Axiomatic Characterization of Data-Driven Influence Measures for Classification

Jakub Sliwinski
ETH Zurich
Zurich, Switzerland
jsliwinski@ethz.ch

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

Yair Zick
National University of Singapore
Singapore
zick@comp.nus.edu.sg

Abstract

We study the following problem: given a labeled dataset and a specific datapoint $\vec{x}$, how did the i-th feature influence the classification for $\vec{x}$? We identify a family of *numerical 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 $\vec{x}$. This family, which we term *monotone influence measures (MIM)*, is uniquely derived from a set of desirable properties, or axioms. The MIM family constitutes a provably sound methodology for measuring feature influence in classification domains; the values generated by MIM are based on the dataset alone, and do not make any queries to the classifier. While this requirement naturally limits the scope of our framework, we demonstrate its effectiveness 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. Indeed, 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 are hard-pressed to explain their behavior. By obfuscating their reasoning, data-driven classifiers expose human stakeholders to risks. These may include incorrect decisions (e.g. a loan application that was wrongly rejected due to system error), information leaks (e.g. an algorithm inadvertently uses information it should not have used), or discrimination (e.g. biased decisions against certain ethnic or gender groups). 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 making algorithms more transparent by using numerical influence measures: methods for measuring the importance of every feature in a dataset. However, these works, by and large, do not justify *why* their particular methodology is sound. Our work takes an axiomatic approach to influence measurement in data-driven domains. Starting from a set of desiderata (or *axioms*), we uniquely derive a class of measures satisfying these axioms. Thus, our work provides a

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

1.1 Our Contribution

Numerical influence measures are functions that assign a value ϕ_i to every feature i; ϕ_i corresponds to the predicted effect of this feature on the outcome. We identify specific properties (axioms) that any reasonable influence measure should satisfy (Section 3), and derive a class of influence measures, dubbed *monotone influence measures* (MIM), uniquely satisfying these axioms (Section 4). Next, we show that MIM can be interpreted as the solution to a natural optimization problem, further grounding our methodology (Section 5). 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 Section 6 in the supplementary material) are heavily reliant on having access to counterfactual information: what would the classifier have done if some features were changed? This may be a strong assumption: 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 provide a general methodology for measuring influence; indeed, many of the methods described in Section 1.2 are unusable in the absence of classifier access, or when the underlying classification algorithm is not known. We show that despite our rather limiting conceptual framework, MIM does surprisingly well on a sparse image dataset, and provides an interesting analysis of a recidivism dataset (see results in the full version of this paper (Sliwinski, Strobel, and Zick 2018)). We compare the outputs of MIM to other measures, and provide interpretable results. Additional results in the full version of this work (Sliwinski, Strobel, and Zick 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.

relate MIM, new influence measures in a statistical cost sharing domain (Balkanski, Syed, and Vassilvitskii 2017), and classic game-theoretic measures (Banzhaf 1965).

1.2 Related Work

Algorithmic transparency has been debated and called for by government bodies (Goodman and Flaxman 2017; Smith, Patil, and C. 2016a; 2016b), the legal community (Suzor 2015; Charruault 2013), and the media (Hofman, Sharma, and Watts 2017; Angwin 2016; Mittelstadt et al. 2016; Winerip, Schwirtz, and Gebeloff 2016). The AI/ML research community is paying attention: algorithmic fairness, accountability and transparency is quickly gaining traction in the CS community, with new conferences (e.g. FAT* and AIES), numerous workshops and dozens of publications in mainstream AI/ML conferences. Several ongoing research efforts are informing the design of explainable AI systems (e.g. Kroll et al. (2017), Zeng, Ustun, and Rudin (2017)), as well as tools that explain the behavior of existing black-box systems (see Weller (2017) for an overview); we focus 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, Sen, and Zick (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 7, 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, 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, Sen, and Zick (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. Another line of work using black-box access (Ribeiro, Singh, and Guestrin 2016b; 2016a) 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 and 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, Greenside, and Kundaje 2017; Sundararajan, Taly, and Yan 2017), or computer programs (Datta et al. 2017b); others apply explanations for generating more accurate predictions (Ross, Hughes, and Doshi-Velez 2017).

2 The Model

A dataset $\mathcal{X} = \langle \vec{x}_1, \dots, \vec{x}_m \rangle$ 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 Axioms for Data-Driven Influence

We are now ready to define our axioms; these are simple properties that we believe any reasonable influence measure should satisfy.

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 $\phi_i(\vec{x}, \mathcal{X}) \leq \phi_i(\vec{x}, \mathcal{X} \cup \{\vec{y}\})$ if $\vec{y}$ strengthens i w.r.t. $\vec{x}$, and $\phi_i(\vec{x}, \mathcal{X}) \geq \phi_i(\vec{x}, \mathcal{X} \cup \{\vec{y}\})$ if $\vec{y}$ 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

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

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. 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.

3.1 The Case for Monotonicity

Monotonicity is a key defining property for characterizing our family of influence measures. Intuitively, it is a consistency requirement: if one is to argue that a person's old age caused their bank loan to be rejected, 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. We mention that monotonicity coupled with flip invariance implies the converse argument as well: adding younger persons whose loans were accepted should increase the influence of age, and adding younger persons whose loans were rejected would decrease it. Of course, in order for the monotonicity property to make any sense, feature states must satisfy some natural order: they should be numerical quantities (e.g. income, age, scores in a test, or shades of a color), states with a natural progression (e.g. education level, or disease severity), or binary states (e.g. gender). Monotonicity does not easily apply to features whose states cannot be naturally ordered (e.g. profession, ethnicity, species). That said, our characterization result holds whenever the dataset has at least one feature whose states can be naturally ordered.

3.2 Non-Bias: Measuring Influence vs. Measuring Noise

The Non-Bias axiom states that when labels are randomly generated, no feature should have any influence in expectation. We argue that this requirement is absolutely necessary: any influence measure that fails this test exhibits an inherent bias towards some features, even when labels are completely unrelated to the data. As we show in Section 6 in the supplementary material, some measures in the literature fail the non-bias test.

4 Characterizing Monotone Influence Measures

In what follows, we show that influence measures satisfying the axioms in Section 3 must follow a specific formula, described in Theorem 4.2. Below, $\mathbb{1}(p)$ is a $\{1, -1\}$-valued indicator (i.e. 1 if p is true and -1 otherwise), and $\|\vec{x}\|$ is the Euclidean length of $\vec{x}$; our analysis admits other distances over $\mathbb{R}^n$, but we stick with $\|\cdot\|$ for concreteness. We begin by showing a simple technical lemma.

Lemma 4.1. *If an influence measure ϕ satisfies both monotonicity and rotation faithfulness, then for any dataset $\mathcal{X}$, any datapoint $\vec{x} \in \mathcal{X}$, and any $\vec{y}$ where $\vec{y}$ and $\vec{x}$ differ in some feature, there exists some $a \in \mathbb{R}$ such that*

$$\phi(\vec{x}, \mathcal{X} \cup \{\vec{y}\}) - \phi(\vec{x}, \mathcal{X}) = a(\vec{y} - \vec{x}); \tag{1}$$

furthermore, $a \geq 0$ if $c(x) = c(y)$, and $a \leq 0$ otherwise.

Proof. Suppose for contradiction that there are $\mathcal{X}, \vec{x} \in \mathcal{X}$, and $\vec{y} \in \mathbb{R}^n$ with $c(\vec{x}) = c(\vec{y})$ such that

$$\forall a \in \mathbb{R}_+ : \underbrace{\phi(\vec{x}, \mathcal{X}) - \phi(\vec{x}, \mathcal{X} \cup \{\vec{y}\})}_{:=\vec{l}} \neq a(\vec{x} - \vec{y})$$

Let A be rotation matrix such that $(A\vec{l})_1 < 0$ and $A(\vec{x} - \vec{y})_1 > 0$; such a matrix exists since either the two vectors are linearly independent, or $\vec{l} = -b(\vec{x} - \vec{y})$ for some $b \in R_+$. Since ϕ satisfies Axiom 2 (Rotation), we get

$$\phi_1(\vec{A}x, A\mathcal{X}) - \phi_1(A\vec{x}, A\mathcal{X} \cup \{A\vec{y}\}) < 0,$$

which contradicts the first case of Axiom 5 (Monotonicity). The case where $c(\vec{x}) \neq c(\vec{y})$ can be derived symmetrically. □

We are now ready to prove our main result.

Theorem 4.2. *An influence measure ϕ satisfies axioms 1 to 6 iff it is of the form*

$$\phi(\vec{x}, \mathcal{X}, c) = \sum_{\vec{y} \in \mathcal{X} \setminus \vec{x}} (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|) \mathbb{1}(c(\vec{x}) = c(\vec{y})) \tag{2}$$

where α is any non-negative-valued function.

Proof. Suppose ϕ satisfies Axioms 1 to 6. We prove the statement by induction on $k = |\mathcal{X}|$. First assume that $k = 1$. When $k = 1$, $\mathcal{X} = \langle \vec{x} \rangle$. By shift invariance, $\phi(\vec{x}, \mathcal{X}) = \phi(\vec{0}, \langle \vec{0} \rangle)$. The vector $\vec{0}$ and $\langle \vec{0} \rangle$ are invariant under rotation; hence, by rotation faithfulness, $\phi(\vec{0}, \langle \vec{0} \rangle) = \vec{0}$, the only vector invariant under rotation. In other words, whenever the dataset has a single point, all features have zero influence.

When $k = 2$, we have $\mathcal{X} = \langle \vec{x}, \vec{y} \rangle$. If $\vec{x} = \vec{y}$ all features have zero influence (this is irrespective of whether $c(\vec{x}) = c(\vec{y})$ or $c(\vec{x}) \neq c(\vec{y})$). Further, note that any set of two points can be translated by shift and rotation to any other set of two points with the same labels and the same euclidean distance between them. Hence, by shift invariance, rotation faithfulness and Lemma 4.1,

$$\phi(\vec{x}) = \begin{cases} (\vec{y} - \vec{x}) \alpha_1(\|\vec{y} - \vec{x}\|) & \text{if } c(\vec{x}) = c(\vec{y}) \\ (\vec{y} - \vec{x}) \alpha_2(\|\vec{y} - \vec{x}\|) & \text{if } c(\vec{x}) \neq c(\vec{y}), \end{cases}$$

where α_1 (α_2) is some non-negative (non-positive) valued function that depends only on $\|\vec{y} - \vec{x}\|$. By random labels and flip faithfulness, $\alpha_1 = -\alpha_2$, thus $\phi(\vec{x}, \mathcal{X}) = (\vec{y} - \vec{x}) \alpha(\|\vec{y} - \vec{x}\|) \mathbb{1}(c(\vec{x}) = c(\vec{y}))$, where α depends only on $\|\vec{y} - \vec{x}\|$.

Suppose the hypothesis holds when $|\mathcal{X}| \leq k$. Consider any dataset $\mathcal{Y}$ of size $k + 1$. The cases where the dataset $\mathcal{Y}$

does not contain at least three different points are handled in a manner similar to when $k = 1, 2$. Suppose $\mathcal{Y}$ contains at least two distinct datapoints $\vec{y}, \vec{z} \neq \vec{x}$. We prove the hypothesis for the case where $\vec{y} - \vec{x}$ and $\vec{z} - \vec{x}$ are linearly independent; the case where they are linearly dependent follows from continuity (we can 'perturb' the points slightly to avoid linear dependency). By Lemma 4.1

$$\begin{aligned} \phi(\vec{x}, \mathcal{Y}) \in A =& \{\phi(\vec{x}, \mathcal{Y} \setminus \{\vec{y}\}) + a(\vec{y} - \vec{x}) : a \in \mathbb{R}\} \\ \text{and } \phi(\vec{x}, \mathcal{Y}) \in B =& \{\phi(\vec{x}, \mathcal{Y} \setminus \{\vec{z}\}) + a(\vec{z} - \vec{x}) : a \in \mathbb{R}\}. \end{aligned}$$

Further by the inductive hypothesis, $\phi(\vec{x}, \mathcal{Y} \setminus \{\vec{y}\})$ equals

$$\phi(\vec{x}, \mathcal{Y} \setminus \{\vec{y}, \vec{z}\}) + (\vec{z} - \vec{x})\alpha(\|\vec{z} - \vec{x}\|)\mathbb{1}(c(\vec{x}) = c(\vec{z}))$$

and $\phi(\vec{x}, Y \setminus \{\vec{z}\})$ equals

$$\phi(\vec{x}, Y \setminus \{\vec{y}, \vec{z}\}) + (\vec{y} - \vec{x})\alpha(\|\vec{y} - \vec{x}\|)\mathbb{1}(c(\vec{x}) = c(\vec{y})).$$

Since $\vec{y} - \vec{x}$ and $\vec{z} - \vec{x}$ are linearly independent we get,

$$\begin{aligned} \phi(\vec{x}, \mathcal{Y}) \in A \cap B = \{&\phi(\vec{x}, \mathcal{Y} \setminus \{\vec{y}, \vec{z}\}) \\ &+ (\vec{z} - \vec{x})\alpha(\|\vec{z} - \vec{x}\|)\mathbb{1}(c(\vec{x}) = c(\vec{z})) \\ &+ (\vec{y} - \vec{x})\alpha(\|\vec{y} - \vec{x}\|)\mathbb{1}(c(\vec{x}) = c(\vec{y}))\} \end{aligned}$$

concluding the inductive step. □

We refer to measures satisfying Equation (2) 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. In what follows, we propose a method of selecting the α parameter, by viewing MIM as a solution to an optimization problem, in a similar manner to Ribeiro, Singh, and Guestrin (2016b).

5 Choosing Optimal MIM Parameters

Is MIM a 'good' way of measuring influence? If the reader is convinced that the axioms proposed in Section 3 make sense, then our work here is done. In this section we make an additional case for MIM, showing that it is an optimal solution to a natural optimization problem. The results in this section serve an additional important purpose. Our characterization result (Theorem 4.2) identifies a *family of measures* (MIM), not a unique function, parameterized by the α function in Equation (2). Theorem 4.2 only requires that α is a function of $\|\vec{x} - \vec{y}\|$, but does not indicate what choice of α is appropriate. As we now show, MIM can be seen as a solution to an underlying optimization problem, the parameters of which may indicate the appropriate choice of α.

We are given a dataset $\mathcal{X}$ and a point of interest $\vec{x}$. Consider any potential influence vector ϕ; intuitively, ϕ should be a direction, such that moving $\vec{x}$ along ϕ will 'increase the chance' or 'positively contribute' to the label of $\vec{x}$ being $c(\vec{x})$. For any point $\vec{y} \in \mathcal{X}$ s.t. $c(\vec{y}) = c(\vec{x})$, it is desirable that ϕ points towards $\vec{y}$; if $c(\vec{y}) \neq c(\vec{x})$, ϕ should point away from $\vec{y}$.

Local points should be assigned more influence than further ones. Assume a function $\alpha_0 : \mathbb{R} \to \mathbb{R}$ whose input is $\|(\vec{y} - \vec{x})\|$; its output is a weightage representing the importance of $\vec{y}$; intuitively, α_0 should be monotone decreasing in its input, assigning higher values to points in a local neighborhood of $\vec{x}$ and lower importance to points further away. Hence, $\phi(\vec{x}, \mathcal{X})$ should maximize

$$\sum_{\vec{y} \in \mathcal{X}} \alpha_0(\|\vec{y} - \vec{x}\|) \cos(\vec{y} - \vec{x}, \phi)\mathbb{1}(c(\vec{x}) = c(\vec{y})) \qquad (3)$$

Equation (3) can be thought of as a weighted variant of the total cosine similarity optimization target.

Theorem 5.1. *MIM with the α parameter in* (2) *set to $\alpha(\|\vec{y} - \vec{x}\|) = \frac{\alpha_0(\|\vec{y}-\vec{x}\|)}{\|\vec{y}-\vec{x}\|}$, maximizes* (3).

Proof. For ease of exposition we assume that $\mathcal{X}$ has been shifted so that $\vec{x} = \vec{0}$; since MIM is shift invariant this is no loss of generality. Note that (3) treats a point $\vec{y}$ with $c(\vec{y}) \neq c(\vec{x})$ as it would the point $\vec{z} = -\vec{y}$ with $c(\vec{z}) = c(\vec{x})$. To simplify further, we assume that all points with a different label than $\vec{x}$ are swapped for their negatives with the same label as $\vec{x}$, resulting in a simplified formula

$$\phi(\vec{x}, \mathcal{X}) := \arg\max_{\phi \in \mathbb{R}^n} \sum_{\vec{y} \in \mathcal{X}} \cos(\vec{y}, \phi)\alpha_0(\|\vec{y}\|). \qquad (4)$$

Equation (4) pertains to the direction of ϕ. Intuitively, the length $\|\phi\|$ should correspond to how well the problem can be optimized. If the dataset is random, no direction should be particularly good, resulting in a short ϕ; that is, small influence values for every feature. In the case of the opposite extreme — all points with the same label as $\vec{x}$ are in a similar region, and all points with a different label in another — ϕ should be long, indicating high influence towards the points with the same label. Hence, the most natural way to specify the length of ϕ, again assuming $\vec{x} = \vec{0}$ for simplicity, is:

$$\|\phi(\vec{x}, \mathcal{X})\| := \max_{\phi \in \mathbb{R}^n} \sum_{\vec{y} \in \mathcal{X}} \cos(\vec{y}, \phi)\alpha_0(\|\vec{y}\|). \qquad (5)$$

To show that MIM maximizes (3), we require the following lemma (see the full version of this work (Sliwinski, Strobel, and Zick 2018) for the proof).

Lemma 5.2. *Let $f : \mathbb{R}^n \to \mathbb{R}$. Given a dataset $\mathcal{X}$,*

$$\sum_{\vec{y} \in \mathcal{X}} \cos(\vec{y}, \phi) f(\vec{y}) = \|\sum_{\vec{y} \in \mathcal{X}} \frac{\vec{y} \cdot f(\vec{y})}{\|\vec{y}\|}\| \cos\Big(\phi, \sum_{\vec{y} \in \mathcal{X}} \frac{\vec{y} \cdot f(\vec{y})}{\|\vec{y}\|}\Big).$$

Using Lemma 5.2 and substituting $\alpha_0(\vec{x}) = \|\vec{x}\|\alpha(\|\vec{x}\|)$, the right-hand side of Equation (5) becomes

$$\max_{\phi \in \mathbb{R}^n} \left(\|\sum_{\vec{y} \in \mathcal{X}} \vec{y}\alpha(\|\vec{y}\|)\| \cdot \cos\big(\phi, \sum_{\vec{y} \in \mathcal{X}} \vec{y}\alpha(\|\vec{y}\|)\big) \right).$$

Hence, $\|\phi(\vec{x}, \mathcal{X})\| = \|\sum_{\vec{y} \in \mathcal{X}} \vec{y}\alpha(\|\vec{y}\|)\|$. Combining that with Equation (4) we get $\phi(\vec{x}, \mathcal{X}) := \sum_{\vec{y} \in \mathcal{X} \setminus \{\vec{x}\}} \vec{y}\alpha(\|\vec{y}\|)$. Accounting for the simplifications we assumed, we get the general formula:

$$\phi(\vec{x}, \mathcal{X}) := \sum_{\vec{y} \in \mathcal{X}} (\vec{y} - \vec{x})\alpha(\|\vec{y} - \vec{x}\|)\mathbb{1}(c(\vec{x}) = c(\vec{y})).$$

□

Intuitively, given a point of interest $\vec{x} \in \mathcal{X}$, a monotone influence vector will point in the direction that has the 'most' points in $\mathcal{X}$ that share a label with $\vec{x}$. The value $\|\phi\|$ can be thought of as one's confidence in the direction: if $\|\phi\|$ 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); in the case that $\|\phi\|$ is small, the direction of ϕ is not a particularly strong indication of where other vectors of the same type can be found. In terms of choosing the right α parameter, Lemma 5.2 provides a few useful insights: if we select $\alpha(\|\vec{y}-\vec{x}\|) = 1$, then the resulting MIM measure maximizes the function $\sum_{\vec{y}\in\mathcal{X}} \cos(\vec{y}-\vec{x}, \phi)\|\vec{y}-\vec{x}\|$; in other words, we put *more* weight on vectors in $\mathcal{X}$ that are more distant from $\vec{x}$. Similarly, if we choose $\alpha(\|\vec{y}-\vec{x}\|) = \frac{1}{\|\vec{y}-\vec{x}\|}$ then we place equal importance on all points in the dataset, whereas if we set $\alpha(\|\vec{y}-\vec{x}\|) = \frac{1}{\|\vec{y}-\vec{x}\|^2}$, vectors that are farther away from the point of interest are weighted by $\frac{1}{\|\vec{y}-\vec{x}\|}$. This choice of α informs our implementation in Section 7.

6 Comparison to Existing Measures

In this section we provide an overview of some existing influence measures in data domains, and compare them to MIM. Measuring influence in data domains for algorithmic transparency is a relatively new approach, and has seen a veritable explosion of literature in recent years; we believe it is important to keep abreast of known methodologies and understand the domains where they are most appropriate.

6.1 Parzen

The main idea behind the approach followed by Baehrens et al. (2010) is to approximate the labeled dataset with a *potential function* and then use the derivative of this function to locally assign influence to features. Given a locality measure $\sigma \in \mathbb{R}_+$ and a kernel function

$$k_\sigma(\vec{x}) = \frac{1}{\sqrt{\pi\sigma^2}} \exp\left(\frac{-\sum_{i=1}^{n} x_i^2}{2\sigma^2}\right) \tag{6}$$

The Parzen measure $\phi_{\text{Parzen}_\sigma}(\vec{x}, \mathcal{X})$, is given by the derivative of the potential function below at $\vec{x}$.

$$\mathbb{P}(c(\vec{x}) = 1|\vec{x}) = \frac{\sum_{\vec{y}\in\mathcal{X} c(\vec{y})=1} k_\sigma(\vec{x}-\vec{y})}{\sum_{\vec{y}\in\mathcal{X}} k_\sigma(\vec{x}-\vec{y})}.$$

It is easy to check that $\phi_{\text{Parzen}_\sigma}$ satisfies Axioms 1 to 4. However, Parzen is neither monotonic, nor does it satisfy non-bias. To understand why Parzen fails monotonicity it helps to look at (6). In Figure 1, we have a single feature ranging from 0 to 2; we are measuring influence for the point $\vec{x}_0$ (marked with a green circle). When we add two more positive labels slightly to its right, monotonicity requires that the value of $\phi_{\text{Parzen}_\sigma}(\vec{x}_0, \mathcal{X})$ should not decrease; however, this addition 'flattens' the potential function, decreasing the influence of the feature. Non-bias is violated on any dataset with at least two distinct points. The underlying problem is the same: $\phi_{\text{Parzen}_\sigma}$ measures only change in labels, so data points with the same label lead to zero influence. This leads to $\phi_{\text{Parzen}_\sigma}$ assigning influence to random noise.

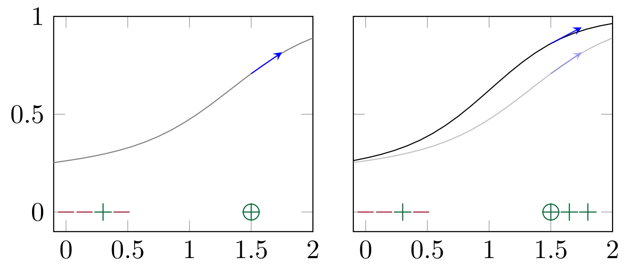


Figure 1: Parzen violates monotonicity; the point of interest $\vec{x}_0$ is marked with a green circle. Its influence is the slope of the blue arrow above it.

6.2 LIME

The approach followed by Ribeiro, Singh, and Guestrin (2016b) is based on the idea of using an interpretable classifier approximating the original in a region around $\vec{x}$; this simpler classifier then can be thought of as an explanation. This approach is termed Local Interpretable Model-agnostic Explanation (LIME).

Ribeiro, Singh, and Guestrin provide a concrete applicable framework, providing explanations in specific application domains. Some parts of this framework, however, lead to obvious violations of the axioms in Section 3. As an example, LIME maps datapoints to a binary explanation space, rather than considering them directly. This mapping aims to ensure that the result is human-interpretable; however, it clearly violates Axioms 1 to 4. On the other hand, one can draw a close connection between the theoretical framework underlying LIME, and the MIM formulation. In order to do so, it is useful to think of an influence measure as a linear classifier that approximates the data in a region close to the point of interest $\vec{x}$. We define the classifier based on an influence measure ϕ simply as $c_\phi(\vec{y}) = \mathbb{1}(\phi\vec{y} \geq \phi\vec{x})$

We rewrite the core optimization problem in Ribeiro, Singh, and Guestrin (2016b), when a linear classifier is used as an explanation:

$$\phi_{\text{LIME}}(\vec{x}) = \arg\min_{\phi\in\mathbb{R}^n} \sum_{\vec{y}\in\mathcal{X}} \alpha(||\vec{y}-\vec{x}||)(c(\vec{y}) - c_\phi(\vec{y}))^2 \tag{7}$$

where α is some non-negative function and we assume for simplicity $c(\vec{x}) = 1$.

Comparing this to Section 5 one can see that at its core, LIME minimizes the mean-squared error, whereas MIM maximizes cosine similarity (see Section 5). We note that other implementations of LIME (appearing in its source code), use cosine similarity rather than mean-squared error as the target; our results (namely Theorem 4.2) indicate that using cosine similarity offers certain theoretical guarantees over other approaches.

6.3 Counterfactual Influence

Datta et al. (2015) initiate the axiomatic analysis of influence in data domains. Unlike other measures in this section, their approach does not measure feature influence for a given point of interest; rather, it measures the *overall influence* of a feature for a given dataset. Following our notation, one can formulate the measure they propose as follows:

$$\eta_i(\mathcal{X}) = \frac{1}{|\mathcal{X}|} \sum_{\vec{x}\in\mathcal{X}} \sum_{y_i:(\vec{x}_{-i},y_i)\in\mathcal{X}} |c(\vec{x}) - c(\vec{x}_{-i}, y_i)| \quad (8)$$

In other words, the measure proposed by Datta et al. (2015) does the following: when measuring the influence of the i-th feature; for every point $\vec{x} \in \mathcal{X}$, it counts the number of points in $\mathcal{X}$ which differ from $\vec{x}$ by only the i-th feature, and in their classification outcome. This follows the idea of *counterfactual influence*: the importance of feature i is equivalent to its aggregate ability to change the outcome for points in $\mathcal{X}$, assuming that one is only allowed to change the i-th coordinate of $\vec{x}$. The axioms satisfied by (8) turn out to be too stringent: first, the counterfactual measure requires a dataset that contains datapoints differing by only one feature. Second, in many types of data, it is extremely unlikely that changing the state of a single feature will result in a change to the classification outcome (as noted by Datta, Sen, and Zick (2016)); indeed, on the dataset we study (Section 7), Equation (8) outputs zero influence for all features: no two points differ by only one feature.

6.4 Quantitative Input Influence

Datta, Sen, and Zick (2016) propose an influence measure generalizing counterfactual influence. Instead of measuring the effect of changing a single feature, they examine the *expected influence of changing a set of features*. More formally, given a set of features S, let $v(S;\vec{x}) = \mathbb{E}_{\vec{y}}[c(\vec{x}_{-S}, \vec{y}_S)]$ where $(\vec{x}_{-S}, \vec{y}_S)$ is the vector resulting from replacing the values of features in S with those of features in $\vec{y}$ ($\vec{y}$ is sampled from the empirical distribution of $\mathcal{X}$). In other words, $v(S;\vec{x})$ measures the expected effect of randomizing the values of features in S on the classification outcome of $\vec{x}$, with samples drawn according to the empirical distribution of S values in the dataset. Given this notion of 'value' for a set of features, Datta, Sen, and Zick (2016) use *the Shapley value (Shapley 1953)*, a well-known economic measure of influence from coalitional game theory. More formally, given a subset of features S, and a feature $i \notin S$, let $m_i(S;\vec{x}) = v(S \cup \{i\};\vec{x}) - v(S;\vec{x})$; that is, $m_i(S;\vec{x})$ is the *marginal effect* of randomizing i, given that we have randomized S. Let $\mathcal{N}_k^i = \{S \subseteq N \setminus \{i\} : |S| = k\}$; the influence measure defined by Datta, Sen, and Zick (2016) is then

$$QII_i(\vec{x}) = \frac{1}{n!}\sum_{k=0}^{n-1} k!(n-k-1)! \left(\sum_{S\in\mathcal{N}_k^i} m_i(S;\vec{x})\right) \quad (9)$$

$QII_i(\vec{x})$ is simply the Shapley value of feature i under the coalitional game defined by $v(S;\vec{x})$. By using the Shapley value, QII immediately guarantees several desirable properties 'for free' (as the Shapley value satisfies them); moreover, the Shapley value (and thus, QII) is the *only* way of measuring influence that can satisfy these properties. However, QII suffers from two major drawbacks. The first is that when computing $v(S;\vec{x})$, one assumes the ability to query the classifier on points that are not in the dataset (in particular, when computing $c(\vec{x}_{-S}, \vec{y}_S)$). Secondly, computing QII is computationally intensive, both when deriving the value of a set of features in $v(S;\vec{x})$ and when aggregating marginal effect in (9) (Chen et al. (2018) propose workarounds to these issues).

6.5 Black-Box Access Vs. Data-Driven Approaches

Influence measures in data domains seem to follow either one of two paradigms. One class of methods relies on *black-box access to the underlying classifier*; for example, QII (Datta, Sen, and Zick 2016) requires classifier queries in order to compute $v(S;\vec{x})$; LIME makes such queries to sample a local region of $\vec{x}$. *Data-driven methods* (e.g. Parzen, MIM) do not require black-box access.

Is it valid to assume black-box access to a classifier? This depends on the implementation domain one has in mind. On the one hand, having more access, measures such as QII and LIME 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 will work on any given dataset; however, they will naturally not be particularly informative in sparse regions of the dataset. That said, data-driven models subsume ones assuming black-box access: any data-driven method can be used after an initial black-box query phase: in this way, we add more points to the dataset $\mathcal{X}$ as a preprocessing step (for example, in order to obtain a dense region around the point of interest), and then run the data-driven method.

7 Experimental results

In what follows, we apply MIM, Parzen and a version of LIME on a facial expression dataset. We ran our experiments using a workstation with a quad core Intel i7 CPU, and 16GB of RAM. We were able to compute each influence vector in $4 - 5$ seconds. 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. A parametric Parzen influence measure with $\sigma = 4.7$ and a monotone influence measure with $\alpha(d) = \frac{1}{d^2}$ were run on some of the images. Further, we used a black-box data version of LIME as described in detail in the full version of this work (Sliwinski, Strobel, and Zick 2018). For the α parameter in Equation 7, we choose $\alpha_\rho(d) = \sqrt{\exp(-d^2/\rho^2)}$ with $\rho = 3$ as a Kernel function.[2]

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 by MIM, Parzen and LIME.

[2] For a discussion on the effect of the parametrization as well as the analysis of a second dataset see (Sliwinski, Strobel, and Zick 2018).

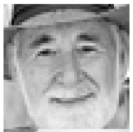
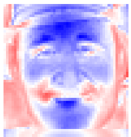
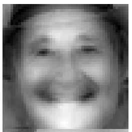
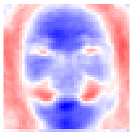
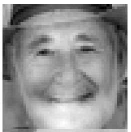
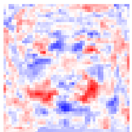
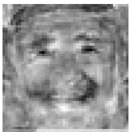
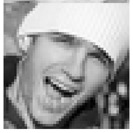
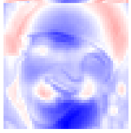
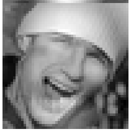
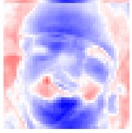
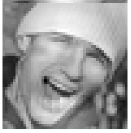
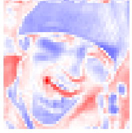
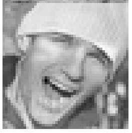
	MIM		Parzen		LIME	
POI	Influence	Shifted	Influence	Shifted	Influence	Shifted
						
						

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

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 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.[3]

8 Conclusions and Future Work

We present a novel characterization of data-driven influence measurement. Our measure is uniquely derived from a set of reasonable properties; what’s more, it optimizes a natural objective function.

Taking a broader perspective, axiomatic influence analysis in data domains is an important research direction: it allows us to rigorously discuss the *underlying norms* that govern our explanations. Different axioms result in alternative measures, and mathematically justifying 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 proofs in this paper might be rather obscure to those without the requisite background, the axioms we use 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.

Current numerical influence measures limit their explanations to individual features; they do not capture joint effect, let alone more complex feature interactions (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 provably sound methods for measuring the effect of pairwise (or k-wise) interactions amongst features is a major challenge. Non-linear explanations naturally trade-off *accuracy* and *interpretability*. A linear explanation is easy to understand, but lacks the explanatory power of a measure that captures k-wise interactions.

Finally, it is important to translate our numerical measure to an actual human-readable report. Datta, Sen, and Zick (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. 2017a; Singh, Ribeiro, and Guestrin 2016). Mapping numerical measures to actual human-interpretable explanations is an important open problem; we believe that analyses such as ours form the fundamental basis for making black-box systems transparent, and ultimately more accountable.

[3] Further examples in the full version (Sliwinski, Strobel, and Zick 2018) support this hypothesis.

acknowledgments The authors thank the AAAI 2019 reviewers and the FAT/ML 2018 participants for their useful suggestions. The authors are supported by an NRF fellowship #R-252-000-750-733.

References

Adler, P.; Falk, C.; Friedler, S. A.; Rybeck, G.; Scheidegger, C.; Smith, B.; and Venkatasubramanian, S. 2016. Auditing black-box models for indirect influence. In *Proc.of the 16th ICDM*, 1–10.

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

Angwin, J. 2016. Make algorithms accountable. *New York Times*.

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

Balkanski, E.; Syed, U.; and Vassilvitskii, S. 2017. Statistical cost sharing. In *Proc.of the 30th NIPS*, 6222–6231.

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

Charruault, M. 2013. N° de pourvoi: 12-17591. Cour de cassation.

Chen, J.; Song, L.; Wainwright, M. J.; and Jordan, M. I. 2018. L-Shapley and C-Shapley: Efficient model interpretation for structured data. *CoRR* abs/1808.02610.

Datta, A.; Datta, A.; Procaccia, A. D.; and Zick, Y. 2015. Influence in classification via cooperative game theory. In *Proc.of the 24th IJCAI*.

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

Datta, A.; Fredrikson, M.; Ko, G.; Mardziel, P.; and Sen, S. 2017b. Use privacy in data-driven systems: Theory and experiments with machine learnt programs. In *Proc.of the 2017 CCS*, 1193–1210.

Datta, A.; Sen, S.; and Zick, Y. 2016. Algorithmic transparency via quantitative input influence. In *Proc.of the 37th IEEE S&P (Oakland)*.

Goodfellow, I.; Erhan, D.; Carrier, P.-L.; Courville, A.; Mirza, M.; Hamner, B.; Cukierski, W.; Tang, Y.; Thaler, D.; Lee, D.-H.; Zhou, Y.; Ramaiah, C.; Feng, F.; Li, R.; Wang, X.; Athanasakis, D.; Shawe-Taylor, J.; Milakov, M.; Park, J.; Ionescu, R.; Popescu, M.; Grozea, C.; Bergstra, J.; Xie, J.; Romaszko, L.; Xu, B.; Chuang, Z.; and Bengio, Y. 2013. Challenges in representation learning: A report on three machine learning contests.

Goodman, B., and Flaxman, S. R. 2017. European union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine* 38(3):50–57.

Hofman, J.; Sharma, A.; and Watts, D. 2017. Prediction and explanation in social systems. *Science* 355(6324):486–488.

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proc.of the 34th ICML*, 1885–1894.

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

Mittelstadt, B. D.; Allo, P.; Taddeo, M.; Wachter, S.; and Floridi, L. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3(2):1–21.

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

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016b. "Why should I trust you?": Explaining the predictions of any classifier. In *Proc.of the 22nd KDD*, 1513–1522.

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proc.of the 26th IJCAI*, 2662–2670.

Shapley, L. 1953. A value for n-person games. In *Contributions to the Theory of Games, vol. 2*, Annals of Mathematics Studies, no. 28. Princeton University Press. 307–317.

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

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

Sliwinski, J.; Strobel, M.; and Zick, Y. 2018. A characterization of monotone influence measures for data classification. *CoRR* abs/1708.02153.

Smith, M.; Patil, D.; and C., M. 2016a. Big data: A report on algorithmic systems, opportunity, and civil rights. White House Report.

Smith, M.; Patil, D.; and C., M. 2016b. Big risks, big opportunities: the intersection of big data and civil rights. *White House Blog*.

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

Suzor, N. 2015. Google defamation case highlights complex jurisdiction problem. *The Conversation*.

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

Winerip, M.; Schwirtz, M.; and Gebeloff, R. 2016. For blacks facing parole in new york state, signs of a broken system. *New York Times*.

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