A Baseline for Shapley Values in MLPs: from Missingness to Neutrality

Cosimo Izzo, Aldo Lipani, Ramin Okhrati, Francesca Medda

University College London, London, United Kingdom

Abstract. Deep neural networks have gained momentum based on their accuracy, but their interpretability is often criticised. As a result, they are labelled as black boxes. In response, several methods have been proposed in the literature to explain their predictions. Among the explanatory methods, Shapley values is a feature attribution method favoured for its robust theoretical foundation. However, the analysis of feature attributions using Shapley values requires choosing a baseline that represents the concept of missingness. An arbitrary choice of baseline could negatively impact the explanatory power of the method and possibly lead to incorrect interpretations. In this paper, we present a method for choosing a baseline according to a neutrality value: as a parameter selected by decision-makers, the point at which their choices are determined by the model predictions being either above or below it. Hence, the proposed baseline is set based on a parameter that depends on the actual use of the model. This procedure stands in contrast to how other baselines are set, i.e. without accounting for how the model is used. We empirically validate our choice of baseline in the context of binary classification tasks, using two datasets: a synthetic dataset and a dataset derived from the financial domain.

1   Introduction and Background

In disciplines such as economics, finance and healthcare, the ability to explain predictions is as important as having a model that performs well [1, 2]. A solution to the problem is provided by feature attribution methods, which are used to indicate how much each feature contributes to the prediction for a given example. A theoretically grounded feature attribution method is provided by Shapley values and their approximations [3]. When explaining a prediction with Shapley values, we need to perform two steps. First, we define a *baseline*. Then, we compute the Shapley values for a given example. While most works focused on the latter, the former has not been sufficiently explored, despite the implications that the baseline definition carries to correctly interpret Shapley values.

For a neural network G_θ with parameters θ, and n input features, the contribution of the feature j calculated according to the Shapley value for the input $x = [x_1, x_2, \ldots, x_n]$ is given by:

$$\sum_{S \subseteq P \setminus \{j\}} \frac{|S|!(|P| - |S| - 1)!}{|P|!} \left(G_\theta(\tilde{x}_{S \cup \{j\}}) - G_\theta(\tilde{x}_S) \right), \qquad (1)$$

where P is the collection of all feature indexes, the element i of the vector $\tilde{x}_S$ is given by $\tilde{x}_{S,i} = x_i \mathbb{1}_{\{i \in S\}} + b_i \mathbb{1}_{\{i \notin S\}}$ (similarly for $\tilde{x}_{S \cup \{j\}}$), and b_i is the baseline

value for the feature i. The *baseline* models the missingness of a feature, i.e., it replaces that feature when it is absent. As it is argued by Sturmfels et al. [4], the concept of missingness is not well defined and explored in machine learning. The standard practice in setting up the baseline is to assign a vector of **zeros** [5, 6, 7, 8] for all features, which coincides with the **average** vector baseline when features are standardised. However, this choice could be misleading. For example, in a classification task, with binary features representing the presence or absence of an entity, given an example and its prediction value, such a baseline would always measure a null contribution for each feature with value equal to zero. A way to address this zero-baseline insensitivity problem is to use the maximum distance baseline (**mdb**) [4]. This baseline consists in taking the furthest observation from the current one in an L^1 norm. This approach unequivocally creates incoherent justifications for the interpretations provided by the model due to the correlation of the baseline with the underlying dataset. Alternatively, one can consider a sample of baselines and average the attributions computed over each baseline, for example by using the underlying empirical distributions of the dataset ($\mathbf{p}_X$) [3, 4, 9]. However, the $\mathbf{p}_X$ baseline increases the computational cost of estimating feature attributions linearly with respect to the number of draws. Moreover, this choice of baseline does not allow the setting of a reference value on the model output when computing the Shapley values. This is important when decisions are taken with respect to a specific value of the model.

The evaluation of explainability methods from a quantitative perspective is difficult due to the lack of a clear definition of what is a correct explanation [6, 8]. Many extrinsic evaluations of explainability power have been developed [10, 11]. These measures may be misled by potential feature correlations. Thus, to avoid this issue, we also perform an analysis using a synthetic dataset with independent features. Further, we evaluate the explainability power across two dimensions. First, along the feature importance dimension by means of ROAR [10], and second, along the information gain/loss dimension by means of *absolute logits*. Once the most important features have been identified on a per example basis, ROAR measures the difference in performance between the original model and a new one trained with these features replaced by their average values. On the other hand, the choice of absolute logits is motivated also by information theory. Indeed, logits can be seen as the difference between two Shannon information. As this measure decreases (increases), the Shannon information differential decreases (increases). Furthermore, when the Shannon information differential is zero, we can argue that the model becomes uninformative.

2   The Neutral Baseline

In this section, we theoretically justify the existence of a baseline according to well defined concepts of neutrality value and fair baselines. Following Bach et al. [12], we argue that the baseline should live on the decision boundary of the classifier.

Definition 1 (Neutrality Value)**.** *Given a model prediction $\hat{y}$ and a decision maker, we say that the value α is neutral if the decision maker's choice is determined by the value of $\hat{y}$ being either below or above α.*

Generally, the concept of missingness is domain specific. However, every time we are faced with a decision boundary and a model, a natural choice is to consider missingness as missing information from the model to the user to take a choice, i.e.: when the model is at the neutrality value. The idea is that this neutrality value can lead to a point in the input domain that could be used as a baseline. However, given a neutrality value and a single-layer perceptron (SLP) with more than one continuous input feature, there are an infinite number of possible combinations of such inputs that lead to the same neutral output. Nevertheless, it is possible to narrow down the set of candidates by being fair in representing each feature in its input space, and given its relation to model $G_\theta(.)$. This ensures that Shapley values are not biased by distributional differences.

Definition 2 (Space of Fair Baselines)**.** *Consider a dataset in $\mathbb{R}^k$, $k \geq 1$, generated by a distribution. The set of fair baselines for a monotonic model $G_\theta(.)$ is given by: $\tilde{B} = \{x^p \in \mathbb{R}^k : x_j^p = C_j^{-1}(\mathbb{1}_{\theta_j > 0} \cdot p + \mathbb{1}_{\theta_j \leq 0} \cdot (1 - p)), p \in [0, 1], j = 1, 2, \ldots, k\}$, where C_j^{-1} is the inverse marginal CDF of x_j $\forall j$.*

Based on the two definitions, neutrality value and space of fair baselines, in what follows we demonstrate the existence of a fair baseline that when given to a SLP returns the neutrality value. Before doing this, we need to state the following two assumptions: **A1**. All activation functions are monotonic and continuous. **A2**. All marginal cumulative distribution functions (CDFs) of the joint CDF of the input features are bijective and continuous.

Proposition 1. *Given a SLP (G_θ) satisfying **A1**, a dataset satisfying **A2**, and a neutrality value α in the image of G_θ, then there exists at least a fair baseline x such that $G_\theta(x) = \alpha$.*

Proof. We need to prove that $\alpha \in G_\theta(\tilde{B})$ where $G_\theta(\tilde{B})$ is the image of $\tilde{B}$ under G_θ. Suppose that I is the image of the SLP. We show that $I \subseteq G_\theta(\tilde{B})$ which proves the result, since $\alpha \in I$. We start by showing that $\inf G_\theta(\tilde{B}) \leq \inf I$ and that $\sup G_\theta(\tilde{B}) \geq \sup I$. Consider vector $x^0 \in \tilde{B}$ defined by $x^0 = \{x_j^0 = C_j^{-1}(\mathbb{1}_{\theta_j \leq 0}),$ for all $j = 1, 2, \ldots, k\}$. So elements of x^0 are the smallest possible if the coefficients are positive, and the largest possible when they are negative. From Assumption **A1**, it follows that $G_\theta(x^0)$ is the smallest value that the SLP can take. Hence, $G_\theta(x^0) \leq \inf I$. Let us now take the vector $x^1 \in \tilde{B}$ which is defined by: $x^1 = \{x_j^1 = C_j^{-1}(\mathbb{1}_{\theta_j > 0}),$ for all $j = 1, 2, \ldots, k\}$. So elements of x^1 are the largest possible if the coefficients are positive, and the smallest possible when they are negative. From assumption **A1**, it follows that $G_\theta(x^1)$ is the largest value that the SLP can take. Hence, $G_\theta(x^1) \geq \sup I$. Suppose that α is in the image of the SLP. Define function $h : [0, 1] \to G(\tilde{B})$ by $h(p) = G_\theta(x^p) = G(\theta \cdot (x^p)^\top)$ where $x^p \in \tilde{B}$, i.e, $x_j^p = C_j^{-1}(\mathbb{1}_{\theta_j > 0} \cdot p + \mathbb{1}_{\theta_j \leq 0} \cdot (1-p)), j = 1, 2, \ldots, k$. From the above argument, we have that $h(0) = G_\theta(x^0) \leq \alpha \leq G_\theta(x^1) = h(1)$.

Since G and C^{-1} are continuous functions by **A1** and **A2**, h is also continuous. By the intermediate value theorem, there is a $p^* \in [0,1]$ such that $h(p^*) = \alpha$, which means that $G_\theta(x^{p^*}) = \alpha$. □

This proof suggests a way to find one neutral and fair baseline for a SLP. An algorithm using empirical CDFs instead of theoretical ones, requires as inputs a SLP (G_θ), a neutrality value (α), a quantile function for each dimension of input features, a granularity level $\delta > 0$, and a tolerance level $\epsilon > 0$. δ and ϵ control the speed of search and the margin of error in finding a baseline such that $|G_\theta(x) - \alpha| < \epsilon$. This algorithm starts the search from the lowest output value, which is when $p = 0$, and it stops when it reaches a point which is close enough to α. This is possible because using the parameters of the SLP we can restrict and define an order in function of p for the set of fair baselines. This allows us to test these baselines from the smallest to the largest SLP value.

Finding a baseline for MLPs is more complicated, because there is no easy way to order the baselines in function of p as in the SLP case, unless the MLP is monotonic in each of the features. Nevertheless, we observe that a MLP with L layers can be rewritten as a function of $\sum_{l=2}^{L} \prod_{l'=l}^{L} k_{l'}$ SLPs. This is done by replicating every node at layer l, k_{l+1} times, i.e., the number of nodes at layer $l+1$, and considering every node in the layer $l+1$ as a SLP with input given by the layer l. Based on this observation, we can recursively apply the algorithm for the SLP backwards through the layers of the model to recover the neutrality values across those SLPs, from the output layer to the input layer. This will provide $\prod_{l=2}^{L} k_l$ baselines, one for each SLP in the first hidden layer. Finally, in order to aggregate these baselines, we define an equivalent sparse representation of a MLP (SparseMLP), which is constructed by concatenating each of the SLPs defined above. An example of this transformation is provided in Fig. 1. This representation allows us to compute the Shapley value for each example-feature pair by using all fair baselines found at once.
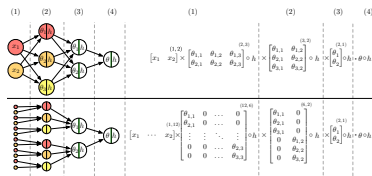

Fig. 1: A MLP (above) and its equivalent sparse representation (below).

3 Experiments

We evaluate in classification tasks the explainability power of our baseline method (**neutral**$_\alpha$) against the **zero**, **average**, $\mathbf{p}_X$, and **mdb** baselines.[1]

---

[1]The code used to run these experiments is available at the following weblink: `https://github.com/cosimoizzo/Neutral-Baseline-For-Shapley-Values`.

Since the output of the trained classifier is probabilistic, i.e., its codomain is in $[0, 1]$ we set the neutrality value α (and so the decision boundary) to 0.5. We use two datasets, a synthetic and a real one. The former to simulate a dataset with independent features and controlled feature importance. The latter to experiment with a real use case.

For the synthetic dataset, we generate 5 independent features from a multivariate normal distribution with 0 mean and unit variance. The importance and sign of each feature are randomly drawn and the partition of the space in two classes is nonlinear. We repeat this 100 times, thus generating 100 synthetic datasets.

The real dataset is about default of credit card clients [13]. The dataset contains one binary target variable and 23 features. The number of observations is 29,351. In order to apply ROAR to such dataset, we need to reduce the number of observations to at least 300. We do so by sampling these observations while keeping the two classes balanced. Additionally, to further reduce the computational cost we use Shapley sampling [14, 15].

We validate on both datasets a MLP with sigmoid activation functions and binary cross-entropy loss, and we use Adam as optimiser. The number of hidden layers and neurons in each layer are chosen via a Monte Carlo sampling of models using the training and validation sets.

(a) Information content synthetic. (b) ROAR synthetic

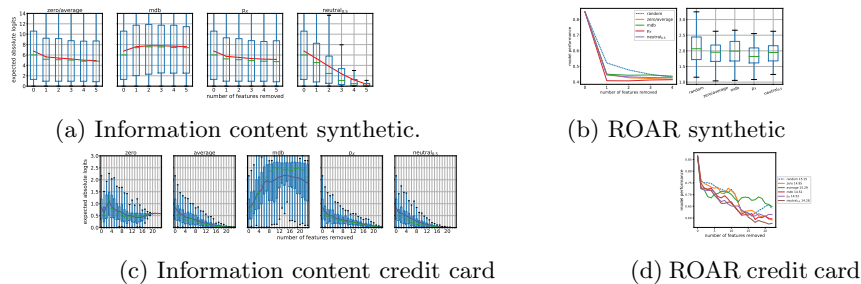(c) Information content credit card (d) ROAR credit card

Fig. 2: Perturbation tests on the synthetic and the credit card dataset.

Fig. 2a, 2c show changes in expected absolute logits. The only baseline that in all datasets guarantees a monotonic decrease to zero in the information content when removing features is the **neutral**$_\alpha$. Thus, this is the only baseline that ensures convergence to a meaningful value when removing features. Fig. 2b, 2d show ROAR scores. While in the synthetic dataset $\mathbf{p}_X$ achieves the best score followed by **zero** and **neutral**$_\alpha$, in the real dataset, it is the **neutral**$_\alpha$ baseline that achieves the best score. Since $\mathbf{p}_X$ and **neutral**$_\alpha$ show similar ROAR scores, the two approaches do equally well in ranking features in order of importance.

4   Conclusion

In this work, we have investigated the identification of baselines for Shapley values based attribution methods and MLPs. We have introduced the concept

of neutrality and fair baselines. Their combination has allowed us to develop a neutral baseline that provides direct interpretation of the Shapley values, being them calculated in relation to the decision threshold of the model. This is in contrast to the baseline methods, where explanations are provided with respect to some arbitrary value with no direct relation to the model. Nevertheless, the computational cost of searching the **neutral**$_\alpha$ baseline increases exponentially with respect to the number of hidden layers. Further, we did not analyse how to apply such method to recurrent networks and how to extend it to regression problems which we leave to future work.

References

[1] Michael Tsang, Dehua Cheng, and Yan Liu. Detecting statistical interactions from neural network weights. In *Proc. of ICLR '18*.

[2] Bryce Goodman and Seth Flaxman. European union regulations on algorithmic decision-making and a right to explanation. *AI Magazine*, 38(3):50–57, 2017.

[3] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proc. of NIPS '17*.

[4] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. *Distill*, 5(1):e22, 2020.

[5] Matthew Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Proc. ECCV '14*.

[6] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proc. of ICML '17*.

[7] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proc. of ICML '17*.

[8] M. Ancona, C. Oztireli, and M. Gross. Explaining deep neural networks with a polynomial time algorithm for shapley values approximation. In *Proc. of ICML '19*, .

[9] Mukund Sundararajan and Amir Najmi. The many shapley values for model explanation. In *Proc. of ICML '20*.

[10] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Proc. of NIPS '19*.

[11] M Ancona, E Ceolini, C Oztireli, and M Gross. A unified view of gradient-based attribution methods for deep neural networks. In *Proc. of NIPS '17*, .

[12] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07 2015.

[13] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36(2):2473–2480, 2009.

[14] Javier Castro, Daniel Gómez, and Juan Tejada. Polynomial calculation of the shapley value based on sampling. *Computers & Operations Research*, 36(5):1726–1730, 2009.

[15] Ramin Okhrati and Aldo Lipani. A multilinear sampling algorithm to estimate shapley values. In *Proc. of ICPR '20*.