy for a human to check that $A(x, p)$ holds unless the set $\mathcal{P}$ is somehow available. The second option is easier to check; in fact, we can treat the regions $B(p, \tau)$ as anchors and then measure consistency and sufficiency using the methods of Section 3.1.4.

3.3.2. COUNTERFACTUALS

A counterfactual explanation of an instance x is another instance x' which is close to x but has a different label, $f(x) \neq$

$f(x')$ (Deutch & Frost, 2019; Mothilal et al., 2020; Slack et al., 2021). To make this concrete, suppose we are performing binary classification and that some distance function d has been chosen for the instance space $\mathcal{X}$. Then the counterfactual explanation for x is the closest point x' that gets the opposite label, that is, $x' = \arg\min_{x':f(x')\neq f(x)} d(x, x')$. The space of explanations is $\mathcal{E} = \mathcal{X}$.

In this case, the "explanation" x' gives information about the nature of predictions in the vicinity of x. Specifically, it asserts that *any point in the open ball* $B(x, d(x, x'))$ *has label* $f(x)$. Therefore, one way to verify faithfulness of these explanations is simply to associate them with scoped rules of this form and to then assess sufficiency as in Section 3.1.4.

4. Evaluating faithfulness of black-box model

In this section, we consider a scenario where we are given a black-box explanation system (f, e) and wish to evaluate its faithfulness. To this end, we develop statistical estimators for consistency and sufficiency given samples $x_1, \ldots, x_n$ from an underlying test distribution μ. How many such samples are needed to accurately assess faithfulness?

4.1. Discrete explanation spaces

Let's begin with the case where the explanation space $\mathcal{E}$ is discrete (that is, countable). We will not assume that we know the entire set $\mathcal{E}$, since this knowledge will not be in general available for a black-box explanation system. Given a few samples from μ, we can look at the resulting explanations and predictions, but it is not trivial to assess the fraction of the explanation space that we have not seen: that is, the *missing mass*. And for any explanation π that we do not see, faithfulness could be arbitrarily bad. With this difficulty in mind, we now turn to our estimators.

A key observation is that although consistency and sufficiency measure different aspects of the explanation system, for the purposes of statistical estimation they can be treated together. To see this, let $R(x, \pi)$ denote an arbitrary relation on $\mathcal{X} \times \mathcal{E}$, and for a given distribution μ on $\mathcal{X}$, define

$$m_\mu^R = \Pr_{X,X'\sim\mu}\left(f(X') = f(X) | R(X', e(X))\right).$$

This generalizes both types of faithfulness: for consistency, take $R(x, \pi)$ to mean $e(x) = \pi$ and for sufficiency take $R(x, \pi) \equiv A(x, \pi)$.

Thus we only need an estimator for m_μ^R. The quality of our estimate will depend upon what fraction of the explanation space we get to see, which in turn depends on $\mathcal{E}$ and μ.

We begin with a few related definitions. Let $p(\pi)$ be the fraction of points for which explanation π is provided, that is, $p(\pi) = \mu(\{x : e(x) = \pi\})$ and let $q(\pi)$ be the fraction for which $R(x, \pi)$ holds: $q(\pi) = \mu(\{x : R(x, \pi)\})$. Thus $p(\pi)$ is a distribution over $\mathcal{E}$ while $q(\pi) \in [0, 1]$ and $q(\pi) \geq p(\pi)$.

Given samples $x_1, \ldots, x_n \sim \mu$, and any y, π, define

$$N_\pi = |\{i : R(x_i, \pi)\}|$$
$$N_{\pi,y} = |\{i : R(x_i, \pi), f(x_i) = y\}|$$

Our estimator for m_μ^R is then

$$\widehat{M} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}(N_{e(x_i)} > 1)\,\frac{N_{e(x_i),f(x_i)} - 1}{N_{e(x_i)} - 1}.$$

We can show the following rate of convergence.

Theorem 1. *The mean-squared error of estimator* $\widehat{M}$ *can be bounded as follows:*

$$\mathbb{E}\left[(\widehat{M} - m_\mu^R)^2\right] \;\leq\; \frac{4}{n} + \left(\sum_\pi p(\pi) e^{-(n-1)q(\pi)}\right)^2.$$

The mean-squared error is the sum of the variance, which is bounded by $4/n$, and the squared bias, the term in parentheses. This bias arises from the inability to correctly assess faithfulness for explanations π that appear 0 or 1 times in the data. One way to make this term small, say $< \epsilon$, is to have n comparable to the size of $\texttt{range}(e)$. In this way, we see that the ease of evaluating the faithfulness of explanations depends on the level of *compression* they achieve.

Corollary 1. *Suppose the unlabeled sample size is at least* $n \geq \texttt{range}(e) \cdot \frac{12}{\epsilon}\log\frac{3}{\epsilon}$. *Then, the mean-squared error of* $\widehat{M}$ *(for either consistency or sufficiency) is at most* ϵ.

4.2. Larger or continuous explanation spaces

The estimator of the previous section needs explanations to appear at least twice before it can begin assessing their faithfulness. This is problematic in continuous explanation spaces, where no explanation might ever be repeated.

One fix, which we later study empirically, is to discretize the space $\mathcal{E}$. We introduce a function $\psi : \mathcal{E} \to \mathcal{E}'$ where $\mathcal{E}'$ is much smaller than $\mathcal{E}$, and consider explanations π, π' to be equivalent if $\psi(\pi) = \psi(\pi')$. An alternative fix is to introduce a distance function d between explanations, and to use a $d(\pi, \pi') \leq \tau$ to determine when π and π' are close enough that they should yield the same prediction.

Explainers that their inner-working is known, their consistency and sufficiency might also be known (e.g., SHAP has perfect consistency). However, the new measures need to be estimated if the inner working is unknown. This section provided conditions where such an estimation is possible. Unfortunately, there are some cases where it is impossible to apply *any* estimation method. One such scenario is where all the explanations are distinct, as the next claim shows.

Claim 2. *(unverifiable explainer) Fix infinite example set $\mathcal{X}$. There are two explainers, e_1 and e_2, a model f, and a distribution over the examples, where on every finite-sample, with probability 1, the explanations are the same, but the sufficiency and consistency of e_1 is 1 while the sufficiency and consistency of e_2 is 0.5.*

5. Experiments

5.1. Canonical properties

We begin with experiments that illustrate basic properties of our faithfulness estimators: (1) they assign low scores to random explanations, (2) they assign higher scores to more faithful explanations as long as the explanation space is not too large, and (3) when the explanation space is huge relative to the amount of unlabeled data, they conservatively assign low scores since they are unable to assess faithfulness.

Highlighted text. To evaluate a variety of *highlighted text* explainers, we began by training a predictor on the `rt-polaritydata` dataset, used for sentiment classification of movie reviews, with 10,433 documents. We represented each document as a bag of words, and used 80% of the data to train a linear model. The remaining documents were used to compare four highlighted text explainers.

We evaluated four explainers. (1) *Top Coefficient* is a white-box explainer that highlights the word in the sentence with the highest absolute coefficient in the linear model. (2) *Anchors* (Ribeiro et al., 2018). (3) *First Word* always highlights the first word in the sentence as the explanation. (4) *All Words* highlights all the words in the sentence as the explanation. We estimated global consistency and sufficiency for each explainer as described in previous sections. We also recorded the *uniqueness* of each explainer, which is the fraction of test data whose explanations were unique.

The results are presented in Table 1. Top Coefficient got the highest consistency and sufficiency scores, as one might expect from an explainer that utilizes its complete knowledge of the model. As Anchors is a black-box explainer that attempts to return a faithful explanation, it produces better results than the last two explainers, which are not designed to be faithful to the model. First Word is close to a random explainer, and thus gets rather low sufficiency and consistency. All Words highlights the entire input and thus has maximal uniqueness (1.0), making it unverifiable. Consequently, its consistency and sufficiency estimates are 0.0, despite the definitions implying a value of 1.0 for both measures.

Decision trees. Next, we used decision trees to study the relationship between the size of the explanation space and the number of samples needed for accurate estimation of faithfulness. Each of our prediction models was a decision tree, and the same tree was used for explanations, implying perfect consistency and sufficiency. We learned six trees of different sizes (2^n leaves, for $n = 6, 7, \ldots, 11$) on the `Adult` dataset (Kohavi et al., 1996), using 66.6% of the examples for training. From the remaining 33.3% of the examples we varied the number of sampled records used to estimate consistency/sufficiency (the two estimates are identical in this setting).

Table 1. The mean±std of the consistency, sufficiency, and uniqueness measures of the four highlighted text explainers, evaluated over 5 samples of 1000 examples.

Explainer	Consistency	Sufficiency	Uniqueness
Top Coefficient	**0.69** $\pm$ 0.01	**0.71** $\pm$ 0.01	0.5 $\pm$ 0.01
Anchors	0.54 $\pm$ 0.01	0.61 $\pm$ 0.01	0.44 $\pm$ 0.01
First Word	0.37 $\pm$ 0.01	0.48 $\pm$ 0.01	0.39 $\pm$ 0.01
All Words	0.0 $\pm$ 0.0	0.0 $\pm$ 0.0	1.0 $\pm$ 0.0

The results appear in Figure 1. For the smallest tree (64 leaves), the global estimator is accurate even with very few samples. However, as the size of the tree grows, there are more possible explanations (root-to-leaf paths), which increases the sample complexity of the estimation task. For example, the largest configuration (2048 leaves) requires 4,300 samples to reach even a 0.9 estimate of sufficiency and consistency. Similar trends were observed for different datasets and when k-nearest neighbors was used as both the model and explainer (Appendix D.3).

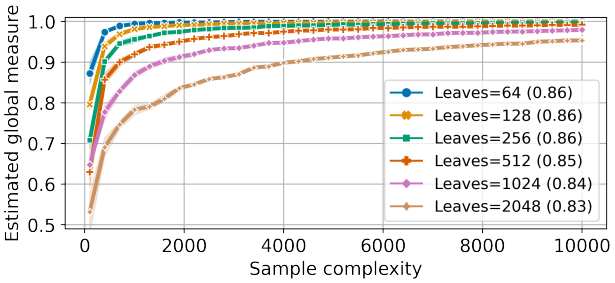


Figure 1. Estimated consistency and sufficiency of decision trees of different sizes over the `Adult` dataset, as a function of the sample complexity. As the sample complexity increases the estimation approaches the ground truth measures (1.0). Larger trees have more leaves and thus a larger explanation space. The decision tree model accuracy over the full test set is reported in the parenthesis in the legend. The displayed results are averaged over 5 executions with a confidence interval of 95%.

5.2. Common explanation systems

We next discuss two important considerations in applying our faithfulness estimators in practice: (1) the effect of explainer parameters on the quality of the estimates, and (2) the use of discretization to reduce explanation uniqueness

and thereby improve estimation. These apply generically for many common explanation systems. For concreteness, the predictors in our experiments are gradient boosted trees, which are frequently used by practitioners (described in Appendix D.2). The analysis is conducted on six standard datasets (described in Appendix D.1).

First, the choice of the explainer's parameters can impact not only sufficiency and consistency but also the accuracy of estimation. For example, a key parameter in Anchors is precision `threshold`, and high `threshold` leads to better sufficiency. Moreover, for high `threshold`, the anchors are typically smaller, as more explanations are possible (increasing the number of explanations worsens the estimators, as seen in Section 5.1). We next illustrate this phenomenon both locally and globally over the `Adult` dataset.

Figure 2, shows an example of the effect on local measures. (2b) shows the explanations, π_1 and π_2, of two different anchors over the record depicted in (2a). The two explainers differ only in their precision `threshold` parameter (0.5 and 0.95). (2c) presents the statistics of these explainers when applied over the record from (2a) and the `Adult` test set. One can see that π_2 refines π_1 since it includes more conditions, and hence $|C_{\pi_1}| > |C_{\pi_2}|$. Moreover, using higher `threshold` improved the sufficiency.

age	38
workclass	Private
fnlwgt	89814
education	HS-grad
education-num	9
marital-status	Married-civ-spouse
occupation	Farming-fishing
relationship	Husband
race	White
sex	Male
capital-gain	0
capital-loss	0
hours-per-week	50
native-country	United-States
y	$\leq 50K$
prediction	$\leq 50K$

(a) Record from `Adult` dataset

π_1 `threshold` $= 0.5$

`education-num` ≤ 9.00

π_2 `threshold` $= 0.95$

`education-num` ≤ 9.00 `and` `capital-gain` ≤ 0.00 `and` `fnlwgt` ≤ 116736

(b) Anchors π_1 and π_2

	π_1	π_2
N^c_π	2364	53
$N^c_{\pi,\leq 50K}$	2364	53
N^s_π	7438	1690
$N^s_{\pi,\leq 50K}$	7049	1655
$\hat{m}^c$	1.0	1.0
$\hat{m}^s$	0.95	0.98

(c) Statistics of π_1 and π_2

Figure 2. Two Anchors explainers with different precision `threshold` parameters, and their performance statistics over an example record from the `Adult` dataset.

Moving to global measures, Figure 3 shows faithfulness estimates for Anchors applied to gradient boosted trees trained on the `Adult` dataset (Appendix D.4 has results for other datasets), as a function of the precision. As expected, as the precision increases, so do sufficiency and uniqueness. Note that higher uniqueness reduces estimator accuracy.

Second, discretizing the output seems to be an effective way to mitigate uniqueness. This is illustrated in Table 2, which shows the results of 5 different discretization methods of SHAP values (described in Appendix D.5) and a non-discretized baseline over 6 datasets. As one would expect

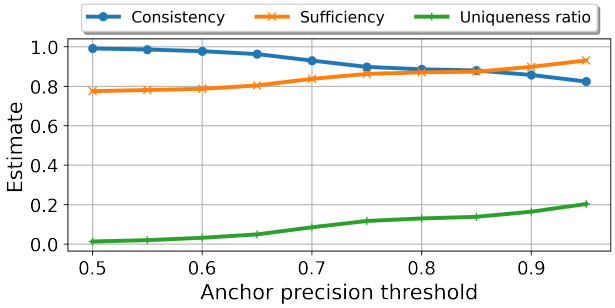


Figure 3. Estimated global consistency and sufficiency and the number of unique explanations of the Anchors explainer over gradient boosted trees model for the `Adult` dataset as a function of precision `threshold` parameter.

(based on Claim 2), without discretization the measures are extremely low. Moreover, while all examined discretization methods improve on the non-discretized baselines, the optimal method depends on the dataset and explainer at hand. Hence, one is encouraged to experiment with different methods to identify the best approach. In Appendix D.5 we also provide the uniqueness ratio of each discretization method, along with discretizations of LIME and Counterfactuals that exhibit similar behavior.

Table 2. SHAP consistency scores for various discretizations, averaged over 5 executions (std is lower than 0.01 in all cases).

Dataset	Original	2-FP	1-FP	Sign	Rank	Sign-of-top-5
Heart	0.0	0.0	**0.48**	0.02	0.02	0.39
Chess	0.0	0.0	0.0	0.33	0.32	**0.35**
Avila	0.01	0.01	0.05	**0.71**	0.56	0.58
Bank marketing	0.03	0.40	**0.93**	0.49	0.38	0.86
Adult	0.02	0.11	**0.95**	0.68	0.15	0.89
Covtype	0.01	0.03	**0.68**	0.13	0.09	0.41

5.3. Explanation quality is data-dependent

The consistency and sufficiency definitions imply that the faithfulness of an explainer depends on the test distribution. When two explanation methods are available, one might be more faithful for some populations, while the other works better for other populations. Moreover, as data distribution changes over time, explanation methods must also adapt. It is not advisable to deploy explainers in real-life settings without verifying faithfulness on the target distribution.

In Figure 4 we demonstrate this by splitting the `Adult` test-set into two different populations. Each negative example is randomly assigned to the first population with probability 0.75 and each positive example is assigned to the first population with probability 0.25. The second population contains all examples not assigned to the first population. As a result, the first population contained $\sim 90\%$ records labeled as "$\leq 50K$" and the second population was almost

balanced. We then used 4 different explainers on the two populations (Anchors with `threshold` of 0.7, SHAP and LIME with 1-FP discretization, and Counterfactuals with discretization of the sign of modification). We estimated the consistency of each population for 5 repetitions and recorded the average and standard deviation. For all the explainers, the consistency is different between the two populations. We remark that Anchors has the highest consistency in the first population (0.934 ± 0.003), while SHAP has the highest consistency in the second (0.885 ± 0.006). Note that the proposed measures serve as a tool for comparing different explanation system, while the exact values of the measures (which depends on the dataset) is of only secondary importance.

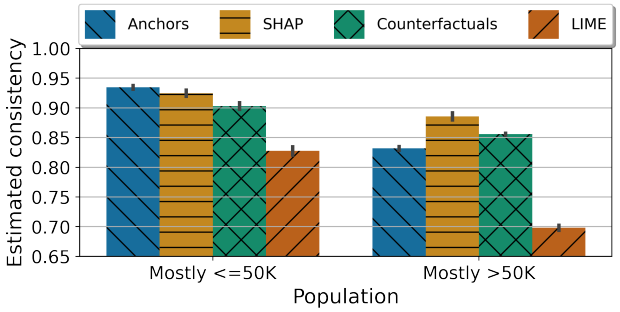


Figure 4. Estimated global consistency of four explanation systems very extensively between two distributions over the `Adult` dataset. The displayed results are the mean of 5 executions with std bars.

6. Conclusion and open problems

We suggest two new measures evaluating the faithfulness of explanations, both locally and globally. These are the consistency and sufficiency measures. We showed estimators for these measures and bounded the sample complexity of the global measures by an *unlabeled* sample of size $O(\texttt{range}(e))$, for constant ϵ error. We analyzed these measures on several known methods: decision trees, Anchors, highlighted text, SHAP, LIME, gradient-based method, k-nn, and counterfactuals. We empirically examined these measures, highlighting essential properties, e.g., faithfulness can be unverifiable if there are too many explanations and faithfulness quality is data-dependent.

In this paper, no assumptions about the explainer were made. However, additional assumptions might pave the way for better estimators. This can be especially important for continuous explanation spaces, where the bound on the estimator, $\texttt{range}(e)$, is ill-defined. As a concrete example, focus on a feature importance explainer, $e : \mathcal{X} \to \mathbb{R}^d$. In this case $\texttt{range}(e)$ can be infinite. Nonetheless, assuming that similar examples have similar feature importance (which can be formalized with the Lipschitz assumption) might allow the design of estimators with superior bounds. On the practical side, a primary mission is to apply these measures in real-life applications.

Acknowledgements

We would like to thank Yoav Goldberg for introducing us to the problem of faithfulness in NLP which initiated this project.

Funding transparency statement

Sanjoy Dasgupta has been supported by NSF CCF-1813160 and NSF IIS-1956339. Nave Frost has been funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 804302). Michal Moshkovitz has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grant number 993/17), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pp. 9505–9515, 2018.

Agarwal, C., Johnson, N., Pawelczyk, M., Krishna, S., Saxena, E., Zitnik, M., and Lakkaraju, H. Rethinking stability for attribution-based explanations. *arXiv preprint arXiv:2203.06877*, 2022.

Agarwal, S., Jabbari, S., Agarwal, C., Upadhyay, S., Wu, Z. S., and Lakkaraju, H. Towards the unification and robustness of perturbation and gradient based explanations. *arXiv preprint arXiv:2102.10618*, 2021.

Alvarez-Melis, D. and Jaakkola, T. S. On the robustness of interpretability methods. *arXiv preprint arXiv:1806.08049*, 2018.

Amparore, E., Perotti, A., and Bajardi, P. To trust or not to trust an explanation: using LEAF to evaluate local linear XAI methods. *PeerJ Computer Science*, 7:e479, 2021.

Ancona, M., Ceolini, E., Öztireli, C., and Gross, M. Towards better understanding of gradient-based attribution methods for deep neural networks. *arXiv preprint arXiv:1711.06104*, 2017.

Anders, C., Pasliev, P., Dombrowski, A.-K., Müller, K.-R., and Kessel, P. Fairwashing explanations with off-

manifold detergent. In *International Conference on Machine Learning*, pp. 314–323. PMLR, 2020.

Barceló, P., Monet, M., Pérez, J., and Subercaseaux, B. Model interpretability through the lens of computational complexity. *arXiv preprint arXiv:2010.12265*, 2020.

Bhatt, U., Weller, A., and Moura, J. M. Evaluating and aggregating feature-based model explanations. *arXiv preprint arXiv:2005.00631*, 2020.

Blackard, J. A. and Dean, D. J. Comparative accuracies of artificial neural networks and discriminant analysis in predicting forest cover types from cartographic variables. *Computers and electronics in agriculture*, 24(3):131–151, 1999.

Carmichael, Z. and Scheirer, W. J. On the objective evaluation of post hoc explainers. *arXiv preprint arXiv:2106.08376*, 2021.

Chen, J., Song, L., Wainwright, M., and Jordan, M. Learning to explain: An information-theoretic perspective on model interpretation. In *International Conference on Machine Learning*, pp. 883–892. PMLR, 2018.

Dasgupta, S., Frost, N., Moshkovitz, M., and Rashtchian, C. Explainable k-means and k-medians clustering. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5587–5597, 2020.

De Stefano, C., Maniaci, M., Fontanella, F., and di Freca, A. S. Reliable writer identification in medieval manuscripts through page layout features: The "avila" bible case. *Engineering Applications of Artificial Intelligence*, 72:99–110, 2018.

Deutch, D. and Frost, N. Constraints-based explanations of classifications. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pp. 530–541. IEEE, 2019.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Fel, T., Vigouroux, D., Cadène, R., and Serre, T. How good is your explanation? algorithmic stability measures to assess the quality of explanations for deep neural networks. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 720–730, 2022.

Hase, P., Xie, H., and Bansal, M. The out-of-distribution problem in explainability and search methods for feature importance explanations. *Advances in Neural Information Processing Systems*, 34, 2021.

Hu, X., Rudin, C., and Seltzer, M. Optimal sparse decision trees. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/ac52c626afc10d4075708ac4c778ddfc-Paper.pdf.

Jacovi, A. and Goldberg, Y. Aligning faithful interpretations with their social attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310, 2021.

Janosi, A., Steinbrunn, W., Pfisterer, M., and Detrano, R. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 64(5):304–310, 1989. ISSN 0002-9149. doi: https://doi.org/10.1016/0002-9149(89)90524-9. URL https://www.sciencedirect.com/science/article/pii/0002914989905249.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., and Gama, J. How can i choose an explainer? an application-grounded evaluation of post-hoc explanations. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 805–815, 2021.

Kim, B. and Doshi-Velez, F. Machine learning techniques for accountability. *AI Magazine*, 42(1):47–52, 2021.

Kohavi, R. et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207, 1996.

Lakkaraju, H., Arsov, N., and Bastani, O. Robust and stable black box explanations. In *International Conference on Machine Learning*, pp. 5628–5638. PMLR, 2020.

Leavitt, M. L. and Morcos, A. Towards falsifiable interpretability research. *arXiv preprint arXiv:2010.12016*, 2020.

Lipton, Z. C. The mythos of model interpretability. *Queue*, 16(3):31–57, 2018.

Lundberg, S. M. and Lee, S.-I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pp. 4765–4774, 2017.

Margot, V. and Luta, G. A new method to compare the interpretability of rule-based algorithms. *arXiv preprint arXiv:2004.01570*, 2020.

Mohseni, S., Block, J. E., and Ragan, E. D. A human-grounded evaluation benchmark for local explanations of machine learning. *arXiv preprint arXiv:1801.05075*, 2018.

Molnar, C. *Interpretable Machine Learning*. Lulu. com, 2019. https://christophm.github.io/interpretable-ml-book/.

Moro, S., Cortez, P., and Rita, P. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62:22–31, 2014.

Moshkovitz, M., Yang, Y.-Y., and Chaudhuri, K. Connecting interpretability and robustness in decision trees through separation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7839–7849, 2021.

Mothilal, R. K., Sharma, A., and Tan, C. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617, 2020.

Nie, W., Zhang, Y., and Patel, A. A theoretical explanation for perplexing behaviors of backpropagation-based visualizations. In *International Conference on Machine Learning*, pp. 3809–3818. PMLR, 2018.

Pang, B. and Lee, L. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, 2005.

Poppi, S., Cornia, M., Baraldi, L., and Cucchiara, R. Revisiting the evaluation of class activation mapping for explainability: A novel metric and experimental analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2299–2304, 2021.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–52, 2021.

Pruthi, D., Bansal, R., Dhingra, B., Soares, L. B., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022.

Quinlan, J. R. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

Ribeiro, M. T., Singh, S., and Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144. ACM, 2016.

Ribeiro, M. T., Singh, S., and Guestrin, C. Anchors: High-precision model-agnostic explanations. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Rivest, R. L. Learning decision lists. *Machine learning*, 2 (3):229–246, 1987.

Shapley, L. S. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

Shrikumar, A., Greenside, P., Shcherbina, A., and Kundaje, A. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint arXiv:1605.01713*, 2016.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Slack, D., Hilgard, S., Lakkaraju, H., and Singh, S. Counterfactual explanations can be manipulated. *arXiv preprint arXiv:2106.02666*, 2021.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

Tomsett, R., Harborne, D., Chakraborty, S., Gurram, P., and Preece, A. Sanity checks for saliency metrics. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 6021–6029, 2020.

Ustun, B. and Rudin, C. Learning optimized risk scores. *Journal of Machine Learning Research*, 20(150):1–75, 2019.

Velmurugan, M., Ouyang, C., Moreira, C., and Sindhgatta, R. Developing a fidelity evaluation approach for interpretable machine learning. *arXiv preprint arXiv:2106.08492*, 2021.

Wang, E., Khosravi, P., and Van den Broeck, G. Towards probabilistic sufficient explanations. In *Extending Explainable AI Beyond Deep Models and Classifiers Workshop at ICML (XXAI)*, 2020a.

Wang, J., Tuyls, J., Wallace, E., and Singh, S. Gradient-based analysis of nlp models is manipulable. *arXiv preprint arXiv:2010.05419*, 2020b.

Wolf, L., Galanti, T., and Hazan, T. A formal approach to explainability. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 255–261, 2019.

Yeh, C.-K., Hsieh, C.-Y., Suggala, A., Inouye, D. I., and Ravikumar, P. K. On the (in) fidelity and sensitivity of explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Yoon, J., Jordon, J., and van der Schaar, M. Invase: Instance-wise variable selection using neural networks. In *International Conference on Learning Representations*, 2018.

Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593, 2021.

A. Proofs

A.1. Proof of Claim 1

For the sake of completeness, we repeat some of the definitions. For a fixed explanation π, the probability that it resulted from an instance labeled y is equal to

$$\Pr(y|\pi) = \sum_{x:f(x)=y \wedge e(x)=\pi} \frac{\Pr(x)}{\Pr(\pi)},$$

where $\Pr(\pi) = \sum_{x:e(x)=\pi} \Pr(x)$ and the distribution over the instances x's is μ. There are two natural ways to define decoders from explanations to labels:

- Gibbs decoder

$$d_G(\pi) = y \text{ with probability } \Pr(y|\pi)$$

- Optimal deterministic decoder

$$d_O(\pi) = \arg\max_y \Pr(y|\pi)$$

The error of any decoder d is equal to

$$\sum_\pi \Pr(\pi) \sum_{y \in \mathcal{Y}} \Pr(y|\pi) \Pr(d(\pi) \neq y|\pi).$$

Specifically, the error of the Gibbs decoder is equal to

$$E_G = \sum_\pi \Pr(\pi) \sum_{y \in \mathcal{Y}} \Pr(y|\pi)(1 - \Pr(y|\pi)).$$

The error of the optimal deterministic decoder is equal to

$$E_O = \sum_\pi \Pr(\pi)(1 - \max_y \Pr(y|\pi)).$$

Now we are ready to prove the claim that

$$E_O \leq E_G \leq 2E_O.$$

For ease of notations, arrange the probabilities $(\Pr(y|\pi))_{y \in \mathcal{Y}}$ in decreasing order $p_1 \geq p_2 \geq \ldots p_{|\mathcal{Y}|}$.

We first prove the left inequality in the claim. We will show that for every explanation π it holds that

$$1 - p_1 \leq \sum_j p_j(1 - p_j).$$

Or equivalently, we will show that

$$\sum_j p_j^2 \leq p_1.$$

The latter holds because

$$\sum_j p_j^2 \leq \sum_j p_j \cdot p_1 = p_1.$$

Now we move on to proving the right inequality in the claim. We will show

$$\sum_j p_j(1 - p_j) \leq 2(1 - p_1).$$

The LHS is equal to

$$\begin{aligned} p_1(1 - p_1) \quad &+ \quad \sum_{j>1} p_j(1 - p_j) \leq 1 - p_1 + \sum_{j>1} p_j \\ &= \quad 2(1 - p_1) \end{aligned}$$

A.2. Proof of Theorem 1

Recall that $R(x, \pi)$ is an arbitrary relation on $\mathcal{X} \times \mathcal{E}$. For a distribution μ on $\mathcal{X}$, we wish to estimate

$$m_\mu^R = \Pr_{X,X' \sim \mu}(f(X') = f(X) | R(X', e(X))).$$

To begin with, let $q(y|\pi)$ denote the probability that a random point $x \sim \mu$ has predicted label y given $R(x, \pi)$:

$$q(y|\pi) = \Pr_{X \sim \mu}(f(X) = y | R(X, \pi)).$$

Then we can rewrite our generic faithfulness measure as

$$m_\mu^R = \mathop{\mathbb{E}}_{X \sim \mu}[q(f(X)|e(X))].$$

Let $p(\pi)$ be the fraction of points for which explanation π is provided, that is, $p(\pi) = \mu(\{x : e(x) = \pi\})$ and let $q(\pi)$ be the fraction for which $R(x, \pi)$ holds: $q(\pi) = \mu(\{x : R(x, \pi)\})$. Note that $p(\pi)$ is a distribution over $\mathcal{E}$ whereas $q(\pi) \in [0, 1]$ and $q(\pi) \geq p(\pi)$.

Given samples $x_1, \ldots, x_n \sim \mu$, and any y, π, define

$$N_\pi = |\{i : R(x_i, \pi)\}|$$
$$N_{\pi,y} = |\{i : R(x_i, \pi), f(x_i) = y\}|$$

Our estimator for m_μ^R is then

$$\widehat{M} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(N_{e(x_i)} > 1) \frac{N_{e(x_i),f(x_i)} - 1}{N_{e(x_i)} - 1}.$$

We start by deriving the expected value of $\widehat{M}$.

Theorem 2. *For estimator $\widehat{M}$,*

$$\mathbb{E}[\widehat{M}] = \mathop{\mathbb{E}}_{X \sim \mu}\left[(1 - (1 - q(e(X)))^{n-1}) q(f(X)|e(X))\right].$$

Proof. Fix any $i \in [n]$. The term $\mathbb{E}_{\setminus i}$ denotes expectation over all points other than i. We will also use $\mathbb{E}_i$ to denote expectation over point i alone. Let $y = f(x_i)$ and $\pi = e(x_i)$, and let k be the number of *other* points (that is, $j \neq i$) to which π also applies: that is, $k = N_\pi - 1$. Suppose these points are $x_{i_1}, \ldots, x_{i_k}$. If $k > 0$, then $\mathbb{E}_{\setminus i}\left[\frac{N_{\pi,y}-1}{N_\pi - 1} \middle| N_\pi = k + 1\right]$ is equal to

$$\frac{1}{k} \sum_{j=1}^{k} (\Pr(f(x_{i_j}) = y | R(x_{i_j}, \pi)) = q(y|\pi).$$

We then have that $\mathbb{E}[\widehat{M}]$ is equal to

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\mathbf{1}(N_{e(x_i)} > 1) \frac{N_{e(x_i),f(x_i)} - 1}{N_{e(x_i)} - 1}\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{i}\left[\mathbf{1}(N_{e(x_i)} > 1) \mathop{\mathbb{E}}_{\setminus i}\left[\frac{N_{e(x_i),f(x_i)} - 1}{N_{e(x_i)} - 1} \middle| N_{e(x_i)} > 1\right]\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \mathop{\mathbb{E}}_{i}\left[\mathbf{1}(N_{e(x_i)} > 1) q(f(x_i)|e(x_i))\right] \\
&= \mathop{\mathbb{E}}_{X \sim \mu}\left[(1 - (1 - q(e(X)))^{n-1}) q(f(X)|e(X))\right],
\end{aligned}$$

as claimed. □

Next, we upper-bound the variance of $\widehat{M}$.

Theorem 3. $\text{var}(\widehat{M}) \leq 4/n$.

Proof. Suppose $\widehat{M}$ is based on n samples $x_1, \ldots, x_n \sim \mu$. It is not hard to check that changing any one sample, $x_i \to x_i'$, can change $\widehat{M}$ by at most $4/n$. Thus $\widehat{M}$ satisfies a bounded-differences property, whereupon its variance can be bounded by a form of the Efron-Stein inequality (Boucheron, Lugosi, Massart, Cor 3.2). □

We then sum the bias and variance to get a bound on mean-squared error. From Theorem 2 and the fact that $q(y|\pi) \in [0, 1]$, we can bound the bias, $\left|\mathbb{E}[\widehat{M}] - m_\mu^R\right|$ of $\widehat{M}$ by

$$\begin{aligned}
&\left|\mathop{\mathbb{E}}_{X}\left[(1-(1-q(e(X)))^{n-1})q(f(X)|e(X))\right] - \mathop{\mathbb{E}}_{X}[q(f(X)|e(X))]\right| \\
&= \mathop{\mathbb{E}}_{X}\left[(1-q(e(X)))^{n-1}q(f(X)|e(X))\right] \\
&\leq \mathop{\mathbb{E}}_{X}\left[e^{-(n-1)q(e(X))}\right] \\
&= \sum_{\pi\in\mathcal{E}} p(\pi)e^{-(n-1)q(\pi)}.
\end{aligned}$$

Theorem 1 then follows by summing the variance and squared bias.

A.3. Proof of Claim 2

Proof. The claim will hold for any model f which is balanced, i.e., there is the same number of examples labeled 1 and examples labeled -1. Take the distribution over the examples to be the uniform one.

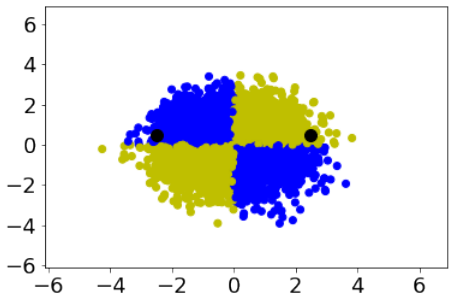
The explainer e_1 returns a different explanation to each example in $\mathcal{X}$. To define the explainer e_2, partition $\mathcal{X}$ into pairs (x_1, x_2) where $f(x_1) \neq f(x_2)$. Such a partition is possible because f is balanced. Each pair receives the exact explanation $e_2(x_1) = e_2(x_2)$.

Suppose that for the two explainers $A(x, \pi) \Leftrightarrow e(x) = \pi$. By definition, the sufficiency and consistency of e_1 is one and the sufficiency and consistency of e_2 is 0.5.

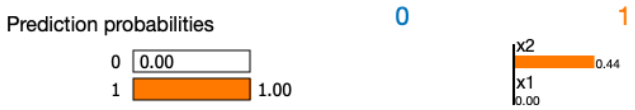
Note that when given a finite sample, since the set of instances $\mathcal{X}$ is infinite, with zero probability, the example set will contain a pair. Those it is impossible to distinguish if the true explainer is e_1 or e_2. □

B. Example where LIME does not have perfect consistency

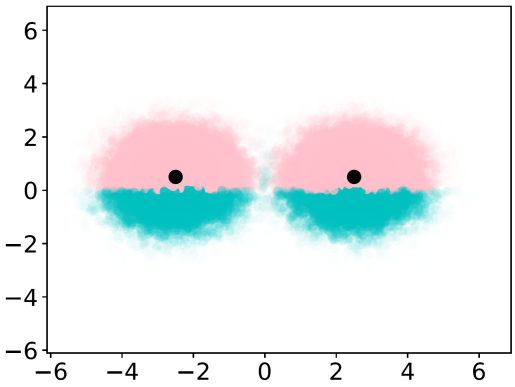
We show a model and two instances that get different labels but the same explanation by LIME. In Figure 5a, we show the XOR model f^{XOR} which is 1 if the two features have the same sign (in blue). We are using the LIME method to explain two instances $(2.5, 0.5)$ and $(-2.5, 0.5)$. The model f^{XOR} assigns these two instances different labels. The output of LIME when given instance $(2.5, 0.5)$ and $(-2.5, 0.5)$ is the same: second feature has the same positive importance of 0.44 on both instances and the first feature does not have importance, see Figures 5b,5d. The reason for such a behavior is that LIME fits a linear classifier around the labeled instance (x, y) where the goal is to predict the class y. From the view point of LIME for both of the instances, a linear classifier is fitted for similar training data, see Figure 5c.



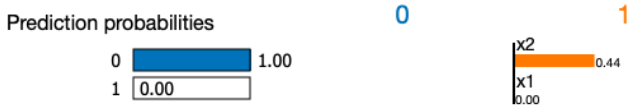
(a) XOR classification



(b) Explanation for $(2.5, 0.5)$



(c) Neighborhoods



(d) Explanation for $(-2.5, 0.5)$

Figure 5. (a) XOR model with two instances (in red) with different labels (b,d) LIME provides the same explanation to these instances: second feature has the same positive importance, 0.44, on both instances and the first feature does not have importance. (c) During the run of LIME explainer on the two instances, the training data supplied to the linear predictor.

C. Local estimators

In this section we explore estimators for the local measures. Namely, Algorithm 1 estimates the local consistency and sufficiency measures of explainer e for model f at instance x. It uses, as an input, *unlabelled* test data S drawn from distribution μ. To estimate consistency it returns the fraction of instances with similar label out of all instances with similar explanation. To estimate sufficiency, it returns the fraction of instances with similar label out of all examples that $e(x)$ applied to.

Algorithm 1 Estimating local consistency and sufficiency

input: model f, instance x, unlabelled test data S
output: estimate of consistency and sufficiency
$con_{counter}, con_{tot} = 0, 0$
$suf_{counter}, suf_{tot} = 0, 0$
for $x' \in S$ **do**
  if $e(x') = e(x)$ **then**
    $con_{counter} += (f(x) == f(x'))$
    $con_{tot} ++$
  end if
  if $A(x', e(x))$ **then**
    $suf_{counter} += (f(x) == f(x'))$
    $suf_{tot} ++$
  end if
end for
return $con_{counter}/con_{tot}, suf_{counter}/suf_{counter}$

If we have a random sample S from C_π then, by Hoeffding's inequality, it is enough to take sample size $|S| = O(1/\epsilon^2)$ to approximate the consistency measure up to an additive error of ϵ with constant probability. This is summarized in the following corollary.

Corollary 2. *Fix $\epsilon \in (0, 1)$ and an instance x. Given a sample of size $O(1/\epsilon^2)$ from $C_{e(x)}$, one can estimate $m^c(x)$ up to an additive error ϵ with probability* 0.9.

The difficulty with the above corollary is the assumption that one can obtain enough samples from $C_{e(x)}$. This assumption is sometimes unrealistic. To get an instance from $C_{e(x)}$, one can use *rejection sampling*. Where instances are received from arbitrary distribution, but then reject any instance that is not in $C_{e(x)}$. Although this is a reasonable technique, it might take a long time till an instance from $C_{e(x)}$ is received.

D. More experimental details

D.1. Datasets

Datasets in the empirical evaluation are depicted in Table 3.

Table 3. Datasets properties

Dataset	# of classes	n	d
Heart (Janosi et al., 1989)	2	303	13
Chess (Dua & Graff, 2017)	17	28,056	6
Avila (De Stefano et al., 2018)	12	20,867	10
Bank marketing (Moro et al., 2014)	2	45,211	16
Adult (Kohavi et al., 1996)	2	48,842	14
Covtype (Blackard & Dean, 1999)	7	581,012	54
rt-polaritydata (Pang & Lee, 2005)	2	10,433	15,888

D.2. Model training

In sections 5.2 and 5.3 we have explained gradient boosted trees models trained over 6 datasets. For each dataset, 66% of it was used for model training and cross-validation. Hyper-parameters were selected based on best mean accuracy over 3 cross-validation executions. The considered hyper-parameters are all combinations of the following:

- `learning_rate`: $2^{-5}, 2^{-4}, \ldots, 2^{2}$.
- `n_estimators`: $50, 100, 150, 200, 250, 300$.
- `max_depth`: $3, 4, 5, 6, 7$.

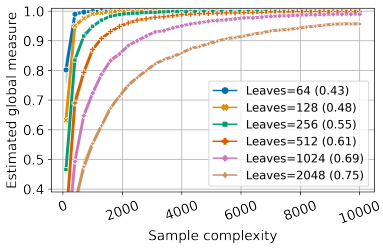
The selected hyper-parameters and test accuracy is presented in Table 4.

Table 4. Gradient boosted trees hyper-parameters and accuracy

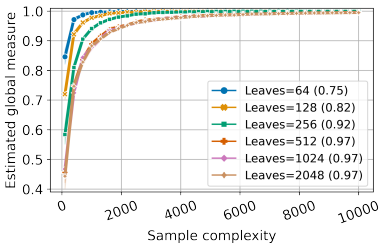
Dataset	`learning_rate`	`n_estimators`	`max_depth`	Test accuracy
Heart (Janosi et al., 1989)	0.0625	250	3	0.8
Chess (Dua & Graff, 2017)	0.0625	300	7	0.9
Avila (De Stefano et al., 2018)	0.125	300	5	0.99
Bank marketing (Moro et al., 2014)	0.0625	250	5	0.91
Adult (Kohavi et al., 1996)	0.25	50	5	0.87
Covtype (Blackard & Dean, 1999)	0.125	300	7	0.94

D.3. Sample complexity experiment

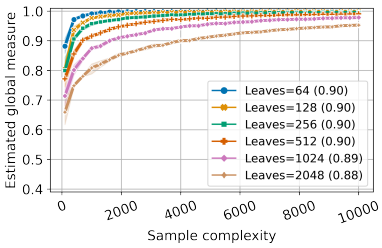
In Section 5.1 Figure 1 we have studied the sample complexity of decision tree model and explainer over `Adult` dataset. Figure 6 depict the same concept over additional datasets.



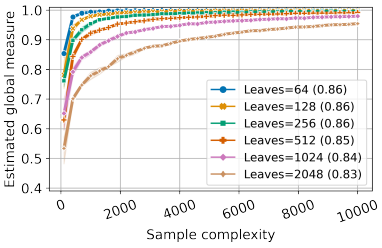
(a) Chess



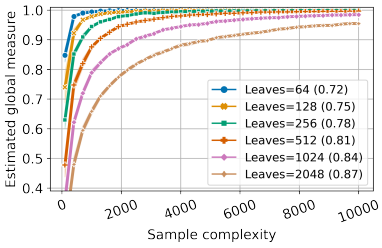
(b) Avila



(c) Bank marketing

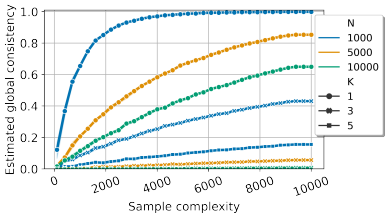


(d) Adult

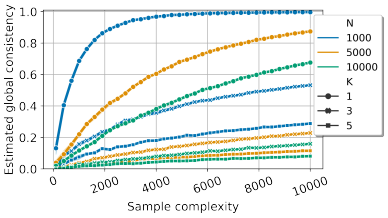


(e) Covtype

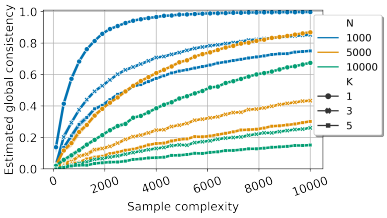
Figure 6. Estimated global consistency & sufficiency of decision trees with different sizes on 5 datasets. As sample complexity grows the estimation is getting closer to the ground truth measures (1.0). Larger trees has more leaves, which implies a larger explanations domain. Decision tree accuracy over the full test set is reported in the legend parenthesis. The displayed results are the mean of 5 executions with confidence interval of 95%.



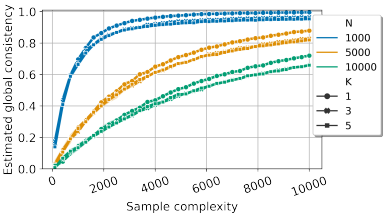
(a) Chess



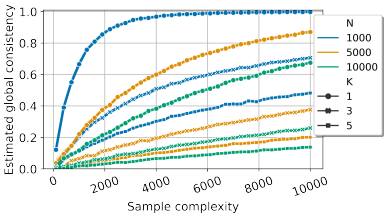
(b) Avila



(c) Bank marketing



(d) Adult



(e) Covtype

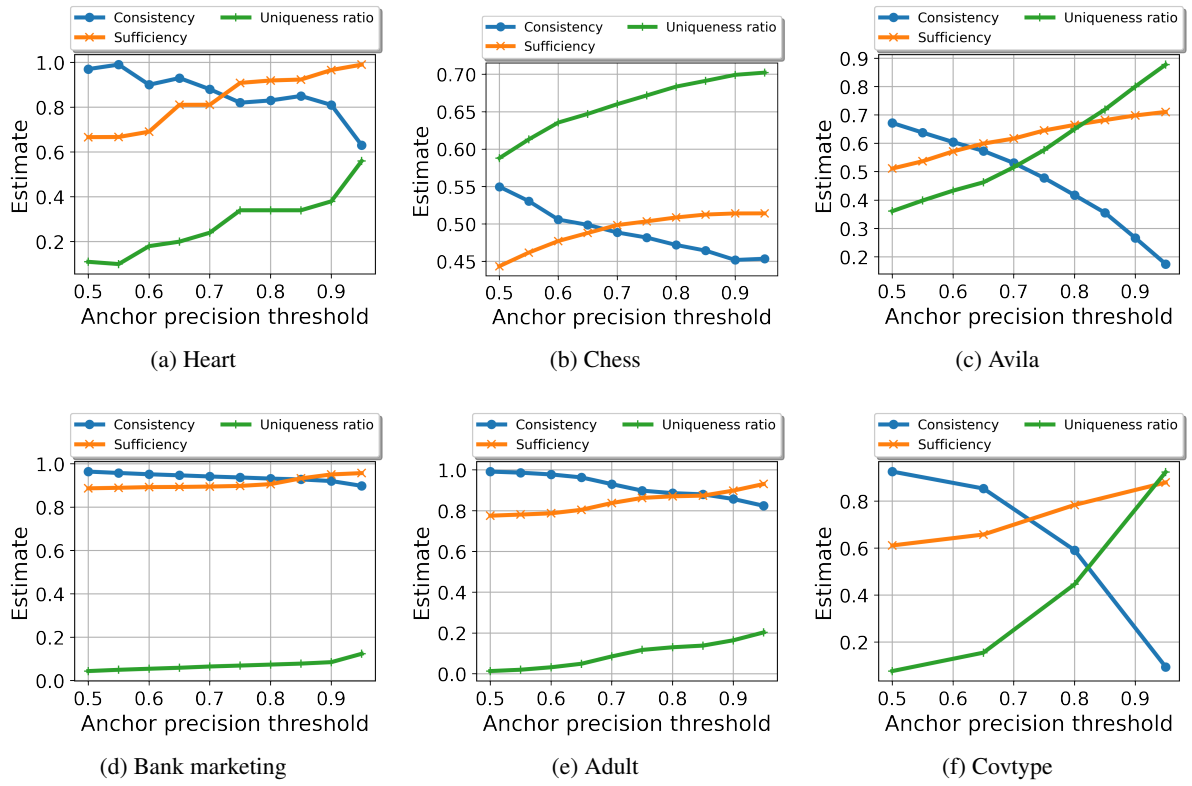
Figure 7. Estimated global consistency of k nearest neighbors with different sizes on 5 datasets. As sample complexity grows the estimation is getting closer to the ground truth measures (1.0). Larger k and larger training set size ,N, implies a larger explanations domain. Accuracy over the full test set is reported in the legend parenthesis. The displayed results are the mean of 5 executions with confidence interval of 95%.

Similarly, Figure 7 depict the sample complexity required for the evaluation of k nearest neighbors model and explainer. As

the explainer and model are the same, the explainer consistency is 1 by definition. Figure 7 shows that as k or N (number of training examples) increases, the explanations space grows, and as a result, more samples are required to accurately estimate the explainer consistency.

D.4. Anchors dependency on precision `threshold` parameter

Figure 8 depict how the explainer's parameters affect global measures. The Figure displays the measures of Anchors explainer, applied over gradient boosted trees trained over six datasets, as a function of the precision `threshold` parameter. Similarly to the findings obtained in Figure 3, one may see that as the precision increases, the sufficiency and uniqueness increases, while the estimated consistency decreases.



(a) Heart (b) Chess (c) Avila

(d) Bank marketing (e) Adult (f) Covtype

Figure 8. Estimated global consistency and sufficiency and the number of unique explanations of the Anchors explainer over gradient boosted trees model for 6 dataset as a function of precision `threshold` parameter. As the required precision grows the number of unique explanations (green) grows as well as the estimated sufficiency.

D.5. Explainers discretization

Next we discuss several discretizations we have evaluated.

Feature importance Recall that for an explanation of type *feature importance*, given an instance $x \in \mathbb{R}^d$ it returns a vector $\phi \in \mathbb{R}^d$ with ϕ_i the importance of the i-th feature. For feature importance explainers, i.e. SHAP and LIME we compared the following discretization methods.

- *Original*: return ϕ as is.
- *2-FP*: discretize ϕ to have 2 floating-points representation, i.e., return $\phi' \in \mathbb{R}^d$, such that $\phi'_i = \frac{\lfloor 100 \cdot \phi_i \rfloor}{100}$.
- *1-FP*: discretize ϕ to have a single floating-point representation, i.e., return $\phi' \in \mathbb{R}^d$, such that $\phi'_i = \frac{\lfloor 10 \cdot \phi_i \rfloor}{10}$.

- *Sign*: return $\phi' \in \{-1, 1\}^d$ such that $\phi'_i = \texttt{sign}(\phi_i)$.
- *Rank*: return $\phi' = \texttt{argsort}(\phi)$.
- *Sign-of-top-5*: let $\phi^+ \in \mathbb{R}^d$ be the vector of absolute values of ϕ, i.e. $\phi^+_i = |\phi_i|$, and let $\phi^{+,R} = \texttt{argsort}(\phi^+)$, i.e. $\phi^{+,R}$ is the rank of ϕ absolute values. Sign-of-top-5 return $\phi' \in \{-1, 0, 1\}^d$ such that $\phi'_i = \begin{cases} \texttt{sign}(\phi_i) & \phi^{+,R}_i > d - 5 \\ 0 & else \end{cases}$.

Tables 5 and 6 depict the consistency and uniqueness ratio of the above discretizations for SHAP and LIME respectively.

Table 5. SHAP consistency scores and uniqueness ratio for various discretizations, averaged over 5 executions (std is lower than 0.01 in all cases).

Dataset	Original		2-FP		1-FP		Sign		Rank		Sign-of-top-5	
	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.
Heart	0.0	1.0	0.0	1.0	**0.48**	0.70	0.02	0.99	0.02	0.99	0.39	0.75
Chess	0.0	1.0	0.0	1.0	0.0	1.0	0.33	0.02	0.32	0.15	**0.35**	0.06
Avila	0.01	0.99	0.01	0.99	0.05	0.97	**0.71**	0.21	0.56	0.49	0.58	0.07
Bank marketing	0.03	0.97	0.40	0.65	**0.93**	0.09	0.49	0.59	0.38	0.68	0.86	0.17
Adult	0.02	0.98	0.11	0.93	**0.95**	0.08	0.68	0.44	0.15	0.89	0.89	0.15
Covtype	0.01	0.99	0.03	0.97	**0.68**	0.36	0.13	0.89	0.09	0.92	0.41	0.03

Table 6. LIME consistency scores and uniqueness ratio for various discretizations, averaged over 5 executions (std is lower than 0.08 in all cases).

Dataset	Original		2-FP		1-FP		Sign		Rank		Sign-of-top-5	
	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.
Heart	0.0	1.0	0.0	1.0	0.34	0.81	0.02	0.99	0.0	1.0	**0.46**	0.66
Chess	0.0	1.0	0.11	0.0	0.11	0.0	**0.13**	0.02	**0.13**	0.18	**0.13**	0.07
Avila	0.0	1.0	0.23	0.0	0.23	0.0	**0.37**	0.17	0.01	0.98	0.28	0.33
Bank marketing	0.0	1.0	0.0	1.0	**0.87**	0.0	0.65	0.48	0.0	1.0	0.84	0.10
Adult	0.0	1.0	0.0	1.0	**0.77**	0.0	**0.77**	0.22	0.04	0.97	0.72	0.13
Covtype	0.0	1.0	0.0	1.0	0.12	0.87	0.0	1.0	0.0	1.0	**0.17**	0.73

Counterfactuals Recall that for *counterfatual* explanation, given an instance $x \in \mathbb{R}^d$ it returns a vector $x' \in \mathbb{R}^d$ such that $f(x) \neq f(x')$ and x' is close to x. To obtain a counterfatual explanations we have used `DiCE` (Mothilal et al., 2020). As the space of explanations $\mathcal{E} = \mathcal{X}$ discretization of $\mathcal{E}$ is essential for estimation of the explainability measures. To this end, we compared the following discretization methods.

- *Original*: return x' as is.
- Δ: return $x' - x$, i.e. consider only the features that were modified.
- Δ-*sign*: return $x'' \in \mathbb{R}^d$, such that $x''_i = \texttt{sign}(x'_i - x_i)$.
- *Is-feature-modified*: return $x'' \in \mathbb{R}^d$ such that $x''_i = \begin{cases} 1 & x_i = x'_i \\ 0 & else \end{cases}$.

Table 7 depict the consistency and uniqueness ratio of the above discretizations.

Table 7. Counterfactuals consistency scores and uniqueness ratio for various discretizations, averaged over 5 executions (std is lower than 0.07 in all cases).

Dataset	Original		Δ		Δ-sign		Is-feature-modified	
	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.	Cons.	Uniq.
Heart	0.0	1.0	0.01	1.0	0.19	0.88	**0.23**	0.66
Chess	0.09	0.70	**0.14**	0.17	**0.14**	0.02	0.13	0.01
Avila	0.20	0.22	0.20	0.23	**0.34**	0.07	0.24	0.0
Bank marketing	0.0	1.0	0.18	0.89	**0.89**	0.04	0.87	0.02
Adult	0.0	1.0	0.07	0.96	**0.91**	0.04	0.81	0.01
Covtype	0.0	1.0	0.38	0.46	**0.57**	0.03	0.52	0.02