Towards Providing Causal Explanations for the Predictions of any Classifier

Adam White and Artur d'Avila Garcez

City, University of London, Northampton Square, London, EC1V 0HB, UK
{adam.white.2, a.garcez}@city.ac.uk

We propose a novel method for explaining the predictions of any classifier. In our approach, local explanations are expected to explain both the outcome of a prediction and how that prediction would change if things-had-been-different. Furthermore, a satisfactory explanation must also be measurable and state how well it can explain a model. It must *know when it does not know* [2]. A system called **C**ounterfactual; **L**ocal **E**xplanations vi**A** **R**egression (CLEAR) is introduced and evaluated. This is based on a concept of counterfactual explanation from the philosophy of science's analysis of causality [6,3]. CLEAR generates w-counterfactuals that state minimum changes necessary to flip a prediction's classification. CLEAR then builds local regression models, using the w-counterfactuals to measure and improve the fidelity of its regressions. By contrast, the popular LIME method [4], which also uses regression to generate local explanations, neither measures its own fidelity nor generates counterfactuals. When applied to multi-layer perceptrons (MLPs) trained on four datasets, CLEAR improves on the fidelity of LIME by approximately 40%. As well as providing local explanations of a classifier, CLEAR can also be used to identify 'real-world' causal relationships that have implicitly been discovered by the classifier.

Perhaps the most influential account of counterfactual explanations comes from Woodward[6]. It is based on Pearl's theory of causation [3], which can be roughly summarised in the idea that variable X is a cause of variable Y, if an ideal intervention on X would change the value of Y. Woodward states that a satisfactory explanation consists in showing patterns of counterfactual dependence. By this he means that it should answer a set of what-if-things-had been-different? questions, which specify how the explanandum (i.e. the phenomenon to be explained) would change if, contrary to the fact, input conditions had been different. It is in this way that a user can understand the relevance of different features, and understand the different ways in which they could change the value of the explanandum. Central to Woodward's notion is the requirement for an explanatory generalization:

> "Suppose that M is an explanandum consisting in the statement that some variable Y takes the particular value y. Then an explanans E for M will consist of (a) a generalization G relating changes in the value(s) of a variable X (where X may itself be a vector or n-tuple of variables X_i) with changes in Y, and (b) a statement (of initial or boundary conditions) that the variable X takes the particular value x."

In Woodward's analysis, X causes Y. For our purposes, Y can be taken as the machine learning system's predictions and X as the system's input features. The required generalization can be a regression equation that captures the machine learning system's local input-output behaviour.

CLEAR treats machine learning systems as black boxes, whose inner working are often complex and beyond the capacities of humans to understand. It explains a machine learning system by explaining its input-output behaviour. CLEAR provides counterfactual explanations by building on the strengths of two state-of-the-art explanatory methods, while at the same time addressing their weaknesses. The first is by Wachter et al. [5] who argue that single predictions are explained by what we shall term as w-counterfactuals. For example, if

a banking machine learning system declined Mr Jones loan application, a w-counterfactual explanation might be that Mr Jones would have received his loan, if his annual salary had been $35,000 instead of the $32,000 he currently earns. The $3000 increase would be just sufficient to flip Mr Jones to the desired side of the banking systems decision boundary. The second method is by Riberio et al. [4] who argue for Local Interpretable Model-Agnostic Explanations (LIME). These explanations are created by building a regression model that seeks to approximate the local input-output behaviour of the machine learning system.

In isolation,w-counterfactuals do not provide explanatory generalizations relating X to Y and therefore are not satisfactory explanations. For example they do not explain how X's features interact with each other. LIME, on the other hand, does not measure the fidelity of its regressions and does not calculate counterfactuals. Furthermore LIME does not adequately select the data to use in its regressions. In the case of counterfactual explanations this data needs to extend from a target observation to the nearest points of the classifier's decision boundary.

CLEAR is based on the concept of a w-perturbation:

Definition Let $\min_f(\boldsymbol{x})$ denote a vector resulting from applying a minimum change to the value of one feature f in $\boldsymbol{x}$ such that $m(\min_f(\boldsymbol{x})) = y'$ and $m(\boldsymbol{x}) = y$, class$(y) \neq$ class(y'). Let $v_f(\boldsymbol{x})$ denote the value of feature f in $\boldsymbol{x}$. A ***w*-perturbation** is defined as the change in value of feature f for a target class y', that is $\mid v_f(\boldsymbol{x}) - v_f(\min_f(\boldsymbol{x})) \mid$. For example, for the w-counterfactual that Mr Jones would have received his loan if his salary had been $35,000, a w-perturbation for *salary* is $3000. CLEAR compares each w-perturbation with an estimate of that value, call it *estimated* w-perturbation, calculated using its local regression, to produce a *fidelity error*, as follows:

$$\textit{fidelity error} = \mid \text{estimated } w\text{-perturbation} - w\text{-perturbation} \mid$$

CLEAR generates an explanation of prediction y made by machine learning system m for observation $\boldsymbol{x}$ by the following steps:

1. Determine $\boldsymbol{x}$'s w-perturbations for a user-selected set of features. This is achieved by querying m with feature values starting with $\boldsymbol{x}$ and progressively moving away.
2. Generate synthetic observations that are then labelled by m.
3. Create a balanced neighbourhood data set. Synthetic observations that are near to $\boldsymbol{x}$ are selected with the objective of achieving a dense cloud of points around m's decision boundaries.
4. Perform a step-wise regression on the neighbourhood data set. The regression can include second degree terms and interaction terms.
5. Estimate the w-perturbations by substituting $\boldsymbol{x}$'s w-counterfactual values from $\min_f(\boldsymbol{x})$, other than for feature f, into the regression equation and calculating the value of f.
6. Measure the fidelity of the regression coefficients. Fidelity errors are calculated by comparing the actual w-perturbations determined in step 1 with the estimates calculated in step 5.
7. Iterate to best explanation. Because CLEAR produces fidelity statistics, its parameters can be iteratively changed to achieve a better trade-off between interpretability and fidelity.
8. CLEAR also provides the option of adding $\boldsymbol{x}$'s w-counterfactuals, $\min_f(\boldsymbol{x})$, to $\boldsymbol{x}$'s neighbourhood data set. The w-counterfactuals are weighted and act as soft constraints on CLEAR's subsequent regression.

For CLEAR an explanation is a tuple $< w, w', r, e >$, where w and w' are w-perturbations (actual and estimated), r is a regression equation and e are fidelity errors.

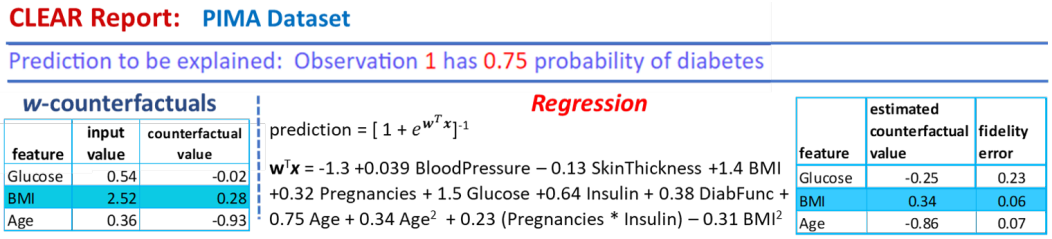

Fig. 1: Example of CLEAR's w-counterfactual report for a single prediction.

Experiments were carried out with four UCI datasets, each being used to train an MLP with a softmax output layer. CLEAR calculated the % of estimated w-perturbations with an error less than a threshold (set at 1/4 standard deviation) In order to enable comparisons with LIME, CLEAR includes an option to run LIME's algorithms for creating synthetic data and generating regression equations; CLEAR then calculates the corresponding w-counterfactuals. CLEAR's regressions were found to be significantly better than LIME's. The best results were obtained by including w-counterfactuals in the neighbourhood data sets (step 8 of the CLEAR method); this was expected, as adding these weighted data points results in a data set capable of representing better the relevant neighbourhood, with CLEAR then being able to provide a regression equation that is more faithful to w-counterfactuals.

Table 1: Comparison of % fidelity of CLEAR and LIME

	Pima	Adult	Credit	Breast
CLEAR- not using w-counterfactuals	57% ± 0.8	80% ± 0.9	39% ± 1.3	54% ± 1.1
CLEAR- using w-counterfactuals	77% ± 0.8	80% ± 0.8	55% ± 1.7	81% ± 1.3
LIME algorithms	27% ± 1.4	26% ± 0.6	12% ± 0.5	14% ± 0.3

CLEAR can also be used in the discovery of 'real-world' causal relationships. CLEAR's regression equations satisfy Woodward's requirements for a causal explanation. Each regression equation captures the local relationship between X and Y, and supports counterfactuals (with the changes in the values of features in step 5 of the CLEAR method corresponding to Pearl's 'ideal-interventions'). Hence CLEAR identifies the local causal relationships implicit in a machine learning system. These will correspond to 'real-world' causal relationships if: (i) the input features for a machine learning system are potential 'real-world' direct causes of Y - this could be verified by an expert with domain knowledge of the mechanisms generating Y (ii) the features in X are independent and (iii) the machine learning system is of high accuracy. CLEAR would then reveal how the 'real-world' causal relationships of the system being modelled can vary by locality, an idea that is being actively researched within the philosophy of causation [1].

References

1. N. Cartwright. *The dappled world. A study of the boundaries of science.* Cambridge University Press, 1999.
2. Zoubin G. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, 2015.
3. J. Pearl. *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York, NY, USA, 2nd edition, 2009.
4. M. Ribeiro, S. Singh, and C. Guestrin. Why should I trust you? Explaining the predictions of any classifier. In *Proc. ACM SIGKDD 2016*, KDD '16, pages 1135–1144, New York, NY, USA, 2016. ACM.
5. S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *CoRR*, abs/1711.00399, 2017.
6. J. Woodward. *Making things happen: a theory of causal explanation.* Oxford University Press, 2003.