When Should We Use Linear Explanations?

Julien Delaunay, Luis Galárraga, Christine Largouët

When Should We Use Linear Explanations?

Julien Delaunay
julien.delaunay@inria.fr
Inria/IRISA
Rennes, France

Luis Galárraga
luis.galarraga@inria.fr
Inria/IRISA
Rennes, France

Christine Largouët
christine.largouet@irisa.fr
L'Institut Agro/IRISA
Rennes, France

ABSTRACT

The increasing interest in transparent and fair AI systems has propelled the research in explainable AI (XAI). One of the main research lines in XAI is post-hoc explainability, the task of explaining the logic of an already deployed black-box model. This is usually achieved by learning an interpretable surrogate function that approximates the black box. Among the existing explanation paradigms, local linear explanations are one of the most popular due to their simplicity and fidelity. Despite their advantages, linear surrogates may not always be the most adapted method to produce reliable, i.e., unambiguous and faithful explanations. Hence, this paper introduces Adapted Post-hoc Explanations (APE), a novel method that characterizes the decision boundary of a black-box classifier and identifies when a linear model constitutes a reliable explanation. Besides, characterizing the black-box frontier allows us to provide complementary counterfactual explanations. Our experimental evaluation shows that APE identifies accurately the situations where linear surrogates are suitable while also providing meaningful counterfactual explanations.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; **Machine learning approaches**; **Learning linear models**; *Instance-based learning*; *Rule learning*.

KEYWORDS

Interpretability, Explainability, Linear Explanations, Rule-based Explanations, Counterfactual Explanations

ACM Reference Format:
Julien Delaunay, Luis Galárraga, and Christine Largouët. 2022. When Should We Use Linear Explanations?. In *Proceedings of the 31st ACM International Conference on Information and Knowledge Management (CIKM '22), October 17–21, 2022, Atlanta, GA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3511808.3557489

1 INTRODUCTION

In the last decade we have witnessed a breakthrough in the capabilities of AI/ML systems, in particular thanks to the emergence of deep learning. This has made ML-based systems ubiquitous, but has also increased the public scrutiny of the ethical aspects of automated decision support. This interest has given rise to initiatives such as the GDPR[1], and have propelled the research in explainable AI (XAI), a branch of AI that focuses on models and systems that can explain their decisions to the layman.

An important line of research in XAI is *post-hoc explainability*, a subfield of XAI that studies the techniques to compute explanations for the answers of an already deployed system. This may be necessary if the system is either too complex or its specifications inaccessible to the user. The explanation usually consists of a surrogate white-box model that mimics the black box, either *globally*, i.e., in the general case, or *locally*, that is, w.r.t. an instance of interest. LIME [17], one of the most popular post-hoc explanation methods, relies on local linear surrogates whose coefficients are used to rank the input features according to their contribution to the black box's outcome on an instance of interest.

Despite the popularity of local linear explanations, they may not always be the most adapted method to explain a black-box outcome. Consider the two cases depicted in Figure 1. In Figure 1a, the instance of interest lies in a zone where there is clearly a single local linear approximation for the black-box classifier. In contrast, the target instance in Figure 1b depicts a scenario where three possible linear explanations are possible. Since these approximations exhibit different inclinations, the attribution scores assigned to the input features are obviously contradictory – a situation that would harden interpretation. While we could provide one of the explanations for Figure 1b, that would tell an incomplete story.

Based on the aforementioned arguments, this article proposes APE, which stands for Adapted Post-hoc Explanations, a novel method to determine *a priori* whether a black-box classifier and a target instance admit a faithful and unambiguous local linear explanation. When this is not the case, APE recommends a different explanation paradigm – a rule-based explanation in our experiments. APE operates by characterizing the classifier's decision boundary, which is achieved by identifying the target's closest counterfactual instance. *Counterfactual instances* (also called *enemies*) are instances that are close to the target instance but are classified differently by the black box. Such instances can be used as contrastive explanations that highlight the minimal changes required on the target instance to change the classifier's outcome. All in all, our contributions are:

- A definition of *suitability* for explanations based on local linear surrogates. This definition builds upon existing notions such as *adherence* and *locality*, which we also define formally.
- The Growing Fields (GF) algorithm for counterfactual search. GF extends the Growing Spheres (GS) algorithm [13] to account for categorical attributes as well as the distribution of the input features using the standardized Euclidean distance as metric.
- The APE oracle, a linear suitability test that tells users whether a black-box classifier can be locally approximated by a single

CIKM '22, October 17–21, 2022, Atlanta, GA, USA

ACM ISBN 978-1-4503-9236-5/22/10…$15.00
https://doi.org/10.1145/3511808.3557489

[1] General Data Protection Regulation, https://gdpr-info.eu/

and faithful linear surrogate. To do so, APE characterizes the distribution of the instances around the decision boundary.
- The APE algorithm that returns a linear explanation if suitable. Otherwise APE proposes a rule-based explanation. In all cases APE computes complementary counterfactual explanations.

The article is structured as follows. After formulating the problem and introducing preliminary concepts in Section 2, Section 3 elaborates on our approach. Then we evaluate APE on a handful of datasets and classifiers in Section 4. This is followed by a survey of the related work in Section 5, and a discussion of our insights.

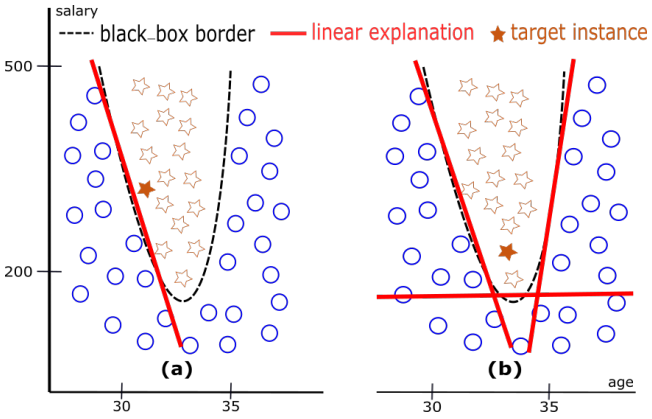


Figure 1: Two explanation scenarios for a classifier and a target instance (the filled star): (a) a suitable single linear explanation; (b) three contradictory linear explanations.

2 PRELIMINARIES

Problem Statement. Given a black-box classifier $f : X \to Y$ trained on a dataset $T \subset X$, and a target instance $x = (x_1, \cdots, x_d) \in X$, our goal is to construct an oracle that tells us whether a linear surrogate g learned on a locality $\Phi \subset X$ (defined below) is suitable to explain $f(x)$. By "suitable" we mean that two *contradictory* linear explanations g, and g' may not have the highest adherence in Φ – the adherence being the outcome agreement between f and g. In this formulation, Y is a finite set of classes, X is a multidimensional domain defined on numerical and categorical features, and Φ is a region of the space that (i) covers x, (ii) is traversed by f's decision boundary, and (iii) is maximal, otherwise stated, the surrogate g cannot attain the quality guarantee $m(g) \geq \tau$ in any locality $\Phi' \supset \Phi$ for some adherence metric m. Table 1 provides an overview of the notation used throughout the paper.

Requirement (i) guarantees that the target instance x is included in the surrogate's training set. Moreover, requirement (ii) ensures that this training set is balanced, that is, it contains both instances inside and outside the class $f(x)$. It follows that the minimal locality satisfying these two requirements should be centered on the decision boundary – more precisely on x's closest counterfactual –, and have the target instance x on the boundary. This is depicted by the inner dotted circle in Figure 2. Requirement (iii) implies that Φ could actually be larger if the surrogate g still attains a good adherence as depicted by the bigger dashed circle in Figure 2. In such a case, the explanation generalizes to larger regions of the data space.

Symbol	Definition	Symbol	Definition
$f(\cdot)$	Black-box classifier	$g(\cdot)$	Linear surrogate
X, x	Input domain, target instance	Y	Output domain
T, t	Input dataset, instance	F	Target's friend instances
E, e	Target's enemies, enemy	Φ, $v_x(\cdot)$	Locality, Locality function
Z, z	Artificial instances, instance	R	Feature-attribution ranking
$m(\cdot)$	Adherence metric	τ	Adherence threshold

Table 1: Notation used in the paper.

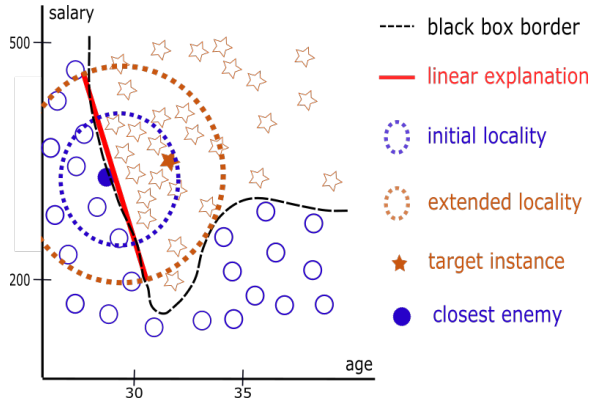


Figure 2: A linear explanation for a classifier and a target instance x. The inner circle (dotted in blue) is the minimal locality Φ that covers x and is traversed by the decision boundary. Locality can be extended (orange circle) and still provide an equally good linear approximation for the black-box. Friends F of x are represented by yellow stars, and enemies E by blue circles.

Linear Explanations and Counterfactuals. In order to explain the outcome $f(x)$ of a classifier f on a target instance x, methods such as LIME [17] or Local Surrogate [14] provide a signed feature-attribution ranking $R(g)$ that consists of ordered sets of features $R^+(g)$ and $R^-(g)$. The features in $R^+(g)$ contribute positively to predicting the class $f(x)$, whereas the features in $R^-(g)$ push towards predicting a different class. The ranking is based on the coefficients of a linear surrogate g that approximates f in a locality or neighborhood around x. This locality is defined by a function $v_x : X \to \{0, 1\}$ such that $v_x(x') = 1$ if x' is a neighbor of x and 0 otherwise. The set of all possible neighbors of x is then defined by $\Phi = v_x(X) = \{x' \in X \mid v_x(x') = 1\}$. The implementation of v_x depends on the explanation method.

In line with existing approaches to compute local explanations [9, 17, 18], we learn g on a sample of instances issued from a generative process that produces *artificial instances* $z \in Z \subset \Phi$ in x's neighborhood. If available, we also consider *real instances* that fall in the neighborhood, i.e., training instances $t \in T \cap \Phi$. We call a *counterfactual* or an *enemy* [13] any instance $e \in E \subset X$ such that $f(e) \neq f(x)$. Conversely, if $f(x') = f(x)$, we say that x' is a *friend* of x. Counterfactuals close to the target instance x can serve as informative contrastive explanations for $f(x)$.

We say two linear explanations g and g' for $f(x)$ are contradictory if they induce different attribute rankings, more formally, if $R(g) \neq R(g')$. We remark that the implementations of existing linear explanation modules may be subject to minor stability issues due

to the non-determinism (randomization) in the instance generation process. This may produce different linear explanations across multiple executions of the module on the same inputs. That said, such issues mostly affect the individual rankings within $R^+(g)$ and $R^-(g)$. In other words, instability episodes will rarely change the sign of the feature attribution. We therefore assume that signs are stable across multiple executions of the explanation module on the same input.

Adherence and Fidelity. The quality of a surrogate model g for a black-box classifier f is evaluated through the notions of adherence and fidelity. The *adherence* of a surrogate model g for a black-box model f is the degree of agreement between f's and g's outcomes. The *fidelity*, on the other hand, assesses the surrogate's ability to identify the features truly employed by the black-box model. When f is a true black box, users can only rely on the adherence to estimate the quality of explanations.

Existing Methods. LIME [17] is the most prominent approach to compute local linear explanations. For tabular data, LIME learns the surrogate g on a weighted neighborhood $Z \subset \Phi$ generated by perturbing the numerical attributes of x according to a μ-centered and σ-scaled normal distribution, where μ and σ are the attribute's mean and std. deviation in the training set. For categorical attributes, LIME uses the empirical distribution of the attribute values. The neighbors' weights are assigned according to an exponential kernel on the l_2-distance to x so that closer neighbors are given more importance when learning the surrogate. It has been shown [14] that we can learn more locally faithful explanations if we apply LIME on a neighborhood traversed by f's decision boundary. In that vibe, the Local Surrogate (LS) approach [14] centers the generative process not on the target instance x but on its closest enemy e – which by itself constitutes a complementary explanation for $f(x)$. LS then learns a linear surrogate on a neighborhood defined by a hyper-sphere centered at e, as depicted by the inner circle in Figure 2. On the downside, LS does not support categorical attributes. One of our contributions – the Growing Fields algorithm – proposes a solution to this limitation.

3 ADAPTED POST-HOC EXPLANATIONS

We now elaborate on APE, our approach to compute adapted post-hoc explanations on tabular data for a target instance x and a black-box classifier f. When the decision frontier of f admits a single local linear surrogate according to our problem statement in Section 2, APE returns a linear-based explanation complemented with a counterfactual explanation. Otherwise, APE recommends a different explanation paradigm such as a rule-based surrogate.

APE is detailed in Algorithm 1. In a first stage (line 1), APE invokes the *Growing Fields* algorithm to find the black-box decision boundary. This is achieved by identifying x's closest enemy – denoted by e. Then, APE generates a set of random instances Z uniformly distributed in a locality around e (line 3). This locality constitutes a *field*, which APE samples using the $\mathcal{F}$ generation process explained later. The size of the field depends on a radius parameter that is proportional to $dist(x, e)$, i.e., the distance between x and its closest enemy. More precisely, we set $r = 1/\delta \times dist(x, e)$, where δ is the farthest distance from x to a real instance in T, i.e., f's training set. By normalizing the radius, we (a) provide users with a data-agnostic notion of distance, and (b) reduce the risk of sampling instances beyond the limits of the attribute domains. By centering the generative process at e with radius r, APE makes sure that Z covers x and contains diverse subsets E and F of friends and enemies of x – in concordance with the requirements (i) and (ii) in the problem statement in Section 2. The $\mathcal{F}$ generation procedure as well as the Growing Fields algorithm are detailed in Section 3.1.

In the next step (line 4), APE characterizes the decision boundary of f. To this end, the algorithm invokes the APE oracle (Section 3.2), which runs efficient unimodality and linear separability tests [20, 23] on E and F to determine whether a linear surrogate is suitable or not. The oracle recommends a linear explanation if both sets E and F exhibit a unimodal distribution, that is, if there is only one cluster per class and we can separate those clusters with a single linear surrogate. In that case, APE returns a linear explanation and the closest enemy of x as a counterfactual explanation for $f(x)$. The linear explanation is learned via an extension of Local Surrogate [14], called LS_{APE}, applied on a superset of Z, consisting of real and artificial instances. Those instances constitute a field with a radius of at least r. We elaborate on those details in Section 3.3.

When the APE oracle deems linear explanations unsuitable, namely because the instances in E or F form multiple clusters, or because Z is not linearly separable, APE proposes a rule-based surrogate. Alternatives are Anchors [18] or shallow decision trees. In the first case, the user obtains a single rule of the form $R : p \Rightarrow f(x)$ where p is a set of conditions, and R has a precision of at least τ [18]. In the second case, the user gets a decision tree trained on a superset of Z. Since the decision boundary may consist of several disconnected instance clusters, APE completes its explanation with a counterfactual instance per cluster in E (see Section 3.4). That way users can have a comprehensive view of the different ways to change the black box's outcome $f(x)$.

In the next sections we elaborate on APE's building blocks, namely the $\mathcal{F}$ instance generation process, the Growing Fields algorithm, the APE oracle, and the procedures to compute the linear and rule-based surrogates.

Algorithm 1 APE

Require: a training dataset $T \subset X$, a target instance $x = (x_1, \ldots, x_d) \in X$, a black-box classifier $f : X \to Y$; number of samples n
Ensure: one or multiple counterfactual instances, a surrogate classifier g
1: $e \leftarrow$ GROWING FIELDS(T, x, f)
2: $r \leftarrow 1/\delta \times dist(x, e)$ *// δ is the largest distance in T*
3: $Z \sim \mathcal{F}(T, r, e)_{i \le n}$
4: **if** APE ORACLE(Z, x, f) **then**
5: **return** e, $LS_{APE}(Z, f, x, e)$ trained on e-centered field of radius $r' \ge r$
6: **else**
7: **return** $\{e_1, \ldots, e_k\} \subset Z$, RULE-BASED SURR.$(f)$
8: **end if**

3.1 Growing Fields

To compute the closest enemies to a target instance x given a classifier f (line 1 in Algorithm 1), APE resorts to an enhancement of the Growing Spheres (GS) algorithm [13] that we call Growing Fields (GF). GS searches for enemies of x by drawing instances uniformly within the volume of a l_2-sphere of radius r centered at x. The value of r is adjusted so that the resulting sphere traverses f's decision

boundary and encompasses enemies of x lying close to the border. GF proceeds likewise, but tackles some of the limitations of GS as explained next.

Attribute-dependant perturbations. By drawing instances uniformly in a l_2-sphere, GS assumes that all numerical attributes should be perturbed at the same rate. In reality, the attributes may have different amplitudes, variances, and distributions. Consequently, in GF the perturbation added to a numerical attribute x_i follows a uniform distribution that depends on both the radius r and the attribute's domain amplitude $A_i(T)$, and at the same time preserves the attribute's std. deviation in the input dataset T – denoted by $\sigma_i(T)$. This implies that the vicinity generated by GF around an instance x is not anymore a sphere, but rather a volume or, as we call it, a *field*. The actual shape of this field depends on the distance function. We highlight that taking into account the data distribution guarantees a data-aware exploration of the space, which results in a speed-up of up to 2 orders of magnitude w.r.t. GS.

Distance. Another limitation of GS is that all attributes have the same impact when computing the distance between two instances. That said, a salary "distance" of 30 EUR is insignificant compared to an age "distance" of 30 years. On those grounds, APE normalizes the contribution of attribute i using the mean μ_i and std. deviation σ_i in the training set, which boils down to the standardized Euclidean distance[2]:

$$dist(x, x') = \sqrt{\sum_{i=1}^{d} \left(\frac{(x_i - \mu_i) - (x'_i - \mu_i)}{\sigma_i}\right)^2} \tag{1}$$

Equation 1 assumes that the categorical attributes have been one-hot encoded.

Support for categorical features. The original GS algorithm does not support categorical attributes such as the sex or the marital status of a person. We can now handle those attributes by treating them as random continuous variables uniformly distributed in $[0, r]$. Consider a field with radius $r = 0.5$ and a target instance with the attribute $sex = F$. If by drawing a random value in $[0, 0.5]$ we obtain for example, a value of 0.2, we interpret it as throwing a biased coin that keeps the sex of the target instance with probability $1 - 0.2 = 0.8$. If the attribute defines more than two categories, e.g., {single, married, divorced, widowed} and we have to change the category, we use the re-adjusted empirical probabilities of the other categories in the input dataset T to randomly choose the new category.

Algorithm 2 details the resulting generation process, called $\mathcal{F}$ (which stands for field), used to draw random artificial instances with both numerical and categorical attributes. The result of integrating $\mathcal{F}$ into GS gives rise to the Growing Fields algorithm detailed in Algorithm 3. Growing Fields starts with an initial field of radius r_0 and reduces it until no enemies are found (lines 3-6). In a second stage, the field is *gradually expanded* until the decision boundary is crossed and close counterfactual instances can be reported (lines 7-10). The algorithm then returns x's closest counterfactual.

[2]This is a special case of the Mahalanobis distance when the covariance matrix is diagonal.

Algorithm 2 The $\mathcal{F}$ instance generation process

Require: a dataset $T \subset X$, a radius $r \in (0, 1]$, an instance $x = (x_1, \ldots, x_d) \in X$
Ensure: An artificial instance $z = (z_1, \ldots, z_d)$
1: **for** $i \in 1 \ldots d$ **do**
2: **if** x_i is numerical **then** // $A_i = max_i - min_i$
3: $a = min(0, r \times A_i(T) - \sigma_i(T))$
4: $b = a + \sigma_i(T)$
5: $z_i \leftarrow x_i + \rho_k$ with $\rho_k \sim \mathcal{U}(a, b)$
6: **else**
7: $z_i \leftarrow (x_i$ with prob. $1 - \rho_k)$ with $\rho_k \sim \mathcal{U}(0, r)$
8: **end if**
9: **end for**
10: **return** z

Algorithm 3 GROWING FIELDS (GF)

Require: a dataset $T \subset X$, a target instance $x = (x_1, \ldots, x_d) \in X$, a classifier $f : X \rightarrow Y$,
Hyper-parameters: $r_0 = 0.1$, $\theta = 1.8$, $n = 2000$ as defined by GS [13]
Ensure: Set Z of instances; resulting field radius r
1: $r \leftarrow r_0$
2: $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
3: **while** $\exists\, e \in Z \mid f(e) \neq f(x)$ **do**
4: $r \leftarrow r/2$
5: Update $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
6: **end while**
7: **while** $\nexists\, e \in Z \mid f(e) \neq f(x)$ **do**
8: $r \leftarrow \min(1, \theta \times r)$
9: Update $Z \sim \mathcal{F}(T, r, x)_{i \leq n}$
10: **end while**
11: **return** $\arg\min_e \{dist(x, e) \mid e \in Z\}$

3.2 APE Oracle

The core of the APE algorithm is the APE oracle described in Algorithm 4. This oracle determines whether the black-box decision boundary is separable by a single linear approximation. To achieve this, the oracle applies the Libfolding unimodality test [20] separately on the sets of friends F and enemies E of the target instance x in Z. If the test is passed, it means that F and E form each a single cluster in Z. This, however, does not suffice for linear separability; ergo the oracle carries out a quick linear separability test to determine whether these clusters of friends and enemies can be told apart with a linear approximation. The test is actually carried out on a balanced sample $Z_b \subseteq Z$. We enforce Z_b to contain an equal number of friends and enemies of x, because Z can be highly imbalanced towards the enemies of x for very small localities.

There are multiple methods to determine whether there exists a linear function that separates a two-class dataset. Such methods range from linear and quadratic programming to approaches based on computational geometry and neural networks [4]. Nevertheless, all these strategies are at least as expensive as running a linear regression on the input dataset. On those grounds, APE resorts to a simple test based on the Thornton's *separability index si* [23]. If $\Gamma_{X'}(x)$ returns the closest neighbor x' of x in a set $X' \subseteq X$, the

separability index measures the ratio of instances for which that closest neighbor is a friend of x. In our setting, this can be computed according to the following formula:

$$si(X') = \frac{\sum_{x' \in X'} \mathbb{1}_{f(\Gamma_{X'}(x'))=f(x')}}{|X'|}.$$

We remark that *si* lies between 0 and 1 and that higher values denote higher separability. Line 2 in Algorithm 4 checks if $si((T \cap \Phi) \cup Z_b) = 1$. That is, the test also considers real instances that fall within the field from which Z was drawn. If the test is passed, the decision boundary is considered linearly separable enough and the oracle returns true.

Algorithm 4 APE ORACLE

Require: instances $Z \subset X$, target instance $x = \{x_1, \ldots, x_d\} \in X$, classifier $f : X \rightarrow Y$
Ensure: Is f linearly separable in Z w.r.t. the class $f(x)$?
1: **if** $E \subset Z$ and $F \subset Z$ are unimodal **then**
2: **if** Z is linearly separable w.r.t. f **then**
3: **return** True
4: **end if**
5: **end if**
6: **return** False

3.3 Linear Explanations

If the APE oracle estimates that f's decision boundary is linearly separable around the target instance x, APE resorts to the routine LS_{APE} (described in Algorithm 5) to learn a linear surrogate g on Z and to provide an explanation for $f(x)$. We could center the generative process to learn g on the target instance x as in standard LIME, or around the decision boundary as in LS. We opt for the latter alternative since LS has been shown to identify more accurately the features that influence the black box locally [14].

We recall that Z is a sample drawn from a field centered on e with radius $r = 1/\delta \times dist(x, e)$ where e is x's closest enemy. We could therefore learn g from the instances used for the linear separability test, because these are exactly what LS needs for training. We highlight, however, that nothing prevents our linear surrogate from attaining a good adherence in larger scopes. In concordance with our maximality requirement (Section 2), LS_{APE} carries out a posteriori expansion of the training field before reporting the linear explanation to the user. While the adherence does not decrease, that is while $m(g) \geq \tau$, LS_{APE} extends the field radius and trains a new linear explanation (line 5-10 in Algorithm 5). The threshold τ is set to the adherence of g in the initial field. The radius is increased using the same expansion strategy of Growing Fields (lines 8-9 in Algorithm 3).

3.4 Rule-based Explanations

If the decision frontier in the vicinity of our target instance is too complex to be approximated with a single linear surrogate, users may apply clustering techniques on the neighborhood Z and provide different linear explanations for each of the instance clusters at the decision boundary. This would provide a complete picture of the black box behavior around the target. However, such an explanation is potentially difficult to grasp for users, because it might consist of potentially contradicting feature-attribution rankings. On those grounds, APE proposes by default a rule-based explanation when linear surrogates are considered unsuitable. Alternatives are anchors or shallow decision trees. Anchors [18] learns a single explanation rule of the form $p \Rightarrow f(x)$ such that p is a conjunction of conditions of maximal coverage and the rule has a precision of at least τ. The decision tree is learned on the set Z containing both friends F and enemies E of x in the field centered on e, the closest enemy of x. We remark, nevertheless, that our framework could be coupled with other explanation approaches [3, 9, 15]. This is an interesting avenue for future research.

Finally, APE complements the rule-based explanation with a set of counterfactual instances. These are the centroids of the clusters defined by an extended set of enemies $E^* \supseteq E$ (generated using the $\mathcal{F}$ generation process from Algorithm 2). This set can be obtained by increasing the field ratio r while the precision of the explanation is above τ. The clusters are computed using K-means [11] and the number of clusters k is determined using the Elbow method [22].

Algorithm 5 EXTENDED LOCAL SURROGATE (LS_{APE})

Require: instances $Z \subset X$ drawn from a field, a classifier $f : X \rightarrow Y$, target and counterfactual instance $x, e \in X$, an adherence metric m; Hyper-parameters: $\theta = 0.05$
Ensure: a linear surrogate classifier g
1: $r \leftarrow 1/\delta \times dist(x, e)$
2: Split Z into Z_{train}, Z_{test}
3: $g \leftarrow$ LINEAR REGRESSION$(Z_{train}, f(Z_{train}))$
4: $a \leftarrow \tau \leftarrow m(g)$ on Z_{test}
5: **while** $a \geq \tau \wedge r < 1$ **do**
6: $r \leftarrow \theta \times r$
7: $Z \sim \mathcal{F}(T, r, e)_{i \leq n}$
8: Split Z into Z_{train}, Z_{test}
9: $g \leftarrow$ LINEAR REGRESSION$(Z_{train}, f(Z_{train}))$
10: $a \leftarrow m(g)$ on Z_{test}
11: **end while**
12: **return** g

4 EXPERIMENTS

We conduct four rounds of experiments to evaluate APE:

- The first round of experiments (Section 4.1) assesses APE's oracle, specifically its ability to distinguish the cases where a linear explanation can yield a single accurate approximation for a given black-box classifier and target instance.
- In the second round, we compare APE's explanations to those of LIME [17] and LS [14] in terms of adherence (Section 4.3).
- In a third round (Section 4.4), we conduct an ablation study of the two components of the APE oracle through an evaluation of their impact on the adherence of APE.
- The last round in Section 4.5 compares the quality of APE's counterfactual explanations – computed with Growing Fields – with those output by Growing Spheres [13].

The source code of APE as well as the experimental datasets and additional results are available on Github[3].

[3] https://github.com/j2launay/APE

4.1 Experimental Setup

Datasets. Table 2 describes our experimental datasets. The list comprises 6 real and 6 synthetic datasets, the latter generated with scikit-learn[4]. Five of those synthetic datasets contain only numerical features. The real datasets were chosen to provide a mix of numerical and categorical features. All datasets define two target classes. We highlight though, that a multi-class classification problem can always be formulated in terms of a set of binary classification problems – one per class.

Name	Features		Instances
	Numerical	Categorical	
Adult	2	10	48842
Blob †	2	0	1000
Blobs †	12	0	5000
Blood	4	0	748
Cat Blobs †	4	4	5000
Cancer	10	20	569
Circles †	2	0	1000
Diabetes	8	0	768
M Blobs †	20	0	7500
Moons †	2	0	1000
Mortality	15	52	1614
Titanic	1	5	1046

Table 2: Experimental datasets († indicates synthetic datasets)

Black-box Classifiers. We evaluate APE on a handful of classifiers of different architectures – i.e., ensemble methods, piecewise-constant functions, smooth functions – implemented in scikit-learn with default values of hyperparameters unless stated otherwise: (i) Gradient Boosting (GB) with 20 tree estimators, (ii) Multi-layer Perceptron (MLP) with a logistic activation function, (iii) Random Forest (RF) with 20 tree estimators, (iv) Gaussian Naive Bayes (NB), (v) Support Vector Machine (SVM) with a balanced class weight, (vi) Decision Tree (DT), (vii) Logistic Regression (LR) and (viii) a Voting ensemble (VC) composed of LR, SVM, and NB classifiers. In addition to the class of an instance, the classifiers can provide class probabilities. The classifiers were trained on 70% of the data points and their accuracy tested on the remaining 30%. They exhibit accuracy scores between 0.65 and 0.99.

Explanation modules. APE and the competitors were tested on a random sample of 100 target instances drawn from the test instances of the experimental datasets. All the explanation modules had access to the training set used to learn the classifiers (argument T in Algorithm 1). We tested APE with Anchors and shallow decision trees (maximal depth of 3) as explanation solutions when linear explanations are considered unsuitable. We denote these variants by APE_a and APE_t. Anchors requires a precision goal τ for rules, that we set to 0.95. Nevertheless, the semantics of τ are purely indicative, because the algorithm will always report an explanation even if this goal is not attainable in the surrogate's training set. In line with LIME and LS, the training instances for learning the linear surrogate are labeled with the class probabilities of the target class $f(x)$ output by the black-box classifiers.

[4]http://scikit-learn.org

Metrics. We measure the adherence of our explanations via the accuracy score of the surrogate models on the region (e.g., field) where they were trained. We use 60% of the generated artificial instances (lines 5 and 7 in Algorithm 1) for training the surrogates and keep 40% for evaluating their accuracy. When we know the features actually used by the input classifier, we measure the explanation fidelity through the precision and the Kendall rank correlation coefficient on the sets of features reported by the explanations. The precision score gives the proportion of features in the explanation that are indeed used by the black-box classifier. The Kendall coefficient quantifies the agreement between the feature attribution rankings of the explanation and the actual contribution ranking in the black-box.

We evaluate the APE oracle by comparing the adherence and fidelity of the linear surrogates learned with LS_{APE} across the two outcomes of the oracle.

4.2 APE Oracle Evaluation

Adherence Evaluation. Table 3 presents the mean adherence (accuracy) of the linear surrogates computed for each black-box classifier across 100 test instances on our experimental datasets. The surrogates were computed using LS_{APE}. For each target instance, the APE oracle determines whether or not the decision boundary admits a single accurate linear approximation (Yes or No). The results show the pertinence of APE's linear suitability test. When the oracle predicts a linearly separable decision boundary, the surrogate's accuracy is on average 0.124 points higher than in the opposite case. Moreover, we observe that the proportion of linearly separable cases is mostly explained by the dataset. That said, the architecture of the black-box classifier can also have an impact on this proportion as suggested by the Adult dataset where 25% of the target instances of the Voting Ensemble (VC) are deemed unsuitable for a linear explanation, in contrast to the other datasets for which this proportion is higher. This happens in contrast to the Gradient Boosting (GB) classifiers where 65% of the target instances do not admit a linear explanation according to the oracle. We remark, however, that even when the oracle rejects linear suitability, the adherence of the linear surrogate can still be high, e.g., Cat Blobs dataset with GB black box. This can be explained by the fact that multimodal, e.g., clustered data, can still exhibit some level of linear separability if the individual clusters contain mostly instances of the same class. In such cases APE favors a rule-based explanation with multiple counterfactual instances in order to highlight the complexity of the decision boundary and illustrate the different ways to change the classifier's outcome. That is why APE tests first for unimodality and then for linear separability.

The interest of the APE oracle can be illustrated through this example drawn from the moons dataset – which contains 2 features. The Libfolding unimodality test on the set of closest enemies E around the target instance $x = [1.37, -0.65]$ detects a multimodal distribution, and the k-elbow method reports three enemy clusters whose centers are $z_1 = [1.23, 0.25], z_2 = [0.90, -0.07]$, and $z_3 = [0.76, 0.26]$. Applying LS_{APE} on those counterfactual instances as centers of the generative process reveals contradictory explanations, since the attribution of the first feature for z_1 is 0.079 whereas it is -0.003 for z_3.

Fidelity Evaluation. To compare the fidelity of the linear surrogates across the two possible outcomes of the oracle, we resort to a set

	Is a Linear Explanation Suitable?													
	GB			MLP			RF			VC			SVM	
	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$	Yes	No	$Prop_{no}$
Adult	**0.555**	0.486	0.65	**0.507**	0.397	0.60	**0.659**	0.483	0.47	**0.334**	0.304	0.25	**0.679**	0.643	0.35
Blob	**0.891**	0.782	0.57	**0.890**	0.760	0.49	**0.874**	0.730	0.56	**0.899**	0.748	0.46	**0.894**	0.744	0.43
Blobs	**0.855**	0.636	0.78	**0.723**	0.606	0.86	**0.783**	0.655	0.82	**0.745**	0.610	0.68	**0.717**	0.599	0.80
Blood	\	0.437	0.99	\	0.497	1.00	\	0.283	1.00	\	0.223	1.00	\	0.622	1.00
Cancer	**0.502**	0.381	0.20	**0.501**	0.499	0.12	0.510	\	0.00	**0.411**	0.382	0.21	0.499	\	0.02
Cat Blobs	**0.910**	0.898	0.70	**0.958**	0.900	0.86	0.874	**0.958**	0.50	**0.967**	0.936	0.72	**0.883**	0.794	0.48
Circles	**0.945**	0.723	0.09	0.958	\	0.00	**0.950**	0.708	0.04	0.948	\	0.00	0.949	\	0.00
Diabetes	**0.630**	0.399	0.92	**0.802**	0.585	0.96	\	0.453	0.98	**0.673**	0.258	0.96	**0.717**	0.518	0.88
M Blobs	\	0.833	0.97	\	0.967	1.00	**0.863**	0.845	0.82	\	0.947	0.99	**0.944**	0.942	0.71
Moons	**0.923**	0.708	0.55	**0.917**	0.802	0.59	**0.918**	0.727	0.42	**0.916**	0.881	0.85	**0.920**	0.750	0.50
Mortality	\	0.826	1.00	\	1.000	1.00	\	0.839	1.00	\	0.518	1.00	\	0.420	1.00
Titanic	**0.761**	0.667	0.06	0.919	\	0.00	0.973	**1.000**	0.04	**0.999**	0.997	0.16	0.715	\	0.00

Table 3: Average accuracy computed on 100 instances per black-box model and dataset of LS_{APE} for both the oracle's outcomes. Columns "Yes" and "No" are the average accuracy of LS_{APE} when the oracle indicates that a linear explanation is suitable or unsuitable. \ denotes a non-meaningful accuracy score, i.e., there were less than 3 instances in that case. Columns $Prop_{no}$ denote the ratio of cases when the oracle does not predict linear suitability. The colors blue, orange, and red indicate $Prop_{no} \leq 33\%$, $33\% > Prop_{no} \geq 66\%$, and $Prop_{no} \geq 66\%$ respectively. Each row reports the results for a particular dataset, such as Adult in the first row.

of "glass" black-box classifiers, i.e., white-box classifiers treated as black boxes. The classifiers are trained on half of the dataset features, which we chose randomly. We restrict our evaluation to datasets with at least 8 features. We apply LS_{APE} and use as explanation the ranking given by the top half features (by the absolute value of the attribution coefficient) of the linear surrogates. Figure 3 depicts the Kendall rank correlation coefficient for gradient boosting (GB), decision tree (DT), random forest (RF), and logistic regression (LR) classifiers. For LR, the ground truth is given by the feature coefficients of the logistic function. Similarly, we can extract the ground truth for DT by collecting the features encountered along the classification path of the instances. For the GB and RF classifiers, we construct feature rankings by means of the Gini importance score [2] provided by scikit-learn.

We observe that whenever the APE oracle predicts linear suitability, the rank correlation is on average very close to 1. This means that LS_{APE} fully recovers the actual importance ranking of the features within the complex model. When the oracle discourages linear explanations, LS_{APE} has indeed difficulties at finding the actual features used by the "glass" black-box classifier. These results confirm that the APE's linear suitability test is a good indicator of the expected quality of a linear surrogate, which translates into faithful explanations for black-box classifiers. Similar results are obtained when using the precision as fidelity metric.

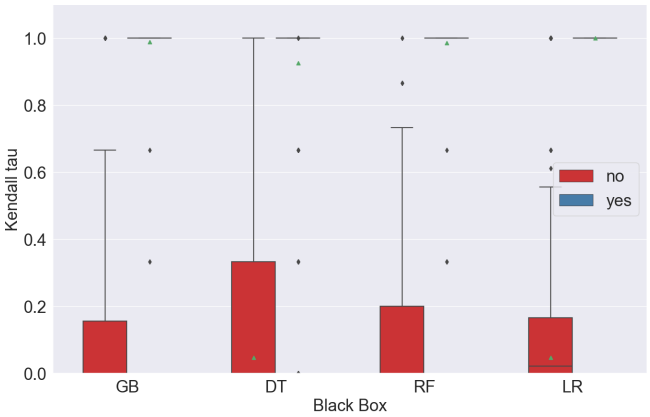


Figure 3: Average Kendall's rank correlation coefficient of the LS_{APE}'s explanations computed on 100 instances for 7 datasets and 4 "glass" black-box models across the oracle's outcomes.

4.3 Competitors Evaluation

We report the average accuracy on 100 target instances for linear surrogates learned with LIME, LS, and the APE's variants APE_a and APE_t. We exclude SHAP [15] from this evaluation because, even though the Kernel SHAP variant resorts to linear regression, it approximates the shapley values unlike linear surrogate such as LIME and LS that compute the gradient of the underlying model [7]. APE's variants return respectively an anchor or a shallow decision tree when the APE oracle does not predict linear suitability, otherwise they both invoke LS_{APE}. The results are shown in Figure 4. For LS, we omit the datasets with categorical attributes since these are not supported by this method.

The results show that regardless of the rule-based surrogate, APE achieves the best accuracy, and that the performance of its two variants depends on the black-box model. On average APE_a offers higher adherence, but also exhibits higher variability. All in all, this evaluation shows that judiciously choosing between linear and rule-based explanations in a per-instance basis brings a fidelity gain of 0.21 points on average when compared to always choosing LIME or LS. We also remark that when APE chooses to report a linear explanation, the decision frontier is indeed linearly separable: this is confirmed by the fact that both APE_a and APE_t outperform LS and LIME by a large margin even for black boxes with a relatively

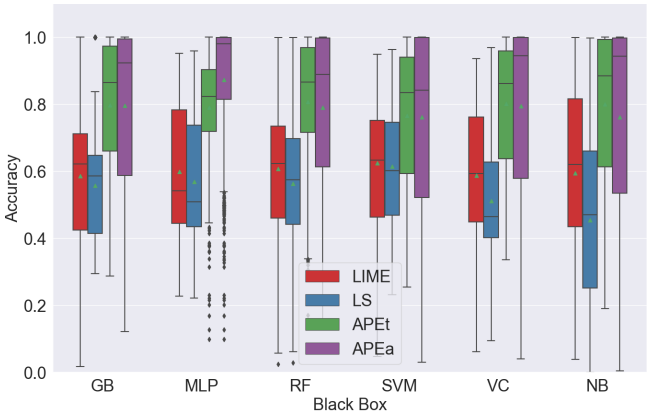


Figure 4: Average accuracy per black-box model on 100 instances of the experimental datasets for APE_a and APE_t.

high proportion of linearly suitable frontiers, e.g., SVM and RF (see Table 3).

4.4 Ablation Study

We now carry an ablation study to assess the contribution of the APE oracle's components, namely the Libfolding unimodality test [20] and the Thornton's linear separability test [23], on the adherence of APE. We report the average accuracy of APE_a and APE_t in Figures 5a and 5b, compared to the same variant of APE excluding the libfolding unimodality test ($APE \setminus \{Libfolding\}$) and the Thornton's separability test ($APE \setminus \{Thornton\}$). The results suggest that the unimodality and separability tests are complementary, and that simply testing for linear separability around the decision boundary is not enough to predict linear suitability as suggested by the accuracy of $APE \setminus \{Libfolding\}$.

4.5 Counterfactuals Evaluation

In line with the literature in counterfactual explanations [25], we assess the quality of APE's reported counterfactual instances for our 100 test instances, by measuring their resemblance to real instances. To this end we resort to the Mahalanobis distance computed against the entire set of enemies in the test instances. The Mahalanobis distance measures to which extent our counterfactual explanations are outliers w.r.t. the distribution of non-synthetic enemies. We report the quality of the counterfactual instances when computed using Growing Fields and Growing Spheres [13]. The distance values show that overall, APE finds more realistic counterfactual instances than GS, i.e., the instances lie in average 0.356 points closer to actual instances. This originates from the fact that, unlike GS, APE – more precisely Growing Fields – takes into account the variance and amplitude of the attributes when generating synthetic instances. This also incurs a speed-up of two orders of magnitude because taking into account the data distribution guarantees a data-aware expansion speed for the field radius during the quest for enemies (runtime results are provided on Github.)

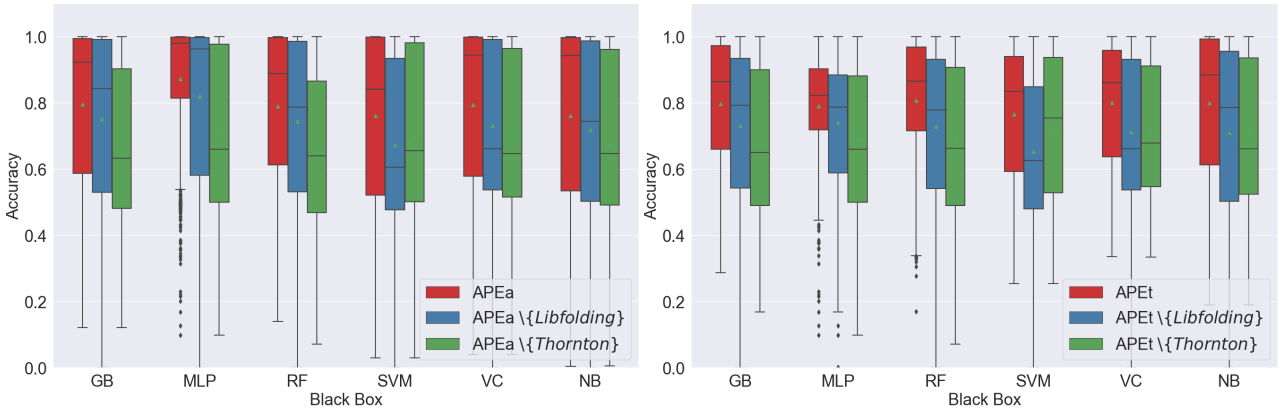
5 STATE OF THE ART

The fundamental question of what makes an explanation suitable for a particular use case lies at the junction of XAI and cognitive sciences. For this reason, this research question has not been addressed from a holistic perspective but rather from different, still complementary, angles.

On the one hand, the XAI community has put emphasis on the development of post-hoc explanation paradigms and methods [10], e.g., attribution scores, linear surrogates, rule-based surrogates, counterfactual explanations, sensitivity coefficients, etc. All these approaches aim to identify the features that play a role in the predictions of an AI model. Among those, feature attribution rankings based on linear surrogates such as LIME [17] or LS [14] enjoy notable popularity, because they can provide accurate per-instance explanations. Besides, practitioners from most disciplines are familiar with linear models. While SHAP [15] – more precisely its variant Kernel SHAP – may resort to linear regression to compute attribution scores, the obtained coefficients are not a linear approximation of the black box, but actually approximations of the Shapley values of the input features. These values are based on coalitional game theory and measure the average change in the model's expected prediction when conditioning on each feature. Shapley values are therefore akin to discrete gradients as computed by methods such as DeepLift [19] or Integrated Gradients [21]. All these explanations models are learned so that they optimize for user-agnostic criteria such as the adherence, which is usually an accurate proxy for fidelity [8]. Adherence is generally quantified by means of classical ML scores that depend on the black box's main task, e.g., classification, regression, etc. Other desiderata for explanations include low complexity [6] and stability [5, 26], however the bulk of the literature in classical XAI has pushed the state of the art towards novel approaches – or improvements of existing ones – that primarily optimize for fidelity *in the general case*. None of these works tackles the question of when a linear surrogate is objectively a reliable explanation. This is the primary driver of our work that focuses on adherence and fidelity for surrogate classifiers in a per-case basis.

At the other side of the spectrum, cognitive and social sciences study the subjective and human aspects of explaining AI models. In that spirit, the suitability of an explanation is characterized by its comprehensibility and plausibility [6]. Comprehensibility captures the extent to which a user grasps an explanation and can use it to accomplish well-defined tasks [1], e.g., determine the features used by the black-box system, predict the black box's answer, etc. That "understanding" is operationalized via objective measures on execution time or accuracy w.r.t. those tasks. On the other hand, the plausibility dimension models the cognitive preferences and background of the users. As pointed out by several studies [6, 12], users can reject an explanation if it contradicts common sense, for instance, if the explanation is too simplistic given that the underlying problem is deemed complex. The consensus seems to indicate that showing plausible and sound explanations increases trust in AI systems [12, 24], whereas the effects on comprehensibility and task efficiency are mixed.

While the XAI and cognitive science communities may appear somehow unreconciled, the relevance of the quality dimensions targeted by classical XAI methods has been justified by user studies. It



(a) Average accuracy of APE_a **(b) Average accuracy of APE_t**

Figure 5: Average accuracy per black box computed on 100 instances of the experimental datasets for APE_a in (a) and for APE_t in (b) when we remove the libfolding unimodality and linear separibility tests from the APE oracle.

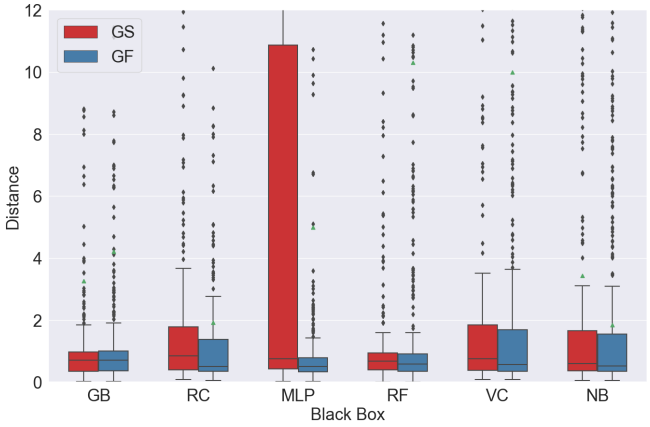


Figure 6: Average Mahalanobis distance between the counterfactual instances generated by Growing Spheres (GS) and Growing Fields (GF).

has been suggested [12] that in the context of recommender systems, low adherence harms trust in explanations. Evidence also suggests that multi-paradigm explanations can have a positive impact on comprehensibility [16, 27]. In particular, counterfactual explanations can be a complement to attribution-based or rule-based explanations. In this line of thought, our approach APE (a) informs the user of whether a use case is explainable with a faithful and unambiguous linear approximation, and (b) enriches the resulting explanation with counterfactual instances. When the decision boundary is unadapted to a linear surrogate, APE offers the possibility of computing a rule-based explanation.

6 DISCUSSION AND CONCLUSION

We have presented a method to decide *a priori* the pertinence of a local linear explanation for a given use case. Our decision is driven by standard user-agnostic desirata, namely the adherence and fidelity of the explanations. The experimental results suggest that it is possible to characterize the decision boundary of a black-box classifier around a target instance and select between linear and rule-based explanations. In that spirit, the answers of APE can provide valuable insights to the users of AI systems and linear surrogates. If APE discourages a linear explanation, then we can conclude that the classification boundary is probably complex and that a unique explanation based on feature attribution will be incomplete or inaccurate. Moreover, our use of counterfactual explanations provides users with a diverse and representative set of scenarios that can change the classifier's output. That being said, we emphasize that the most adapted explanation for a use case must take into account the human and cognitive aspects of explaining complex AI systems to end users. We focused on local linear explanations, because practitioners often resort to these models without questioning their pertinence.

Existing studies [16] suggest that multifaceted explanations, e.g., an anchor plus a counterfactual, can be more effective than single-paradigm explanations at illustrating the logic behind a classifier. Since this argument does not exclude the combination of attribution-based and rule-based explanations, this work does not discourage such a conjunction of paradigms. Instead it provides hints about the nature of the classifier's decision border. This could be useful in scenarios where the goal is to replace the black-box model, e.g., for reverse engineering or when a single and complete unambiguous explanation is required.

As future work we envision to port APE to other data types, e.g., text, and ML tasks, e.g., regression. Moreover, we would like to adapt our framework to other explanation paradigms such as rules, Shapley values, and different sorts of discrete gradients. This could be formulated in two ways: (i) by replacing rule-based surrogates with a different paradigm in APE, or (ii) by defining new oracles that can tell us when a particular explanation type is adapted to a classifier and target instance. Another interesting research avenue is the integration of the notions of coverage, complexity, and plausibility when deciding for the best explanations for a given use case.

Acknowledgments. This research was funded by the *Agence Nationale de la Recherche* (ANR) under grant agreement ANR-19-CE23-0019-01 and by the TAILOR Network (EU Horizon 2020 research and innovation program under grant agreement 952215).

REFERENCES

[1] Adrien Bibal, Bruno Dumas, and Benoît Frénay. 2019. User-Based Experiment Guidelines for Measuring Interpretability in Machine Learning. In *Workshop on Advances in Interpretable Machine Learning and AI*.

[2] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. 1984. *Classification and Regression Trees*. Wadsworth.

[3] Julien Delaunay, Luis Galárraga, and Christine Largouët. 2020. Improving Anchor-based Explanations. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*. 3269–3272. https://doi.org/10.1145/3340531.3417461

[4] David A. Elizondo. 2006. The linear separability problem: some testing methods. *IEEE Trans. Neural Networks* 17, 2 (2006), 330–344. https://doi.org/10.1109/TNN.2005.860871

[5] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. 2019. ILIME: Local and Global Interpretable Model-Agnostic Explainer of Black-Box Decision. In *European Conference on Advances in Databases and Information Systems*. Springer, 53–68.

[6] Johannes Fürnkranz, Tomáš Kliegr, and Heiko Paulheim. 2020. On cognitive preferences and the plausibility of rule-based models. *Machine Learning* 109, 4 (01 Apr 2020), 853–898. https://doi.org/10.1007/s10994-019-05856-5

[7] Damien Garreau and Ulrike von Luxburg. 2020. Explaining the Explainer: A First Theoretical Analysis of LIME. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 1287–1296. http://proceedings.mlr.press/v108/garreau20a.html

[8] Romaric Gaudel, Luis Galárraga, Julien Delaunay, Laurence Rozé, and Vaishnavi Bhargava. 2022. s-LIME: Reconciling Locality and Fidelity in Linear Explanations. In *International Symposium on Intelligent Data Analysis*. Springer, 102–114.

[9] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Dino Pedreschi, Franco Turini, and Fosca Giannotti. 2018. Local Rule-Based Explanations of Black Box Decision Systems. *CoRR* (2018). http://arxiv.org/abs/1805.10820

[10] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2019. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* 51, 5 (2019), 93:1–93:42. https://doi.org/10.1145/3236009

[11] Anil K. Jain and Richard C. Dubes. 1988. *Algorithms for Clustering Data*. Prentice-Hall.

[12] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways Explanations Impact End Users' Mental Models. In *IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10.

[13] Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. 2017. Inverse Classification for Comparison-based Interpretability in Machine Learning. *CoRR* abs/1712.08443 (2017). arXiv:1712.08443 http://arxiv.org/abs/1712.08443

[14] Thibault Laugel, Xavier Renard, Marie-Jeanne Lesot, Christophe Marsala, and Marcin Detyniecki. 2018. Defining Locality for Surrogates in Post-hoc Interpretablity. *CoRR* (2018). arXiv:1806.07498 http://arxiv.org/abs/1806.07498

[15] Scott M. Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems*. 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions

[16] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artif. Intell.* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[17] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2016), 1135–1144.

[18] Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*. 1527–1535. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16982

[19] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. In *Proceedings of the 34th International Conference on Machine Learning*. 3145–3153. http://proceedings.mlr.press/v70/shrikumar17a.html

[20] Alban Siffer, Pierre-Alain Fouque, Alexandre Termier, and Christine Largouët. 2018. Are your data gathered?. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, 2210–2218. https://doi.org/10.1145/3219819.3219994

[21] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70. 3319–3328. http://proceedings.mlr.press/v70/sundararajan17a.html

[22] Robert L. Thorndike. 1953. Who belongs in the family? *Psychometrika* 18, 4 (01 Dec 1953), 267–276. https://doi.org/10.1007/BF02289263

[23] Chris Thornton. 2002. *Truth from trash: How learning makes sense*. Mit Press.

[24] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404. https://www.sciencedirect.com/science/article/pii/S0004370220301533

[25] Sahil Verma, John P. Dickerson, and Keegan Hines. 2020. Counterfactual Explanations for Machine Learning: A Review. *CoRR* abs/2010.10596 (2020). https://arxiv.org/abs/2010.10596

[26] Giorgio Visani, Enrico Bagli, and Federico Chesani. 2020. OptiLIME: Optimized LIME Explanations for Diagnostic Computer Algorithms. In *Proceedings of the CIKM 2020 Workshops*, Vol. 2699. http://ceur-ws.org/Vol-2699/paper03.pdf

[27] Sandra Wachter, Brent D. Mittelstadt, and Chris Russell. 2017. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *CoRR* (2017). arXiv:1711.00399 http://arxiv.org/abs/1711.00399