An Axiomatic Approach to Linear Explanations in Data Classification

Jakub Sliwinski
ETH Zurich
Zurich, Switzerland
jakvbs@gmail.com

Martin Strobel
National Univ. of Singapore
Singapore
mstrobel@comp.nus.edu.sg

Yair Zick
National Univ. of Singapore
Singapore
dcsyaz@nus.edu.sg

ABSTRACT

In this work, we focus on *local explanations* for data analytics; in other words: given a datapoint $\vec{x}$, how important was the i-th feature in determining the outcome for $\vec{x}$? The literature has seen a recent emergence of various analytical answers to this question. We argue for a *linear influence measure* explanation: given a datapoint $\vec{x}$, assign a value $\phi_i(\vec{x})$ to every feature i, which roughly corresponds to feature i's importance in determining the outcome for $\vec{x}$. We present a family of measures called MIM (monotone influence measures), that are uniquely derived from a set of axioms: desirable properties that any reasonable influence measure should satisfy. Departing from prior work on influence measures, we assume no knowledge — or access — to the underlying classifier labeling the dataset. In other words, our influence measures are based on the dataset alone and do not make any queries to the classifier. We compare MIM to other linear explanation models in the literature and discuss their underlying assumptions, merits, and limitations.

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: User Interfaces—*Theory and methods*

Author Keywords

Influence Measures, Explainable ML, Algorithmic Transparency

INTRODUCTION

An individual is denied a bank loan; knowing that they are in good financial standing, they demand that the bank explain its decision. However, the bank uses an ML algorithm that automatically rejected the loan application. How should the bank explain its decision? This example is more than anecdotal; recent years have seen the widespread implementation of data-driven algorithms making decisions in increasingly high-stakes domains, such as healthcare, transportation, and public safety. Using novel ML techniques, algorithms are able to process massive amounts of data and make highly accurate predictions; however, their inherent complexity makes it increasingly difficult for humans to understand *why* certain decisions were made. By obfuscating the underlying decision-making processes, such algorithms potentially expose human stakeholders to risks. These risks could include incorrect decisions (e.g. Alice's application was wrongly rejected due to a system bug), information leaks (e.g. the algorithm was inadvertently given information about Alice that it should not have seen), or discrimination (e.g. the algorithm is biased against female applicants). Indeed, government bodies and regulatory authorities have recently begun calling for *algorithmic transparency*: providing human-interpretable explanations of the underlying reasoning behind large-scale decision-making algorithms. Our work represents a first formal axiomatic analysis of automatically generated explanations of black-box classifiers.

ExSS '18, March 11, Tokyo, Japan.

Our Proposal

We propose utilizing simple mathematical frameworks for an explanation via *influence measures*: these are functions that, given a dataset, assign a value to every feature; this value should roughly correspond to the feature's importance in affecting the classification outcome for individual data points. Slightly more formally, we are given a dataset $\mathscr{X}$ containing n dimensional vectors, whose data points are labeled by a binary classifier c, such that $c(\vec{y}) = \pm 1$ for all $\vec{y} \in \mathscr{X}$; now, given a *point of interest* $\vec{x} \in \mathscr{X}$, we wish to identify the features in $\vec{x}$ that are 'responsible' for it being labeled the way it was. This is done via a mapping ϕ whose input is the dataset $\mathscr{X}$, its labels (given by c), and the point of interest $\vec{x}$; its output is a vector $\phi(\vec{x}) \in \mathbb{R}^n$, where $\phi_i(\vec{x})$ corresponds to the influence of feature i on the label of $\vec{x}$. Intuitively, a large positive value of $\phi_i(\vec{x})$ should mean that feature i was highly important in determining the label of $\vec{x}$; a large negative value for $\phi_i(\vec{x})$ should mean that *despite* the value of i at $\vec{x}$, $\vec{x}$ was assigned this label. This approach carries several important benefits. First of all, it is completely generic, requiring no assumptions on the underlying classification model; secondly, linear explanation models are simple and straightforward, even for a layperson to understand (e.g. 'Alice was denied her loan because of the high importance the algorithm placed on her low monthly income, and despite her never having to file for bankruptcy'). The appeal of linear explanations has been recognized by the research community; recent years have seen a moderate boom of papers proposing linear explanations in data-driven domains (see Section 1.2). However, this poses a new problem for end users that wish to apply these methodologies: which linear explanation is the 'right' one to choose? In other words,

> ...which linear explanations are guaranteed to satisfy certain desirable properties?

We argue for an axiomatization of influence measures in classification domains. The axiomatic approach is common in the economics literature: first one reasons about simple, reasonable properties (axioms) which should be satisfied by any function (say, methods for dividing revenue amongst collaborators, or agreeing on an election winner given voters' preferences); next, one should prove that there exists a *unique* function satisfying these simple mathematical properties. The axiomatic approach allows one to rigorously reason about the types of influence measures one should use in a given setting: if the axioms set forth make sense in this setting, there is but one method of assigning influence in the given domain. It is, in some sense, an *explanation of an explanation method*, a provable guarantee that the method is sound; in fact, uniqueness implies that it is the only sound method one can reasonably use in a domain.

In a recent line of work, we identify specific properties that any reasonable influence measure should satisfy (Section 3); using these axioms, we mathematically derive a class of influence measures, dubbed *monotone influence measures* (MIM), which uniquely satisfy these axioms (Section 4). Unlike most existing influence measures in the literature, we assume neither knowledge of the underlying decision-making algorithm, nor of its behavior on points outside the dataset. Indeed, some methodologies (see Related Work in Section 1.2) are heavily reliant on having access to counterfactual information: what would the classifier have done if some features were changed? This is a rather strong assumption, as it assumes not only access to the classifier but also the potential ability to use it on nonsensical data points[1]. By making no such assumptions, we are able to provide a far more general methodology for measuring influence; indeed, many of the tools described in Section 1.2 will simply not be usable when queries to the classifier are not available, or when the underlying classification algorithm is not known. Finally, grounding the measure in the dataset ensures the distribution of data is accounted for, rather than explaining the classification in terms of arbitrarily chosen data points. The points can be very unlikely or impossible to occur in practice, and using them can demonstrate a behavior the algorithm will never exhibit in its actual domain. Despite their rather limiting conceptual framework, our influence measures do surprisingly well on a sparse image dataset. We show that the outputs of our influence measure are comparable to those of other measures, and provide interpretable results.

Related Work

Axiomatic approaches for influence measurement are common in economic domains. Of particular note are axiomatic approaches in cooperative game theory [9, 12, 3].

The first axiomatic characterization of an influence measure for datasets is provided in [4]; however, they interpret influence as a global measure (e.g., what is the overall importance of gender for decision making). Moreover, one of the axioms proposed in [4] turned out to be too strong, severely limiting the explanation power of the resulting measure. Indeed, as [6] show, the measure proposed by [4] outputs undesirable values (e.g. zero influence) in many real instances. [1] propose an empirical influence measure that relies on a potential-like approach. However, as we show, their methodology fails to satisfy reasonable properties even on simple datasets. Other approaches in the literature either rely on black-box access to the classifier [6, 8], or assume domain knowledge (e.g. that the classifier is a neural network whose layers are observable) [11]. Another notable axiomatic treatment of influence in data-driven domains appears in [6]; in this work, it is shown that a Shapley value based approach is the only way influence can be measured when one assumes counterfactual access to the black-box classifier. This result is confirmed in [7].

[1]For example if the dataset consists of medical records of men and women, the classifier might need to answer how it would handle pregnant men

THE FORMAL MODEL

A dataset $\mathscr{X} = \langle \vec{x}_1, \ldots, \vec{x}_m \rangle$ is given as a list of vectors in $\mathbb{R}^n$ (each dimension $i \in [n]$ is a feature), where every $\vec{x}_j \in \mathscr{X}$ has a unique label $c_j \in \{-1, 1\}$; given a vector $\vec{x} \in \mathscr{X}$, we often refer to the label of $\vec{x}$ as $c(\vec{x})$. For example, $\mathscr{X}$ can be a dataset of bank loan applications, with $\vec{x}$ describing the applicant profile (age, gender, credit score etc.), and $c(\vec{x})$ being a binary decision (accepted/rejected). An *influence measure* is simply a function ϕ whose input is a dataset $\mathscr{X}$, the labels of the vectors in $\mathscr{X}$ denoted by c, and a specific point $\vec{x} \in \mathscr{X}$; its output is a value $\phi_i(\vec{x}, \mathscr{X}, c) \in \mathbb{R}$; we often omit the inputs $\mathscr{X}$ and c when they are clear from context. The value $\phi_i(\vec{x})$ should roughly correspond to the importance of the i-th feature in determining the outcome $c(\vec{x})$ for $\vec{x}$.

AXIOMS FOR EMPIRICAL INFLUENCE MEASUREMENT

We are now ready to define our axioms; these are simple properties that we believe any reasonable influence measure should satisfy. We take a geometric interpretation of the dataset $\mathscr{X}$; thus, several of our axioms are phrased in terms of geometric operations on $\mathscr{X}$.

1. **Shift Invariance:** let $\mathscr{X} + \vec{b}$ be the dataset resulting from adding the vector $\vec{b} \in \mathbb{R}^n$ to every vector in $\mathscr{X}$ (not changing the labels). An influence measure ϕ is said to be *shift invariant* if for any vector $\vec{b} \in \mathbb{R}^n$, any $i \in [n]$ and any $\vec{x} \in \mathscr{X}$,

$$\phi_i(\vec{x}, \mathscr{X}) = \phi_i(\vec{x} + \vec{b}, \mathscr{X} + \vec{b}).$$

In other words, shifting the entire dataset by some vector $\vec{b}$ should not affect feature importance.

2. **Rotation and Reflection Faithfulness:** let A be a rotation (or reflection) matrix, i.e. an $n \times n$ matrix with $\det(A) \in \pm 1$; let $A\mathscr{X}$ be the dataset resulting from taking every point $\vec{x}$ in $\mathscr{X}$ and replacing it with $A\vec{x}$. An influence measure ϕ is said to be *faithful to rotation and reflection* if for any rotation matrix A, and any point $\vec{x} \in \mathscr{X}$, we have $A\phi(\vec{x}, \mathscr{X}) = \phi(A\vec{x}, A\mathscr{X})$. In other words, rotating or reflecting the entire dataset results in the influence vector rotating in the same manner.

3. **Continuity:** an influence measure ϕ is said to be *continuous* if it is a continuous function of $\mathscr{X}$.

4. **Flip Invariance:** let $-c$ be the labeling resulting from replacing every label $c(\vec{x})$ with $-c(\vec{x})$. An influence measure is *flip invariant* if for every point $\vec{x} \in \mathscr{X}$ and every $i \in [n]$ we have $\phi_i(\vec{x}, \mathscr{X}, c) = \phi_i(\vec{x}, \mathscr{X}, -c)$.

5. **Monotonicity:** a point $\vec{y} \in \mathbb{R}^n$ is said to *strengthen* the influence of feature i with respect to $\vec{x} \in \mathscr{X}$ if $c(\vec{x}) = c(\vec{y})$ and $y_i > x_i$; similarly, a point $\vec{y} \in \mathbb{R}^n$ is said to *weaken* the influence of i with respect to $\vec{x} \in \mathscr{X}$ if $y_i > x_i$ and $c(\vec{x}) \neq c(\vec{y})$. An influence measure ϕ is said to be *monotonic*, if for any data set $\mathscr{X}$, any feature i and any data point $\vec{x} \in \mathscr{X}$ we have $\phi_i(\vec{x}, \mathscr{X}) \leq \phi_i(\vec{x}, \mathscr{X} \cup \{\vec{y}\})$ whenever $\vec{y}$ strengthens i w.r.t. $\vec{x}$, and $\phi_i(\vec{x}, \mathscr{X}) \geq \phi_i(\vec{x}, \mathscr{X} \cup \{\vec{y}\})$ whenever $\vec{y}$ weakens i w.r.t. $\vec{x}$.

6. **Random Labels:** an influence measure ϕ is said to satisfy the *random labels* axiom, if for any dataset $\mathscr{X}$, if all labels are assigned i.i.d. uniformly at random (i.e. for all $\vec{x} \in \mathscr{X}$, $\Pr[c(\vec{x}) = 1] = \Pr[c(\vec{x}) = -1]$); we call this label distribution $\mathscr{U}$. Then, for all $\vec{x} \in \mathscr{X}$ and all i we have

$$\mathbb{E}_{c \sim \mathscr{U}}[\phi_i(\vec{x}, \mathscr{X}, c) \mid c(\vec{x}) = 1] =$$
$$\mathbb{E}_{c \sim \mathscr{U}}[\phi_i(\vec{x}, \mathscr{X}, c) \mid c(\vec{x}) = -1] = 0$$

In other words, when we fix the label of $\vec{x}$ and randomize all other labels, the expected influence of all features is 0.

Let us briefly discuss the latter two axioms. Monotonicity is key in defining what influence means: intuitively, if one is to argue that Alice's old age caused her loan rejection, then finding *older* persons whose loans were similarly rejected should strengthen this argument; however, finding older persons whose loans were not rejected should weaken the argument. The Random Labels axiom states that when labels are randomly generated, no feature should have any influence in expectation; any influence measure that fails this test is inherently biased towards assigning influence to some features, even when labels are completely unrelated to the data.

CHARACTERIZING MONOTONE INFLUENCE MEASURES

Influence measures satisfying the Axioms in Section 3 must follow a simple formula, described in Theorem 4.1; the full proof of Theorem 4.1 appears in a full version of this work.[2] Below, $\mathbb{1}(p)$ is a $\{1,-1\}$-valued indicator (i.e. 1 if p is true and -1 otherwise), and $\|\vec{x}\|_2$ is the Euclidean length of $\vec{x}$; note that we can admit other distances over $\mathbb{R}^n$, but stick with $\|\cdot\|_2$ for concreteness.

THEOREM 4.1. *Axioms 1 to 6 are satisfied iff ϕ is of the form*

$$\phi(\vec{x}, \mathscr{X}) = \sum_{\vec{y} \in \mathscr{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|_2) \mathbb{1}(c(\vec{x}) = c(\vec{y})) \quad (1)$$

where α is any non-negative-valued function.

We refer to measures satisfying Equation (1) as *monotone influence measures* (MIM). MIM uniquely satisfy a set of reasonable axioms; moreover, they maximize the total cosine similarity objective function. Intuitively, given a vector $\vec{x} \in \mathscr{X}$, an MIM vector $\phi(\vec{x}, \mathscr{X})$ will point in the direction that has the 'most' vectors in $\mathscr{X}$ sharing a label with $\vec{x}$. The value $\|\phi\|_2$ can be thought of as one's confidence in the direction: if $\|\phi\|_2$ is high, this means that one is fairly certain where other vectors sharing a label with $\vec{x}$ are (and, correspondingly, this means that there are at least some highly influential features identified by ϕ); a small value of $\|\phi\|_2$ implies low explanation strength.

[2]The main paper is currently under review.

EXISTING MEASURES

In this section, we provide an overview of some existing methodologies for measuring influence in data domains and compare them to MIM.

Parzen

The main idea behind the approach followed by [1] is to approximate the labeled dataset with a *potential function* and then use the derivative of this function to locally assign influence to features. Parzen satisfies Axioms 1 to 4. However, it is neither monotonic nor can it efficiently detect random labels.

LIME

The measure in [8] is based on the idea of finding a best local fit for the classifier in a region around $\vec{x}$. At its core, LIME fits a classifier by minimizing the mean-squared error, whereas MIM maximizes cosine similarity.

The Counterfactual Influence Measure

[4] initiated the axiomatic treatment of influence in data analysis; they propose a counterfactual aggregate influence measure for black-box data domains. Unlike other measures in this section, [4] do not measure local feature influence; rather, they measures the *overall influence* of a feature for a given dataset. The measure proposed by [4] does the following: when measuring the influence of the i-th feature; for every point $\vec{x} \in \mathscr{X}$, it counts the number of points in $\mathscr{X}$ who differ from $\vec{x}$ by only the i-th feature, and in their classification outcome. Given its rather restrictive notion of influence, this methodology only measures non-zero influence in very specific types of datasets: it assigns zero influence to all features in datasets that do not contain data points that differ from one another by only one feature; moreover, it only measures influence when a change in the state of a single feature changes the classification outcome.

Quantitative Input Influence

[6] propose a general framework for influence measure in datasets, generalizing counterfactual influence. Instead of measuring the effect of changing a single feature on point $\vec{x} \in \mathscr{X}$, they examine the *expected effect of changing a set of features*. The resulting measure, named QII (Quantitative Input Influence) is based on the Shapley value [9], a method of measuring the importance of individuals in collaborative environments. QII allows access to counterfactual information; moreover, it is computationally intensive in practice, and under its current implementation, will not scale to domains having more than a few dozen features.

Black-Box Access Vs. Data-Driven Approaches

Some measures above assume black-box access to the classifier (e.g. QII and LIME); others (e.g. Parzen and MIM) make no such assumption. Is it valid to assume black-box access to a classifier? This depends on the implementation domain one has in mind and the strength of explanations that one wishes to arrive at. On the one hand, having more access, measures such as QII and LIME can offer better explanations in a sparse data domain; however, they are essentially unusable when one does not have access to the underlying classifier. Data-driven approaches such as MIM, the counterfactual measure, and Parzen are more generic and can be applied on any

given dataset; however, they will naturally not be particularly informative in sparse regions of the dataset.

DISCUSSION AND FUTURE WORK

In this paper, we argue for the axiomatic treatment of linear influence measurement. We present a measure uniquely derived from a set of reasonable properties which also optimizes a natural objective function. Our characterization subsumes known influence measures proposed in the literature. In particular, MIM becomes the Banzhaf index in cooperative games and is also related to formal models of causality. Furthermore, MIM generalizes the measure proposed by [2] for measuring influence in a data-dependent cooperative game setting. Taking a broader perspective, axiomatic influence analysis in data domains is an important research direction: it allows us to rigorously discuss the *underlying desirable norms* we'd like to see in our explanations. Indeed, an alternative set of axioms is likely to result in other novel measures, that satisfy other desirable properties. Being able to mathematically justify one's choice of influence measures is important from a legal/ethical perspective as well: when explaining the behavior of classifiers in high-stakes domains, having *provably sound* measures offers mathematical backing to those using them.

While MIM offers an interesting perspective on influence measurement, it is but a first step. There are several interesting directions for future work; first, our analysis is currently limited to binary classification domains. It is possible to naturally extend our results to regression domains, e.g. by replacing the value $\mathbb{1}(c(\vec{x}) = c(\vec{y}))$ with $c(\vec{x}) - c(\vec{y})$; however, it is not entirely clear how one might define influence measures for multiclass domains. It is still possible to retain $\mathbb{1}(c(\vec{x}) = c(\vec{y}))$ as the measure of 'closeness' between classification outputs — i.e. all points that share $\vec{x}$'s output offer positive influence, and all those who do not offer negative influence — but we believe that this may result in a somewhat coarse influence analysis. This is especially true in cases where there is a large number of possible output labels. One possible solution for the multiclass case would be to define a distance metric over output labels; however, the choice of metric would greatly impact the outputs of MIM (or any other influence measure).

Another major issue with MIM (and several other measures) is that their explanations are limited to the influence of individual features; they do not capture joint effect, let alone more complex synergistic effects of features on outputs (the only exception to this is LIME, which, at least in theory, allows fitting non-linear classifiers in the local region of the point of interest). It would be a major theoretical challenge to axiomatize and design 'good' methods for measuring the effect of pairwise (or k-wise) interactions amongst features. This also allows one to have a natural tradeoff between the *accuracy* and *interpretability* of a given explanation. A linear explanation (e.g. LIME, QII, or this work) is easy to understand: each feature is assigned a number that corresponds to their positive or negative effect on the output of $\vec{x}$; a measure that captures k-wise interactions would be able to explain much more of the underlying feature interactions, but would naturally be less human interpretable. Indeed, a measure that captures all levels of feature interactions would be equivalent to a local approximation of the original classifier, which may not be feasible to achieve, nor easy to interpret. A better understanding of this behavior would be an important step in the design of influence measures. Finally, it is important to translate our numerical measure to an actual human-readable report. [6] propose using linear explanations as *transparency reports*; however, more advanced methods which assume access to the classifier source code propose mapping back to specific subroutines for explanations [5, 10]. Indeed, while the transition from data to numerical explanations is an important step, mapping these to actual human-interpretable explanations is an open problem.

REFERENCES

1. D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. 2010. How to explain individual classification decisions. *Journal of Machine Learning Research* 11 (2010), 1803–1831.
2. E. Balkanski, U. Syed, and S. Vassilvitskii. 2017. Statistical Cost Sharing. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*. 6222–6231.
3. J.F. Banzhaf. 1965. Weighted Voting Doesn't Work: a Mathematical Analysis. *Rutgers Law Review* 19 (1965), 317–343.
4. A. Datta, A. Datta, A. D. Procaccia, and Y. Zick. 2015. Influence in Classification via Cooperative Game Theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.
5. A. Datta, M. Fredrikson, G. Ko, P. Mardziel, and S. Sen. 2017. Proxy Non-Discrimination in Data-Driven Systems. *CoRR* abs/1707.08120 (2017).
6. A. Datta, S. Sen, and Y. Zick. 2016. Algorithmic Transparency via Quantitative Input Influence. In *Proceedings of the 37th IEEE Conference on Security and Privacy (Oakland)*.
7. S.M. Lundberg and S. Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*. 4768–4777.
8. M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD)*. 1513–1522.
9. L.S. Shapley. 1953. A Value for n-Person Games. In *Contributions to the Theory of Games, vol. 2*. Princeton University Press, 307–317.
10. S. Singh, M. T. Ribeiro, and C. Guestrin. 2016. Programs as Black-Box Explanations. *CoRR* abs/1611.07579 (2016).
11. M. Sundararajan, A. Taly, and Q. Yan. 2017. Axiomatic Attribution for Deep Networks. *arXiv preprint arXiv:1703.01365* (2017).
12. H.P. Young. 1985. Monotonic solutions of cooperative games. *International Journal of Game Theory* 14, 2 (1985), 65–72.