A Characterization of Monotone Influence Measures for Data Classification

Jakub Sliwinski[1], Martin Strobel[2], and Yair Zick[1]

[1] National University of Singapore
[2] Nanyang Technological University

Abstract. In this work we focus on the following question: how important was the i-th feature in determining the outcome for a given datapoint? We identify a family of *influence measures*; functions that, given a datapoint $\boldsymbol{x}$, assign a value $\phi_i(\boldsymbol{x})$ to every feature i, which roughly corresponds to that i's importance in determining the outcome for $\boldsymbol{x}$. This family is 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. While this requirement naturally limits the scope of explanations we provide, we show that it is effective on real datasets.

1 Introduction

Alice applied for a bank loan and was denied; knowing that she is in good financial standing, she demands that the bank explains its decision. However, the bank has recently implemented an ML algorithm that filters some applications, and has automatically rejected Alice's. How should the bank explain its decision? This example is more than anectdotal; 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 run the risk of exposing 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.

1.1 Our Contribution

In this work, we investigate *influence measures*: these are functions that, given a dataset, assign a value to every feature, roughly corresponding to its importance in affecting the

classification outcome for individual datapoints. We identify specific properties that any reasonable influence measure should satisfy (Section 3); next, 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 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[3]. 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 datapoints. 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 real datasets. In the appendix, we show that the outputs of our influence measure are comparable to those of other measures, and provide interpretable results.

1.2 Related Work

Algorithmic transparency has been called for by several government agencies [13, 16, 9, 18, 19]; in addition, recent court rulings have also required the opacity and neutrality of automatic decision systems [15, 5, 21, 6]. Last but not least, algorithmic transparency has been widely discussed in the media [17, 8, 2, 1, 22]. The AI and ML community has answered this call. Researchers are designing better explainable AI systems, as well as developing tools to explain the behavior of existing systems; our work is focused on the latter.

[10] axiomatically characterize an influence measure for datasets; however, they interpret influence as a global measure (e.g., what is the overall importance of gender in making decisions); on the other hand, we measure feature importance for individual data-points. Moreover, as [11] show, the measure proposed by [10] outputs undesirable values (e.g. zero influence) in many real instances. [3] propose an empirical influence measure that relies on a potential vector like approach. However, as we show in the appendix, their methodology fails to satisfy our axioms on simple datasets. Other approaches in the literature either rely on black-box access to the classifier [11, 14], or assume domain knowledge (e.g. that the classifier is a neural network whose layers are observable) [20].

[3] 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

2 Preliminaries

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

3 Axioms for Empirical Influence Measurement

We are now ready to define our axioms. We take a geometric interpretation of the dataset $\mathcal{X}$; thus, several of our axioms are phrased in terms of geometric operations on $\mathcal{X}$.

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

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

In other words, shifting the entire dataset by some vector $\boldsymbol{b}$ 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 $\boldsymbol{x}$ in $\mathcal{X}$ and replacing it with $A\boldsymbol{x}$. An influence measure ϕ is said to be *faithful to rotation and reflection* if for any rotation matrix A, and any point $\boldsymbol{x} \in \mathcal{X}$, we have

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

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 $\mathcal{X}$.

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

$$\phi_i(\boldsymbol{x}, \mathcal{X}, c) = \phi_i(\boldsymbol{x}, \mathcal{X}, -c).$$

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

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

$$\mathbb{E}[\phi_i(\boldsymbol{x}, \mathcal{X}, c)] = 0.$$

Let us briefly discuss the latter two axioms. Monotonicity is key in defining influence: 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 may assign influence to some features when labels are data independent.

4 Characterization result

In what follows, we show that influence measures satisfying the Axioms in Section 3 must follow a simple formula, described in Theorem 1. Below, $\mathbb{1}(p)$ is a $\{1, -1\}$-valued indicator (i.e. 1 if p is true and -1 otherwise), and $\|\boldsymbol{x}\|_2$ is the euclidean length of $\boldsymbol{x}$.

We begin by showing a simple technical lemma (proof omitted due to space constraints).

Lemma 1. *If an influence measure ϕ satisfies both monotonicity and rotation faithfulness, then for any dataset $\mathcal{X}$, any datapoint $\boldsymbol{x} \in \mathcal{X}$, and any $\boldsymbol{y}$ where $\boldsymbol{y}$ and $\boldsymbol{x}$ differ in some feature, there exists some $a \in \mathbb{R}$ such that*

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

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

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

$$\phi(\boldsymbol{x}, \mathcal{X}) = \sum_{\boldsymbol{y} \in \mathcal{X} \setminus \boldsymbol{x}} (\boldsymbol{y} - \boldsymbol{x}) \alpha(\|\boldsymbol{y} - \boldsymbol{x}\|_2) \mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{y})) \tag{2}$$

where α is any non-negative-valued function.

Proof. Suppose ϕ statisfies Axioms 1 to 6. We prove the statement by induction on $k = |\mathcal{X}|$; some technical points are omitted due to space constraints. When the dataset contains a single point (i.e. $k = 1$), the axioms imply that all features have an influence of 0.

When $k = 2$, we have $\mathcal{X} = \langle \boldsymbol{x}, \boldsymbol{y} \rangle$. If $\boldsymbol{x} = \boldsymbol{y}$ all features have zero influence. 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 1,

$$\phi(\boldsymbol{x}) = \begin{cases} (\boldsymbol{y} - \boldsymbol{x}) \alpha_1(\|\boldsymbol{y} - \boldsymbol{x}\|_2) & \text{if } c(\boldsymbol{x}) = c(\boldsymbol{y}) \\ (\boldsymbol{y} - \boldsymbol{x}) \alpha_2(\|\boldsymbol{y} - \boldsymbol{x}\|_2) & \text{if } c(\boldsymbol{x}) \neq c(\boldsymbol{y}), \end{cases}$$

where α_1 (α_2) is some non-negative (non-positive) valued function. By labels-expectation and flip faithfulness, $\alpha_1 = -\alpha_2$, and then $\phi(\boldsymbol{x}, \mathcal{X}) = (\boldsymbol{y} - \boldsymbol{x})\alpha(\|\boldsymbol{y} - \boldsymbol{x}\|_2)\mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{y}))$, where α depends only on $\|\boldsymbol{y} - \boldsymbol{x}\|_2$.

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 $\boldsymbol{y}, \boldsymbol{z} \neq \boldsymbol{x}$. We prove the hypothesis for the case where $\boldsymbol{y} - \boldsymbol{x}$ and $\boldsymbol{z} - \boldsymbol{x}$ 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 1 we have

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

Further by the inductive hypothesis we have:

$$\begin{aligned}\phi(\boldsymbol{x}, Y \setminus \{\boldsymbol{y}\}) =& \phi(\boldsymbol{x}, Y \setminus \{\boldsymbol{y}, \boldsymbol{z}\}) \\ &+ (\boldsymbol{z} - \boldsymbol{x})\alpha(\|\boldsymbol{z} - \boldsymbol{x}\|_2)\mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{z})) \\ \text{and } \phi(\boldsymbol{x}, Y \setminus \{\boldsymbol{z}\}) =& \phi(\boldsymbol{x}, Y \setminus \{\boldsymbol{y}, \boldsymbol{z}\}) \\ &+ (\boldsymbol{y} - \boldsymbol{x})\alpha(\|\boldsymbol{y} - \boldsymbol{x}\|_2)\mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{y})).\end{aligned}$$

Hence, since $\boldsymbol{y} - \boldsymbol{x}$ and $\boldsymbol{z} - \boldsymbol{x}$ are linearly independent we get,

$$\begin{aligned}\phi(\boldsymbol{x}, Y) \in A \cap B =& \{\phi(\boldsymbol{x}, Y \setminus \{\boldsymbol{y}, \boldsymbol{z}\}) \\ &+ (\boldsymbol{z} - \boldsymbol{x})\alpha(\|\boldsymbol{z} - \boldsymbol{x}\|_2)\mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{z})) \\ &+ (\boldsymbol{y} - \boldsymbol{x})\alpha(\|\boldsymbol{y} - \boldsymbol{x}\|_2)\mathbb{1}(c(\boldsymbol{x}) = c(\boldsymbol{y}))\}\end{aligned}$$

concluding the inductive step.

5 Conclusions and Future Work

In this paper we present a novel characterization of empirical influence measurement. Axiomatic analysis of influence in data domains is an important research direction, as it allows one to discuss *underlying desirable properties*. QII [11] is axiomatically characterized, but LIME(A.2) and PARZEN(A.1) are not. We believe that an axiomatic characterization of other measures would help the research community to better understand the benefits and drawbacks of each method.

Monotone influence measures have interesting connections to other domains. One can show that our measures generalize influence measures in empirical game-theoretic domains [4]; furthermore, our measures are related to mathematical formulations of responsibility and blame, described by [7]. These connections are encouraging, as they pave the way towards a general theory of causal influence across domains.

Acknowledgements

Sliwinski and Zick are supported by a Singapore MOE Grant #R-252-000-625-133; Zick is also supported by a Singapore NRF Fellowship Grant #R-252-000-643-281.

References

1. Angwin, J.: Make algorithms accountable. New York Times (Aug 2016), http://www.nytimes.com/2016/08/01/opinion/make-algorithms-accountable.html
2. Angwin, J., Larson, J., Mattu, S., Kirchner, L.: Machine bias theres software used across the country to predict future criminals. and its biased against blacks. ProPublica (May 2016), https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
3. Baehrens, D., Schroeter, T., Harmeling, S., Kawanabe, M., Hansen, K., Müller, K.R.: How to explain individual classification decisions. Journal of Machine Learning Research 11(Jun), 1803–1831 (2010)
4. Balkanski, E., Syed, U., Vassilvitskii, S.: Statistical cost sharing. CoRR abs/1703.03111 (2017)
5. Blue, T.H.J.: Duffy v. Google Inc (2015), http://www.austlii.edu.au/cgi-bin/sinodisp/au/cases/sa/SASC/2015/170.html, [2015] SASC 170
6. Charruault, M.: N° de pourvoi: 12-17591. Cour de cassation (Jun 2013), https://www.legifrance.gouv.fr/affichJuriJudi.do?oldAction=rechJuriJudi&idTexte=JURITEXT000027596148&fastReqId=468358130
7. Chockler, H., Halpern, J.Y.: Responsibility and blame: A structural-model approach. Journal of Artificial Intelligence Research 22, 93–115 (2004)
8. Citron, D.: (Un)fairness of risk scores in criminal sentencing. Forbes (Jul 2016), http://www.forbes.com/sites/daniellecitron/2016/07/13/unfairness-of-risk-scores-in-criminal-sentencing/#106235e54479
9. Custers, B.and Calders, T., Schermer, B., Zarsky, T.: Discrimination and Privacy in the Information Society: Data Mining and Profiling in Large Databases, vol. 3. Springer Science & Business Media (2012)
10. Datta, A., Datta, A., Procaccia, A.D., Zick, Y.: Influence in classification via cooperative game theory. In: Twenty-Fourth International Joint Conference on Artificial Intelligence (2015)
11. Datta, A., Sen, S., Zick, Y.: Algorithmic transparency via quantitative input influence. In: Proceedings of 37th IEEE Symposium on Security and Privacy (2016)
12. 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., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests (2013), http://arxiv.org/abs/1307.0414
13. Hollande, F.: Pour une république numérique (1) (Oct 2016), https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte, lOI n 2016-1321 NOR: ECFI1524250L
14. Ribeiro, M.T., Singh, S., Guestrin, C.: " Why should I trust you?": Explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 1513–1522. ACM (2016), http://www.kdd.org/kdd2016/subtopic/view/why-should-i-trust-you-explaining-the-predictions-of-any-classifier
15. Roggensack, C.J., Abrahamson, J.: Wisconsin v. Loomis (2016), https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690, case No.: 2015AP157 - CR

16. de Rosnay, M.D.: Algorithmic transparency and platform loyalty or fairness in the french digital republic bill (Apr 2016), http://blogs.lse.ac.uk/mediapolicyproject/2016/04/22/algorithmic-transparency-and-platform-loyalty-or-fairness-in-the-french-digital-republic-bill/, accessed: 2016-11-28
17. Smith, M.: A case is putting the use of data to predict defendants futures on trial. New York Times (Jun 2016), http://www.nytimes.com/2016/06/23/us/backlash-in-wisconsin-against-using-data-to-foretell-defendants-futures.html
18. Smith, M., Patil, D., C., M.: Big data: A report on algorithmic systems, opportunity, and civil rights. White House Report (May 2016), https://www.whitehouse.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf
19. Smith, M., Patil, D., C., M.: Big risks, big opportunities: the intersection of big data and civil rights. White House Blog (2016), https://www.whitehouse.gov/blog/2016/05/04/big-risks-big-opportunities-intersection-big-data-and-civil-rights
20. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. arXiv preprint arXiv:1703.01365 (2017)
21. Suzor, N.: Google defamation case highlights complex jurisdiction problem. The Conversation (Oct 2015), http://theconversation.com/australian-court-holds-google-is-responsible-for-linking-to-defamatory-websites-49883
22. Winerip, M., Schwirtz, M., Gebeloff, R.: For blacks facing parole in new york state, signs of a broken system. New York Times (Dec 2016), http://www.nytimes.com/2016/12/04/nyregion/new-york-prisons-inmates-parole-race.html?mtrref=www.nytimes.com&gwh=6B188E5340042B0E4B4848476BC73AE5&gwt=pay&_r=0

Appendix A Axiomatic analysis of existing measures

As mentioned above, several feature influence measures were proposed in prior work. Most of them, however, fundamentally rely on black-box access to the underlying classifier and cannot be immediately applied to our setting; for example, QII [11] cannot be easily applied without some heavy modifications. In this section we discuss two popular proposed methods: LIME [14] and PARZEN [3]. These methods can be applied to our setting without departing much from their original definition; moreover, they can be seen as typical examples of two fundamentally different ways of looking at this problem.

A.1 Parzen

The main idea behind the approach followed by [3] 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 σ and a kernel function

$$k_\sigma(\boldsymbol{x}) = \frac{1}{\sqrt{\pi\sigma^2}} \exp\left(\frac{\sum_{i=1}^n x_i^2}{2\sigma^2}\right),$$

we can derive the influence measure

Definition 1 (Parzen). *The* parametric parzen influence measure $\phi_{Parzen_\sigma}(\boldsymbol{x}, \mathcal{X})$ *is given by the derivative at* $\boldsymbol{x}$ *of the potential function*

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

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

A.2 LIME

The measure developed by [14] has been shown to work well in some instances. Unfortunately, at its' core is a discretization step which makes it unsuitable for an axiomatic analysis. Through the discretization alone it violates almost all axioms. On the other hand, based on the underlying idea of locally approximating the classification with a linear function, one can design an SVM-like measure more fit for theoretical analysis. However, the more adjustments one makes, the more the measure resambles a monotone influence measure, so the motivation for experimental comparison becomes unclear.

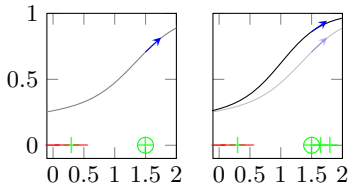
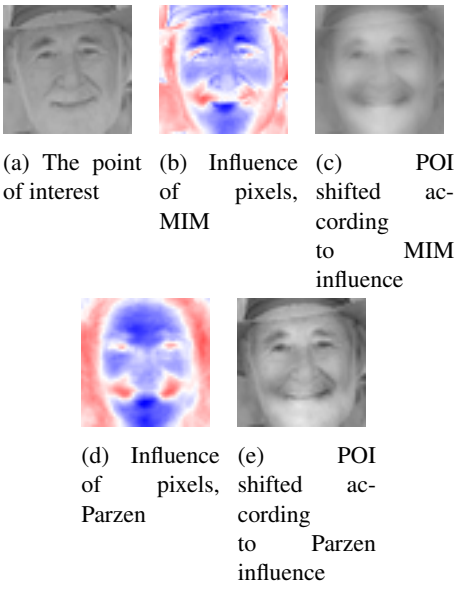


Fig. 1: Parzen violates monotonicity; the point of interest $\boldsymbol{x}_0$ is marked with a green circle. Its influence is the slope of the blue arrow above it.

Appendix B Experimental results

The dataset used to produce the experimental results is a part of the Facial Expression Recognition 2013 dataset described in [12]. The data consists of 12156 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.



(a) The point of interest

(b) Influence of pixels, MIM

(c) POI shifted according to MIM influence

(d) Influence of pixels, Parzen

(e) POI shifted according to Parzen influence

Fig. 2: Example point of interest

Figure 2 shows an example picture of a happy face from the dataset, along with a visualization of the influence vertors as produced by MIM and Parzen. In the images

(a) The point of interest

(b) Influence of pixels, MIM

(c) POI shifted according to MIM influence

(d) Influence of pixels, Parzen

(e) POI shifted according to Parzen influence

Fig. 3: Example point of interest

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. Subfigures 2c and 2e show the point of interest shifted according to the 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.

Figure 3 shows another example of a picture from the dataset and its MIM/Parzen influence vectors; however, both 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 how the influence measures are based only on the dataset provided and indicates it needs to describe the locality of the point of interest reasonably well, if black-box access to the classifier or any domain knowledge cannot be assumed.