Fidelity: the fundamental metric for evaluating post-hoc explanations

Thomas Merritt Smith [1]

Abstract

Deployment of accurate black-box models in high-stakes decision contexts has led to the advent of explanation methods which are designed to engender trust in these models, reassuring the user that the model is making decisions appropriately. However, to date, there is no single quantitative method to evaluate explanation methods. In this paper we define fidelity as the degree to which the explanation accounts for the model output for the explained sample, and argue that the fidelity of an explanation is the most important feature of any explanation. We analyse and compare the fidelity of explanations from the LIME and Layer wise-Relevance Propagation methods, demonstrating that fidelity provides compelling justification for selecting the best explanation method for a model.

1. Introduction

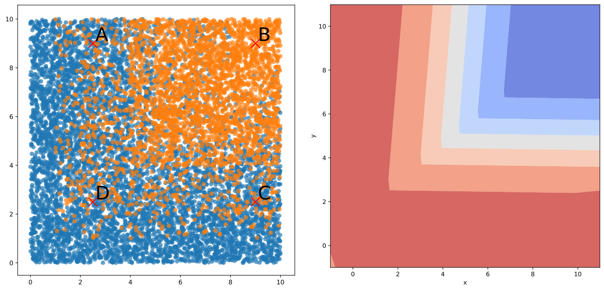
The continued success of machine learning models, and deep learning models in particular, in classification and regression problems across a wide range of domains has prompted reflection about the challenges of using such models without sufficient insight as to how the model output is related to the input features (Rudin, 2019). Such insight engenders trust that the model has learned significant and meaningful relationships between the input data and the response variable, and also enables the model user to 'debug' the modelling approach in cases where the model performance is not satisfactory. The ability to explain the output of a model is especially valuable in high stakes decision contexts, such as medical diagnosis, making parole decisions or fraud detection settings (Lipton, 2018; Collaris et al., 2018; Mittelstadt et al., 2019; Julia Angwin, 2016).

The appeal of explanation techniques is clear: high-accuracy black-box models can be deployed in high risk settings whilst still allowing model users to understand the decision-making process, and in doing so play an active part in that process. In practice, things are not as simple. Many methods to explain the predictions of black-box models have been proposed (Molnar, 2018), but for each method there are different mechanisms for generating explanations, different approaches to evaluation, and even different definitions of what an explanation is, if this is defined at all. These important technical differences can be difficult to communicate to end-users (Collaris et al., 2018). Often, an explanation method is justified by an appeal to qualitative factors, such as whether the explanation corresponds to a human understanding of the problem, or is easy to interpret by humans (Jacovi and Goldberg, 2020; Lipton, 2018; Ribeiro et al., 2016). While these are important features, quantitative measures are necessary to ensure that an explanation does in fact account for the behaviour of the model.

In this paper, we argue that the most important question to answer when attempting to explain the predictions of a model is whether the explanations from the selected approach account for the behaviour of the model. In order to do this, one must clearly state what an explanation is, and have a quantitative method with which to assess how accurate the explanations are in relation to the model. We define explanations as causal statements linking the inputs and outputs of a model, and propose *fidelity* to assess the accuracy of these statements. Since an ”explanation” which is not faithful to the model is essentially a false statement about the model behaviour, fidelity is therefore the fundamental measure that all explanation methods should be evaluated against, without which all the other benefits of explaining models are lost. To illustrate our argument, we calculate the fidelity of explanations from two methods applied to a neural network, trained on a simple, synthetic data set, and demonstrate that this calculation allows for a clear, principled evaluation of, and choice between, explanations.

2. Motivation

Consider the simplified synthetic two dimensional, binary classification data set as shown in Figure 1a, and in particular the four labelled points, as shown in Table 1. Intuitively, the explanation for point A being a member of class 0 is that the x_1 value is too low; and this intuition is the same for point C, but in reverse: if x_2 were larger, the probability of point C being a member of class 1 would increase. For point B, the classification would change if either x_1 or x_2 was less than 5, while for point D both x and y would have to increase to change the classification from class 0 to class 1. The *intuitive* explanation for these points is a counterfactual, causal one: the most important feature (or



(a) Synthetic data set

(b) Decision contours

Figure 1. The data set motivating this discussion is pictured in Figure 1a, showing a mixture of bernoulli binary classification problem. The decision contours for the 1-layer neural network M are shown in Figure 1b

Coordinate	Lime	LRP	δy
A: (2.5, 9)	(0.31, 0.07)	(-1.27, 3.24)	(0.60, 0.04)
B: (9, 9)	(0.21, 0.19)	(1.17, 1.11)	(0.63, 0.57)
C: (9, 2.5)	(0.06, 0.33)	(-3.95, 0.71)	(0.02, 0.68)
D: (2.5, 2.5)	(0.18, 0.14)	(-1.09, 0.71)	(0.01, 0.08)

Table 1. The coordinates A - D show the variation between LIME, LRP and δy. For points A, B or C, δy suggests that changing either x_1 or x_2 will significantly change the model output, while the output for point D will be unaffected by a change in either variable independently. This can also be concluded from the decision contours of M in 1b. With regards to a causal intuition that the most important features to M for a sample are those which would effect a larger change in the model output if they were changed (as presented in 2), we can see that LIME and δy are attuned to this intuition, while LRP is quite different.

features) is the one which, if it were different, would result in the greatest change to the model output. Suppose a user receives the LIME and Layerwise Relevance Propagation (LRP) explanations from Table 1; how should they choose between them? LIME seems to bear more relation to our intuition, but on the other hand perhaps the model is acting counterintuitively - certainly, one of the methods is less accurate than the other, since they do not agree on most of the cases. In a high stakes setting, selection between these results could be the difference between deploying a biased model on false evidence, or acting to ensure it is not deployed before further scrutiny. Crucially, we must separate our intuition about the more plausible explanation that x_1 is more important to the decision when $x_2 > 5$ and vice versa, from our *causal reasoning* underpinning our explanations - when $x_2 > 5$, *if* x_1 *were different, then the model output would change*. If this causal intuition can be quantified, then we can compare our explanation methods in a meaningful way.

3. Related Work

3.1. Evaluating explanations using fidelity

Jacovi and Goldberg (2020) summarise the literature on faithfulness (in a natural language processing context) as resting on one of three assumptions: that the explanation model is faithful if its outputs closely match the outputs of the true model (Model assumption); that a faithful explanation method is one which has similar outputs for similar inputs (referred to as the ”Prediction” assumption, this is similar to robustness, proposed by Alvarez-Melis (2018)); and finally the ”linearity assumption” which assumes that “certain parts of the input are more important to the model than others”, and that the inputs are independent with respect to their importance in the model. While these assumptions all demonstrate valuable features of explanations, we argue that both the ”Model” and the ”Prediction” assumptions are insufficient for evaluating explanation fidelity.

Ribeiro et al (2016) evaluate the fidelity of LIME using the ”Model” assumption by applying it to interpretable models, demonstrating that LIME is able to recover the most relevant features (as determined by model coefficients) in the test models in 90% of the predictions. However, this only demonstrates that LIME can recover the most relevant features of a linear model and a piece-wise linear model, rather than the most relevant features to any model. More generally, the assumption that if two models are shown to have similar predictive performance, they have similar ’rationale’ is not sufficient to demonstrate explanation fidelity to M, because in this case an explanation is actually a statement about the surrogate model: at best, it is a claim that ”feature k was important to M in making the prediction y about input x because a linear model M' which has similar inputs and outputs (locally) to M used k”, but this claim effectively denies the Rashomon effect by assuming that the same explanation applies to any models which return the same or similar outcomes for the same inputs.

The ”Prediction” assumption covers the work of Alvarez-Melis and Jaakkola (2018), who argue that explanations should be robust to small perturbations in input by satisfying a distance inequality, and raise the question of whether an explanation method should be robust if a classifier is not robust. Our definition of fidelity answers this question with an emphatic yes; a faithful explanation should be a representation of how the output of the model changes with respect to its input, and therefore a faithful explanation method should be robust to the same degree as the model itself.

Techniques using the ”Linearity assumption” to assess the fidelity of explanations include using a toy data set where ground-truth relevance is known, and comparing the successive removal or addition of features with either the model performance with reference to ground-truth labels, or the model outputs (Arras et al., 2019; Hooker et al., 2019; Samek et al., 2016; Montavon et al., 2018; Arras et al., 2016; Adebayo et al., 2018; Ancona et al., 2017). While our definition of fidelity shares in common with this assumption the belief that 'parts of the input are more important than others', we dispute the requirement that the contributions from different inputs are independent, with an example in 4.3.1. Furthermore, most of these methods base their evaluation on the change in predictive output (or even change in predictive accuracy). This reliance on the ground-truth labels precludes the explanation of models with poor performance, since it would be unclear whether the explanations are faithful to the model if both the explanation and the model are incorrect on classifications (this critique also applies to the "Model" assumption). It also unnecessarily discretises the outputs of the explanation method, resulting in a more coarse evaluation. On the other hand, evaluating fidelity based on a toy data set avoids this possibility (e.g. Arras et. al (2019)), but whether this successfully extrapolates to real data is unclear.

3.2. Counterfactuals

Counterfactual explanations are specifically designed to answer the "what if things had been different?" question, by providing counterfactual examples for the sample to be explained, and are therefore causal statements about a model (Wachter et al., 2017; Martens and Provost, 2014; Laugel et al., 2018). Counterfactual explanations benefit from being model agnostic since they only require an optimization problem to be solved to identify counterfactual examples (Wachter et al., 2017). They also align with wider research about how humans interpret explanations: as Mittelstadt, Russel and Wachter (2019) argue, "human explanations are contrastive". However, a counterfactual example may lose it's explanatory power if it is not appropriate for the sample space (e.g. "this animal would not be classified as a dog if it had 17 legs"). In their discussion, Laugel et. al(2019) propose justification requirements which counterfactual explanations should satisfy in order to be acceptable, based on a topological definition of path connectedness in the data set, but this concern attempts to address both the fidelity of an explanation to the model and sample being explained, as well as the relationship between the explanation and the underlying data, and Laugel et. al (2019) go so far as to state that "there is no existing satisfying way to provide post-hoc explanations that are both faithful to the classifier and to ground-truth data". In this paper, we focus only on the validity of explanations with respect to the model. On the other hand, counterfactual approaches may suffer from the Rashomon effect, where multiple differing explanations are possible (Molnar, 2018), although this can be avoided to some degree depending on the method used to generate counterfactuals (Wachter et al., 2017). Finally, counterfactual approaches can be challenging if the input space is high-dimensional, for example with categorical features at multiple levels, and different approaches may be required for different data types (Mittelstadt et al., 2019).

3.3. Causal explanations

While an in-depth discussion of this philosophy is outside the scope of this paper, we are mindful of Miller's (2017) warning that understanding and evaluating explanations should not be the remit of machine learning researchers alone [1]. Most relevant to this paper is work on causal explanations through counterfactuals (Lewis, 2013; Scriven, 1975), and we take Woodward's (2003) position that explanations must answer "w-questions", that is, "what-if-things-had-been-different?". Within the machine learning domain, Watson and Floridi (2019) use a game-theoretic approach to develop a framework for generating optimal explanations for a given prediction, drawing on the causal interventionism of Pearl (2009). Zhao and Hastie (2019) demonstrate a link between partial dependence plots and Pearl's do-calculus (2009), but in this case the causal interpretations are with respect to the underlying inputs X and true outputs Y, rather than causal interpretations of predictions from the model. Partial dependence plots (PDP) (and their more recent descendents Individual Conditional Expectation (ICE) plots and Accumulated Local Effects (ALE) plots) (Friedman, 2001; Goldstein et al., 2015; Apley and Zhu, 2016) are visual methods which attempt to explain model behaviour by perturbing input features to M and plotting the results. While PDPs do this globally, marginalising out the effects for a single variable across the data set, ICEs and ALEs accumulate sample-wise perturbations. These methods all share the shortcoming that more than two features are impossible to visualise, and hence each plot can only summarise a part of the model behaviour.

[1] These inmates, at least, will not take charge of the asylum!

4. Methods

4.1. Synthetic data generation method and modelling approach

The data is generated using a mixture of Bernoulli distributions, $y \sim B(1, p = f(X))$:

$$f : X \to \begin{cases} 0 < X \leq 1 & p = 0 \\ 1 < X \leq 2 & p = 0.1 \\ 2 < X \leq 4 & p = 0.2 \\ 4 < X \leq 5 & p = 0.5 \\ 5 < X \leq 7 & p = 0.8 \\ 7 < X \leq 8 & p = 0.9 \\ 8 < X \leq 10 & p = 1 \end{cases}$$

We model the data generating mechanism with a simple, one-layer neural network consisting of 8 nodes with Relu activation and a Softmax layer, using the Tensorflow library (Abadi et al., 2015).

4.2. Explanation methods

1. *Locally-interpretable Model-agnostic explanations:* LIME (2016) is a framework to identify an 'interpretable data representation' of the model input features, and an interpretable surrogate model (this could be a linear model, decision tree, or falling rule lists) which finds a tradeoff between the complexity of the surrogate model (number of coefficients, depth of trees etc.) and the "local fidelity" of the surrogate to the original model on the interpretable data representation - defining fidelity as the distance between the prediction probabilities of the original model and the surrogate model.

2. *Layerwise Relevance Propagation:*This method propagates a prediction backward through a neural network, using the weights and activation values to compute the most important input values for a given input. (Montavon et al., 2019) The propagation is a conservative procedure, in that the prediction is conserved as it is redistributed to the lower layer neurons.

4.3. Calculation of Explanation Fidelity

Given a d-dimensional input space X, output space Y, and a model $M : X \to Y$, a feature-attribution explanation method is a function $E : (x, M) \to e$, where e is a d-dimensional vector of weights such that e_i denotes the 'importance' or 'contribution' of x_i with respect to the model output y. We argue that the only meaningful interpretation of e_i is a causal one: *if x_i had a different value, y would be different.* [2] In order to measure this for individual features, we propose the following:

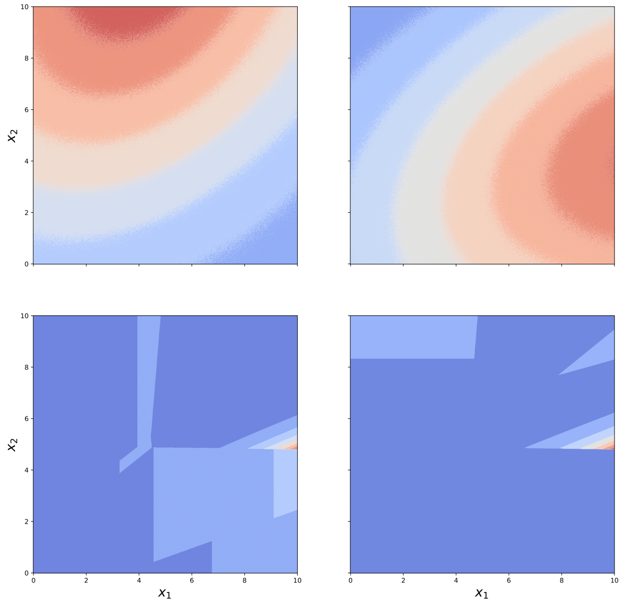


Figure 2. Contour heat maps for the output of LIME (top row) and LRP explanation methods (bottom row) across x_1 (first column) and x_2 (second column), from most important (red) to least important (blue). LIME shows explanations approximately corresponding to the intuition described in 2 and the perturbation contours in in 4, where $x1$ is more important when $x2 > 5$ and vice versa. There is significantly less variation in the output from LRP, with a small region of high importance when $x1$ is near 10, and $x2$ is near 5, and little correspondence with the perturbation contours in 4

4.3.1. Definition

Model-based perturbation value: Given a model M and a sample $x \in X^d$, then for $i = 1, \ldots, d$, we create x'_i as a permutation of x in the ith feature. Since the ith element can be permuted in k distinct ways, let $x'_{i,j}$ denote the jth permutation, such that an estimate of the change in y over all possible permutations of the ith element is given by

$$\delta y_i = \frac{\sum_{j=1}^{k} d(M(x), M(x'_{i,j}))}{k} \tag{1}$$

where d is an appropriate distance metric. Model-based perturbation has much in common with counterfactual methods, but the difference here is in our purpose to evaluate explanations, rather than to provide them. As noted in the

[2] Note that y may not be continuous for all models, or in a classification setting for example. In such situations, a discrete set of high-cardinality, a prediction probability space, or even a post-processing calibration step may be appropriate. Further work is required in this area.

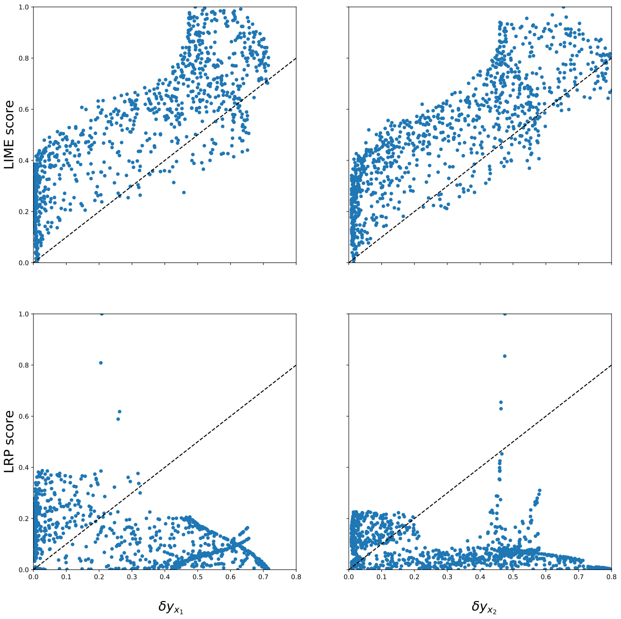
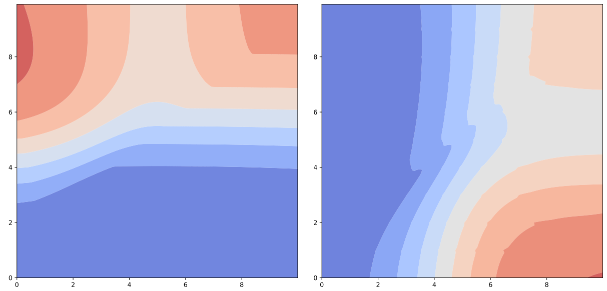


Figure 3. Scatter plots for x_1 and x_2 showing the relationship between δy_{x_i}, the model perturbation values, and the output from LIME and LRP. For the purposes of visual clarity, the LRP and LIME values were normalised since they are essentially dimensionless 'importance' values, but δy_{x_i} values are not normalised since they represent the average change in prediction probability for M when x_i is perturbed and therefore are meaningful in relation to the problem. The dashed line $x_1 = x_2$ is plotted for reference.

related work section, model-based perturbation may not be practical as an explanation method: high-dimensional input spaces make it difficult to calculate or "explanations" can be generated which do not make sense within the domain. However, our goal is less ambitious: given an explanation e as defined above, we can assess the fidelity of this explanation by measuring the correlation between e_i and δy_i, essentially asking how well it answers the "w-question". In this paper we use the Pearson Correlation Coefficient $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$ to calculate the correlation between the model-based perturbations and the explanation values from LIME and LRP.

In using model-based perturbation to calculate fidelity, we therefore accept the "linearity" assumption (Jacovi and Goldberg, 2020) insofar as stating that some parts of the input have more effect on the model output than others. However, it is not clear that feature independence necessarily follows. While any explanation method which is an injective function $X \rightarrow IR$ implies independence of input features, there is no reason as to why this should be the case in general. δy_i as defined above could be calculated for interaction terms between features x_i and x_j for example, but since model-based perturbation is for evaluating e, this would be necessary only when e explicitly makes a



(a) Perturbations of x_1

(b) Perturbations of x_2

Figure 4. Contour heat maps for the perturbation scores of M in x_1 and x_2,

statement about x_i and x_j covarying.

5. Results and Discussion

The results in Table 1 show that our causal intuition about points A - D in the data set shown in Figure 1 1a aligns with the average change in model output, δy. We now present a more in depth analysis across the whole test set, as shown in Figure 2 in order to evaluate the fidelity of LIME and LRP for M on this data set. While in general, we propose model-based perturbations as an evaluation method, the simplicity of this experiment allows us to use perturbation analysis to generate faithful explanations for M, since they directly answer the question of "what-if-things-had-been-different", and can be computed efficiently in this case. The contour maps show that when $x_2 > 5$, x_1 has the larger score, and similarly for x_2 when $x_1 > 5$.

To explain the predictions of M, the LIME framework fits a linear model to a sample of inputs and predictions around the point to be explained, and returns the coefficients as an explanation of the prediction. LIME uses the 'Model assumption' (Ribeiro et al., 2016; Jacovi and Goldberg, 2020), which we have argued above is not sufficient for model fidelity. Despite this concern, the outputs of LIME are quite faithful to the the model output in this example; x_1 is most important when $x_2 > 5$, and vice-versa, and the contour map shows a graduated change in importance which is in line with the decision contour map for M. The PCC between LIME and the pertubation score is 0.85 for x_1, and 0.83 for x_2.

From the LRP algorithm, we deduce that the output is a point-wise decomposition of which "conserves class relevance on a layer and node basis". In the original paper (Bach et al., 2015), justifications for the value of LRP include removing the "most relevant" pixels from MNIST images, noting that "on average over all digits, flipping the highest scoring pixels at first results in a fast decline of the prediction for the true class, and at some point another class is predicted.", so we can conclude that the authors consider LRP explanations to answer the w-question, and furthermore that this explanation method assumes Jacovi's (2020) "Linearity Assumption", that each input is independent as to its effect on the model output. The LRP contour maps for x and y across the test set shows that LRP in fact gives only a small number of differing explanations across the data set, and the PCC score between the LRP values and perturbation values is -0.29 for x, and -0.45 for y. While these values could in fact indicate a negative linear correlation, the scatter plots in Figure 3 show that there is no meaningful relationship.

In this experiment, we can conclude that the LIME model provides a more faithful explanation model for M than LRP, as evidenced by the PCC scores above and scatter plots in Figure 3. While we use the Model perturbation method as a form of explanation in this example, we again acknowledge that perturbation methods run in to problems in higher dimensional spaces, and state that in general the method we propose is for evaluation of fidelity, rather than explaining each point.

6. Conclusions

We argue that explanations for the predictions of a model should be interpreted as causal statements about how a model maps its inputs to its predictions, and that explanations should be evaluated as to how accurately they can provide a causal explanation for model behaviour ahead of any other evaluation - this is the fidelity of the explanation method. In this paper, we present a model trained on a simple synthetic data set, calculate faithful explanations using perturbations of the input data, and compare these with explanations from LIME and Layerwise Relevance Propagation. We discuss the results of this experiment in the context of related work about explanation fidelity and causal interpretations of explanations. While a formal definition of fidelity which is applicable to all models and data types (images, text, tabular etc.) remains a challenge for the research community, this paper demonstrates the importance of attempting to assess fidelity of explanations as causal statements about a model in order to validate the explanations.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y. and Zheng, X. (2015), 'TensorFlow: Large-scale machine learning on heterogeneous systems'. Software available from tensorflow.org.
URL: *https://www.tensorflow.org/*

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M. and Kim, B. (2018), Sanity checks for saliency maps, *in* 'Advances in Neural Information Processing Systems', pp. 9505–9515.

Alvarez-Melis, D. and Jaakkola, T. S. (2018), 'On the robustness of interpretability methods', *arXiv preprint arXiv:1806.08049* .

Ancona, M., Ceolini, E., Öztireli, C. and Gross, M. (2017), 'Towards better understanding of gradient-based attribution methods for deep neural networks', *arXiv preprint arXiv:1711.06104* .

Apley, D. W. and Zhu, J. (2016), 'Visualizing the effects of predictor variables in black box supervised learning models'.

Arras, L., Horn, F., Montavon, G., Müller, K.-R. and Samek, W. (2016), Explaining predictions of non-linear classifiers in NLP, *in* 'Proceedings of the 1st Workshop on Representation Learning for NLP', Association for Computational Linguistics, Berlin, Germany, pp. 1–7.
URL: *https://www.aclweb.org/anthology/W16-1601*

Arras, L., Osman, A., Müller, K.-R. and Samek, W. (2019), Evaluating recurrent neural network explanations, *in* 'Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP', pp. 113–126.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R. and Samek, W. (2015), 'On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation', *PloS one* **10**(7).

Collaris, D., Vink, L. M. and van Wijk, J. (2018), Instance-level explanations for fraud detection: a case study, *in* '2018 ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)'.

Friedman, J. H. (2001), 'Greedy function approximation: A gradient boosting machine', *The Annals of Statistics* **29**(5), 1189–1232.
URL: *http://www.jstor.org/stable/2699986*

Goldstein, A., Kapelner, A., Bleich, J. and Pitkin, E. (2015), 'Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation', *Journal of Computational and Graphical Statistics* **24**(1), 44–65.
URL: *https://doi.org/10.1080/10618600.2014.907095*

Hooker, S., Erhan, D., Kindermans, P.-J. and Kim, B. (2019), A benchmark for interpretability methods in deep neural networks, *in* 'Advances in Neural Information Processing Systems', pp. 9734–9745.

Jacovi, A. and Goldberg, Y. (2020), 'Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness?'.

Julia Angwin, J. L. (2016), 'Machine bias'.
URL: *https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing*

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. and Detyniecki, M. (2018), Comparison-based inverse classification for interpretability in machine learning, *in* 'International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems', Springer, pp. 100–111.

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X. and Detyniecki, M. (2019), 'The dangers of post-hoc interpretability: Unjustified counterfactual explanations', *arXiv preprint arXiv:1907.09294* .

Lewis, D. (2013), *Counterfactuals*, Wiley.
URL: *https://books.google.co.uk/books?id=bCvnk3JMvfAC*

Lipton, Z. C. (2018), 'The mythos of model interpretability', *Commun. ACM* **61**(10), 36–43.
URL: *https://doi.org/10.1145/3233231*

Martens, D. and Provost, F. (2014), 'Explaining data-driven document classifications', *Mis Quarterly* **38**(1), 73–100.

Miller, T., Howe, P. and Sonenberg, L. (2017), 'Explainable ai: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences', *arXiv preprint arXiv:1712.00547* .

Mittelstadt, B., Russell, C. and Wachter, S. (2019), Explaining explanations in ai, *in* 'Proceedings of the Conference on Fairness, Accountability, and Transparency', FAT* '19, Association for Computing Machinery, New York, NY, USA, p. 279–288.
URL: *https://doi.org/10.1145/3287560.3287574*

Molnar, C. (2018), 'A guide for making black box models explainable', *URL: https://christophm. github. io/interpretable-ml-book* .

Montavon, G., Binder, A., Lapuschkin, S., Samek, W. and Müller, K.-R. (2019), Layer-wise relevance propagation: an overview, *in* 'Explainable AI: Interpreting, Explaining and Visualizing Deep Learning', Springer, pp. 193–209.

Montavon, G., Samek, W. and Müller, K.-R. (2018), 'Methods for interpreting and understanding deep neural networks', *Digital Signal Processing* **73**, 1–15.

Pearl, J. (2009), *Causality: Models, Reasoning and Inference*, 2nd edn, Cambridge University Press, USA.

Ribeiro, M. T., Singh, S. and Guestrin, C. (2016), "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', KDD '16, Association for Computing Machinery, New York, NY, USA, p. 1135–1144.
URL: *https://doi.org/10.1145/2939672.2939778*

Rudin, C. (2019), 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence* **1**(5), 206–215.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.-R. (2016), 'Evaluating the visualization of what a deep neural network has learned', *IEEE transactions on neural networks and learning systems* **28**(11), 2660–2673.

Scriven, M. (1975), 'Causation as explanation', *Noûs* **9**(1), 3–16.
URL: *http://www.jstor.org/stable/2214338*

Wachter, S., Mittelstadt, B. and Russell, C. (2017), 'Counterfactual explanations without opening the black box: Automated decisions and the gdpr', *Harv. JL & Tech.* **31**, 841.

Watson, D. and Floridi, L. (2019), 'The explanation game: A formal framework for interpretable machine learning', *Available at SSRN 3509737* .

Woodward, J., Woodward, J., Press, O. U. and (Firm), P. (2003), *Making Things Happen: A Theory of Causal Explanation*, Oxford scholarship online, Oxford University Press.
URL: *https://books.google.co.uk/books?id=LrAbrrj5te8C*

Zhao, Q. and Hastie, T. (2019), 'Causal interpretations of black-box models', *Journal of Business & Economic Statistics* **0**(0), 1–10.
URL: *https://doi.org/10.1080/07350015.2019.1624293*