Axiomatic Characterization of Data-Driven Influence Measures for Classification

Jakub Sliwinski [1] **Martin Strobel** [2] **Yair Zick** [2]

Abstract

Given a labeled dataset and a specific datapoint $\vec{x}$, how did the i-th feature influence the classification for $\vec{x}$? This question can be answered using *influence measures* — functions that, given a datapoint $\vec{x}$, assign a numeric value $\phi_i(\vec{x})$ to every feature i, corresponding to how altering i's value would influence the outcome for $\vec{x}$. We propose *monotone influence measures (MIM)*, which are uniquely derived from a set of desirable properties, or axioms. The MIM family constitutes a provably sound methodology for measuring influence in classification; we show that while MIM is based on the dataset alone, it is demonstrably effective on data.

1. Introduction

Recent years have seen the widespread implementation of data-driven decision making algorithms in increasingly high-stakes domains, such as finance, healthcare, transportation and public safety. Using novel ML techniques, these 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. These algorithms are *black-box decision makers*: their inner workings are either hidden from human scrutiny by proprietary law, or (as is often the case) are so complicated that even their own designers will be hard-pressed to explain their behavior. Opaque data-driven classifiers run the risk of exposing human stakeholders to risks. These may include incorrect decisions, information leaks, or discrimination. 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. Several recent works propose influence measures for transparency: these methods quantify feature importance, offering insights to the roles they play in the underlying decision making process. However, these works, by and large, do not justify *why* their particular methodology is sound. Our work takes a fundamental approach to influence measurement in data-driven domains. We start from a set of desirable properties (or *axioms*), and derive a unique class of influence measures that satisfies these properties. In other words, we provide a

> *...formal axiomatic analysis of automatically generated explanations for black-box classifiers.*

1.1. Our Contribution

We investigate *influence measures*: functions that, given a dataset, assign a real value to every feature; it corresponds to how altering the feature's value is predicted to affect the outcome for individual datapoints. We identify specific properties (axioms) that any reasonable measure should satisfy, yielding the class of *monotone influence measures (MIM)*, uniquely satisfying these axioms (Section 3). 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 are rely on access to counterfactual information: how would the classifier label points that do not appear in the data? This assumption may be too strong, as it requires 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 methods described in Section 1.2 are unusable when queries to the classifier are not available, or when the underlying algorithm is unknown. In addition, MIM is faithful to the original data distribution: some measures are generated by sampling points uniformly at random at a certain region; however, such points can be very unlikely or impossible to occur in practice, and using them assumes a behavior that the classifier will never exhibit. As an illus-

[1]Information Technology and Electrical Engineering Department, ETH Zurich, Zurich, Switzerland [2]School of Computing, National University of Singapore, Singapore. Correspondence to: Martin Strobel <mstrobel@comp.nus.edu.sg>.

Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML), Stockholm, Sweden, 2018.

[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.

tration we show that despite our rather limiting conceptual framework, MIM does surprisingly well on a sparse image dataset (see Section 4). We compare the outputs of MIM to other measures, and provide interpretable results.[2]

1.2. Related Work

Several ongoing research efforts are informing the design of explainable AI systems (e.g. Kroll et al. (2017), Zeng et al. (2017)), as well as tools that explain the behavior of existing black-box systems (see Weller (2017) for an overview); our work focuses on the latter.

The work most closely related to ours is that of Datta et al. (2015). Datta et al. (2015) axiomatically characterize an influence measure for datasets; however, they interpret influence as a global measure (e.g., what is the overall importance of gender for decision making), whereas we measure feature importance for individual datapoints. Moreover, as Datta et al. (2016) show, the measure proposed by Datta et al. (2015) outputs undesirable values (e.g. zero influence) on real data; this is due to the fact that the measure requires the existence of counterfactual data: datapoints that differ by only a single feature. As we show in Section 4, MIM does not require such a dense dataset in order to register influence. Baehrens et al. (2010) propose a data-driven influence measure that relies on a potential-like approach; as we demonstrate in the supplementary material (Section C.1), their methodology fails to satisfy reasonable properties even on simple datasets.

Other approaches in the literature rely on black-box access to the classifier. Datta et al. (2016) use an axiomatically justified influence measure based on an economic fairness paradigm, called QII; briefly, QII perturbs feature values and observes the effect this has on the classification outcome (an equivalent methodology is also used by (Lundberg & Lee, 2017)). Another line of work using black-box access (Ribeiro et al., 2016b;a) uses queries to the classifier in a local region near the point of interest in order to measure influence. Adler et al. (2016) equate the influence of a given feature i with the ability to infer i's value from the rest of features, after it has been obscured; this idea is the basis for a framework for auditing black-box models. However, this approach assumes that one can make predictions on a dataset with some features removed. Koh & Liang (2017) have a different take on influence, identifying key *datapoints* — rather than features — that explain classifier behavior.

Some works study explanations for specific domains, such as neural networks (Ancona et al., 2017; Shrikumar et al., 2017; Sundararajan et al., 2017), or computer programs (Datta et al., 2017a); others apply explanations for generating more accurate predictions (Ross et al., 2017).

[2] We provide additional supplementary material in an anonymous online repository at https://www.dropbox.com/s/4m5u5oc0a85d9sq/MIM.FATML.2018.supp.pdf

2. The Model

A dataset $\mathcal{X} = \langle \vec{x}_1, \ldots, \vec{x}_m \rangle$ is given as a list of vectors in $\mathbb{R}^n$ (each dimension $i \in [n]$ is a feature), where every $\vec{x}_j \in \mathcal{X}$ has a unique label $c_j \in \{-1, 1\}$; given a vector $\vec{x} \in \mathcal{X}$, we refer to the label of $\vec{x}$ as $c(\vec{x})$. An *influence measure* is a function ϕ whose input is a dataset $\mathcal{X}$, vector labels denoted by c, and a specific *point of interest* $\vec{x} \in \mathcal{X}$; its output is a value $\phi_i(\vec{x}, \mathcal{X}, c) \in \mathbb{R}$; we often omit the inputs $\mathcal{X}$ and c when they are clear from context. The value $\phi_i(\vec{x})$ should correspond to how altering the i-th feature is predicted to affect the outcome $c(\vec{x})$ for $\vec{x}$ in the following way: if $\phi_i(\vec{x})$ is positive (negative), then for points similar to $\vec{x}$, increasing the value of the i-th feature increases (decreases) the likelihood of assigning the label $c(\vec{x})$, and the value $|\phi_i(\vec{x})|$ expresses the strength of that effect.

3. Characterizing Monotone Influence Measures

In this section we are going to define our axioms; these are simple properties that we believe any reasonable influence measure should satisfy and state our characterization result.

1. **Shift Invariance:** let $\mathcal{X} + \vec{b}$ be the dataset resulting from adding the vector $\vec{b} \in \mathbb{R}^n$ to every vector in $\mathcal{X}$ (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 $\vec{x} \in \mathcal{X}$,

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

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 $A\mathcal{X}$ be the dataset resulting from taking every point $\vec{x}$ in $\mathcal{X}$ and replacing it with $A\vec{x}$. An influence measure ϕ is said to be *rotation and reflection faithful* if for any rotation matrix A, and any point $\vec{x} \in \mathcal{X}$, we have

$$A\phi(\vec{x}, \mathcal{X}) = \phi(A\vec{x}, A\mathcal{X}).$$

In other words, the influence measure ϕ is invariant under rotation and reflection.

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

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 $\vec{x} \in \mathcal{X}$ and every $i \in [n]$ we have $\phi_i(\vec{x}, \mathcal{X}, c) = \phi_i(\vec{x}, \mathcal{X}, -c)$.

5. **Monotonicity:** a point $\vec{y} \in \mathbb{R}^n$ is said to *strengthen* the influence of feature i with respect to $\vec{x} \in \mathcal{X}$ 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 $\vec{x} \in \mathcal{X}$ 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 $\mathcal{X}$, any feature i and any data point $\vec{x} \in \mathcal{X}$ we have $\phi_i(\vec{x}, \mathcal{X}) \leq \phi_i(\vec{x}, \mathcal{X} \cup \{\vec{y}\})$ whenever $\vec{y}$ strengthens i w.r.t. $\vec{x}$, and $\phi_i(\vec{x}, \mathcal{X}) \geq \phi_i(\vec{x}, \mathcal{X} \cup \{\vec{y}\})$ whenever $\vec{y}$ weakens i w.r.t. $\vec{x}$.

6. **Non-Bias:** suppose that all labels for points in $\mathcal{X}$ are assigned i.i.d. uniformly at random (i.e. for all $\vec{y} \in \mathcal{X}$, $\Pr[c(\vec{y}) = 1] = \Pr[c(\vec{y}) = -1]$). We call this label distribution $\mathcal{U}$; an influence measure ϕ satisfies the *non-bias* axiom if for all $\vec{x} \in \mathcal{X}$ and all $i \in [n]$ we have

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

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

The first four axioms are rather fundamental: indeed, most influence measures in the literature trivially satisfy some variants of these properties. The last two axioms are more interesting (and naturally more debatable). While we strongly believe that there is no one "universally correct" set of axioms that all influence measures should satisfy, we believe that our proposed properties make intuitive sense in many application domains.

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 $\vec{x}$; we can admit other distances over $\mathbb{R}^n$, but stick with $\|\cdot\|_2$ for concreteness.

We are now ready to state our main result.

Theorem 3.1. *An influence measure ϕ satisfies axioms 1 to 6 iff it is of the form*

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

where α is any non-negative-valued function.

Due to space constraints we had to move the proof of Theorem 6 to the supplementary material (Section A).

We refer to measures satisfying Equation (1) as *monotone influence measures* (MIM). We note that MIM is a *family* of influence measures, parameterized by the choice of the function α. It may be natural to assume that α is a monotone decreasing function; that is, the further away the point $\vec{y}$ is from $\vec{x}$, the lower its effect on ϕ should be. However, this assumption does not follow from our analysis.

Further, we could show that MIM is an optimal solution to a natural optimization problem (see Section B in the supplementary material).

4. Experimental results

In what follows, we apply MIM, on a facial expression dataset. We compare it to results obtained fromtwo existing measures Parzen (Baehrens et al., 2010) and a adapted version of LIME (Ribeiro et al., 2016b) (see Section C for further explanation). The dataset used for this experiment is a part of the Facial Expression Recognition 2013 dataset (Goodfellow et al., 2013). The data consists of 12 156 48×48 pixel grayscale images of faces, evenly divided between happy and sad facial expressions. Each pixel is a feature; its brightness level is its parametric value.

The parameters were chosen as follows. For MIM we chose $\alpha(d) = \frac{1}{d^2}$ and for Parzen $\sigma = 4.7$. Finally, for the α parameter in Equation 6 of LIME, we choose $\alpha_\rho(d) = \sqrt{\exp(-d^2/\rho^2)}$ with $\rho = 3$ as a Kernel function.[3]

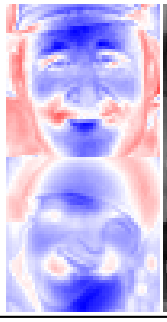
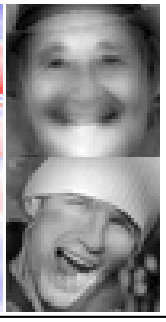
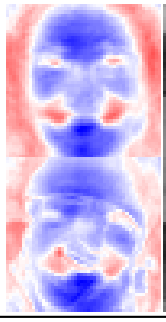
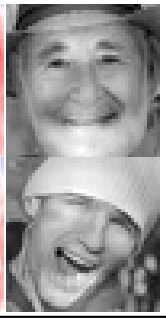
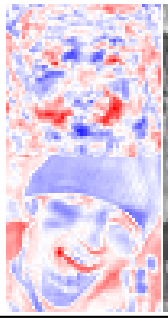
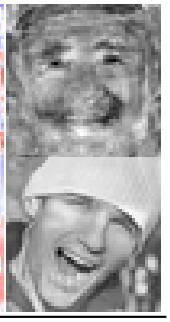
The first row of Table 1 shows an example picture of a happy face from the dataset, along with a visualization of the influence vectors as produced. In the images of influence vectors, the color blue (red) indicates positive (negative) influence; that is, for every pixel, the measures indicate that the brighter (darker) the pixel in the original image, the more 'happy' ('sad') the face. The third, fifth and seventh column show the point of interest shifted according to the respective influence vector, i.e. the pixels with positive influence were brightened, and darkened if their influence was negative.

According to the MIM influence vector, the factors that contribute to this face looking happy, are a bright mouth with darkened corners, bright eyebrows, bright tone of the face, and a darkened background. Shifting the picture along the influence vector seems to make the person in the picture smile wider, and open their mouth slightly. The Parzen vector differs from the MIM vector mainly in that it suggests dark eyes as indicative of the label and does not indicate the eyebrows as strongly. LIME, while generally agreeing with the other two, results in a more 'shattered' image. Seemingly it's better for a classifier to focus it's weights on a smaller set of features, while for MIM and Parzen you can see that neighbouring pixels actually have similar influence.

The second row shows another example picture and its corresponding influence vectors; however here, all measures fail to offer a meaningful explanation. This is likely to be since the face in the image is tilted, unlike the majority of images in the dataset. This is due to the fact that the dataset does not describe the locality of the image well enough; one can expect this to be the case for many images if the dataset is so small (12000) for such a complex feature space ($48 \times 48 = 2304$ features, with each potentially taking 256 different shades of gray). This exemplifies the dependency

[3] For a discussion on the effect of the parameterization as well as the analysis of a second dataset see the supplementary material Section E and F.

Table 1. Influence of two different points of interest (POI)

	MIM		Parzen		LIME	
POI	Influence	Shifted POI	Influence	Shifted POI	Influence	Shifted POI
						

of MIM on the dataset provided, and indicates it needs a relatively dense locality in order to perform reasonably well, if black-box access to the classifier or any domain knowledge cannot be assumed.[4]

5. Conclusions and Future Work

In this paper we present a novel characterization of data-driven influence measurement. We show that our measure is uniquely derived from a set of reasonable properties; what's more, it optimizes a natural objective function. Our work provides an additional perspective on LIME: our results suggest that cosine similarity should be favored over distance metrics as an optimization objective. 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 measures. Being able to mathematically justify one's choice of influence measures makes them more *accountable*: when explaining the behavior of classifiers in high-stakes domains, having *provably sound* measures offers mathematical backing to those using them. More importantly, an axiomatic approach allows one to justify the approach to non-academic stakeholders: while the formula for MIM (as well as other influence measures) might be rather obscure to those without the requisite background, the axioms it is derived from are rather intuitive and can be easily explained.

While MIM offers an interesting perspective on influence measurement, it is but a first step. 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, but we believe that this may result in a somewhat coarse 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 results.

Another major issue with current influence measures is that their explanations are limited to 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). Designing mathematically sound methods for measuring the effect of pairwise (or k-wise) interactions amongst features is a major challenge. This also allows one to have a natural tradeoff between the *accuracy* and *interpretability* of a given explanation. A linear explanation 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.

Finally, it is important to translate our numerical measure to an actual human-readable report. Datta et al. (2016) propose using linear explanations as *transparency reports*; more advanced methods use subroutines from the classifier's source code to explain its behavior (Datta et al., 2017b; Singh et al., 2016). Mapping numerical measures to actual human-interpretable explanations is an important open problem; we believe that analyses such as the one presented in this work form the fundamental basis for making black-box systems transparent, and ultimately more accountable.

[4] Further examples in the supplementary material support this hypothesis.

References

Adler, P., Falk, C., Friedler, S. A., Rybeck, G., Scheidegger, C., Smith, B., and Venkatasubramanian, S. Auditing black-box models for indirect influence. In *Proceedings of the 16th IEEE International Conference on Data Mining (ICDM)*, pp. 1–10, 2016.

Ancona, M., Ceolini, E., Öztireli, A. Cengiz, and Gross, M. H. A unified view of gradient-based attribution methods for deep neural networks. *CoRR*, abs/1711.06104, 2017.

Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., and Müller, K.-R. How to explain individual classification decisions. *Journal of Machine Learning Research*, 11:1803–1831, 2010.

Balkanski, E., Syed, U., and Vassilvitskii, S. Statistical cost sharing. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 6222–6231, 2017.

Banzhaf, J.F. Weighted voting doesn't work: a mathematical analysis. *Rutgers Law Review*, 19:317–343, 1965.

Datta, A., Datta, A., Procaccia, A. D., and Zick, Y. Influence in classification via cooperative game theory. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence. In *Proceedings of the 37th IEEE Conference on Security and Privacy (Oakland)*, 2016.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 1193–1210, 2017a.

Datta, A., Fredrikson, M., Ko, G., Mardziel, P., and Sen, S. Proxy non-discrimination in data-driven systems. *CoRR*, abs/1707.08120, 2017b.

Goodfellow, Ian, Erhan, Dumitru, Carrier, Pierre-Luc, Courville, Aaron, Mirza, Mehdi, Hamner, Ben, Cukierski, Will, Tang, Yuchuan, Thaler, David, Lee, Dong-Hyun, Zhou, Yingbo, Ramaiah, Chetan, Feng, Fangxiang, Li, Ruifan, Wang, Xiaojie, Athanasakis, Dimitris, Shawe-Taylor, John, Milakov, Maxim, Park, John, Ionescu, Radu, Popescu, Marius, Grozea, Cristian, Bergstra, James, Xie, Jingjing, Romaszko, Lukasz, Xu, Bing, Chuang, Zhang, and Bengio, Yoshua. Challenges in representation learning: A report on three machine learning contests, 2013.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pp. 1885–1894, 2017.

Kroll, J.A., Huey, J., Barocas, S., Felten, E., Reidenberg, J.R., Robinson, D.G., , and Yu, H. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

Lundberg, S. M. and Lee, S. A unified approach to interpreting model predictions. In *Proceedings of the 30th Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 4768–4777, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016a.

Ribeiro, M. T., Singh, S., and Guestrin, C. "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)*, pp. 1513–1522, 2016b.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2662–2670, 2017.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. *CoRR*, abs/1704.02685, 2017.

Singh, S., Ribeiro, M. T., and Guestrin, C. Programs as black-box explanations. *CoRR*, abs/1611.07579, 2016.

Sundararajan, M., Taly, A., and Yan, Q. Axiomatic attribution for deep networks. *arXiv preprint arXiv:1703.01365*, 2017.

Weller, A. Challenges for transparency. *CoRR*, abs/1708.01870, 2017.

Zeng, J., Ustun, B., and Rudin, C. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 180(3):689–722, 2017.