Evaluating and Aggregating Feature-based Model Explanations

Umang Bhatt[1,2*] , **Adrian Weller**[1,3] and **José M. F. Moura**[2]

[1]University of Cambridge
[2]Carnegie Mellon University
[3]The Alan Turing Institute

{usb20, aw665}@cam.ac.uk, moura@ece.cmu.edu

Abstract

A feature-based model explanation denotes how much each input feature contributes to a model's output for a given data point. As the number of proposed explanation functions grows, we lack quantitative evaluation criteria to help practitioners know when to use which explanation function. This paper proposes quantitative evaluation criteria for feature-based explanations: low sensitivity, high faithfulness, and low complexity. We devise a framework for aggregating explanation functions. We develop a procedure for learning an aggregate explanation function with lower complexity and then derive a new aggregate Shapley value explanation function that minimizes sensitivity.

1 Introduction

There has been great interest in understanding black-box machine learning models via post-hoc explanations. Much of this work has focused on feature-level importance scores for how much a given input feature contributes to a model's output. These techniques are popular amongst machine learning scientists who want to sanity check a model before deploying it in the real world [Bhatt *et al.*, 2020]. Many feature-based explanation functions are gradient-based techniques that analyze the gradient flow through a model to determine salient input features [Shrikumar *et al.*, 2017; Sundararajan *et al.*, 2017]. Other explanation functions perturb input values to a reference output and measure the change in the model's output [Štrumbelj and Kononenko, 2014; Lundberg and Lee, 2017].

With many candidate explanation functions, machine learning practitioners find it difficult to pick which explanation function best captures how a model reaches a specific output for a given input. Though there has been work in qualitatively evaluating feature-based explanation functions on human subjects [Lage *et al.*, 2019], there has been little exploration into formalizing quantitative techniques for evaluating model explanations. Recent work has created auxiliary tasks to test if attribution is assigned to relevant inputs [Yang and Kim, 2019] and has developed tools to verify if the

features important to an explanation function are relevant to the model itself [Camburu *et al.*, 2019].

Borrowing from the humanities, we motivate three criteria for assessing a feature-based explanation: sensitivity, faithfulness, and complexity. Philosophy of science research has advocated for explanations that vary proportionally with changes in the system being explained [Lipton, 2003]; as such, explanation functions should be insensitive to perturbations in the model inputs, especially if the model output does not change. Capturing relevancy faithfully is helpful in an explanation [Ruben, 2015]. Since humans cannot process a lot of information at once, some have argued for minimal model explanations that contain only relevant and representative features [Batterman and Rice, 2014]; therefore, an explanations should not be complex (i.e., use few features).

In this paper, we first define these three distinct criteria: low sensitivity, high faithfulness, and low complexity. With many explanation function choices, we then propose methods for learning an aggregate explanation function that combines explanation functions. If we want to find the simplest explanation from a set of explanations, then we can aggregate explanations to minimize the complexity of the resulting explanation. If we want to learn a smoother explanation function that varies slowly as inputs are perturbed, we can leverage an aggregation scheme that learns a less sensitive explanation function. To the best of our knowledge, we are the first to rigorously explore aggregation of various explanations, while placing explanation evaluation on an objective footing. To that end, we highlight the contributions of this paper:

- We describe three desirable criteria for feature-based explanation functions: low sensitivity, high faithfulness, and low complexity.

- We develop an aggregation framework for combining explanation functions.

- We create two techniques that reduce explanation complexity by aggregating explanation functions.

- We derive an approximation for Shapley-value explanations by aggregating explanations from a point's nearest neighbors, minimizing explanation sensitivity and resembling how humans reason in medical settings.

---

*Contact Author

2 Preliminaries

Restricting to supervised classification settings, let $\boldsymbol{f}$ be a black box predictor that maps an input $\boldsymbol{x} \in \mathbb{R}^d$ to an output $\boldsymbol{f}(\boldsymbol{x}) \in \mathcal{Y}$. An explanation function $\boldsymbol{g}$ from a family of explanation functions, $\mathcal{G}$, takes in a predictor $\boldsymbol{f}$ and a point of interest $\boldsymbol{x}$ and returns importance scores $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}) = \boldsymbol{\phi}_{\boldsymbol{x}} \in \mathbb{R}^d$ for all features, where $\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i = \phi_{\boldsymbol{x}, i}$ (simplified to ϕ_i in context) is the importance of (or attribution for) feature x_i of $\boldsymbol{x}$. By $\boldsymbol{g}_j$, we refer to a particular explanation function, usually from a set of explanation functions $\mathcal{G}_m = \{\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_m\}$.

We denote $D : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ to be a distance metric over explanations, while $\rho : \mathbb{R}^d \times \mathbb{R}^d \mapsto \mathbb{R}_{\geq 0}$ denotes a distance metric over the inputs. An evaluation criterion μ takes in a predictor $\boldsymbol{f}$, explanation function $\boldsymbol{g}$, and input $\boldsymbol{x}$, and outputs a scalar: $\mu(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x})$. $\mathcal{D} = \{(\boldsymbol{x}^i, y^i)\}_{i=1}^n$ refers to a dataset of input-output pairs, and $\mathcal{D}_x$ denotes all $\boldsymbol{x}^i$ in $\mathcal{D}$.

3 Evaluating Explanations

With the number of techniques to develop feature level explanations growing in the explainability literature, picking which explanation function $\boldsymbol{g}$ to use can be difficult. In order to study the aggregation of explanation functions, we define three desiderata of an explanation function $\boldsymbol{g}$.

3.1 Desideratum: Low Sensitivity

We want to ensure that, if inputs are near each other and their model outputs are similar, then their explanations should be close to each other. Assuming $\boldsymbol{f}$ is differentiable, we desire an explanation function $\boldsymbol{g}$ to have low sensitivity in the region around a point of interest $\boldsymbol{x}$, implying local smoothness of $\boldsymbol{g}$. While [Melis and Jaakkola, 2018] codified the property, [Ghorbani *et al.*, 2019] empirically tested explanation function sensitivity. We follow the convention of the former and define *max sensitivity* and *average sensitivity* in the neighborhood of a point of interest $\boldsymbol{x}$.

Let $\mathcal{N}_r = \{\boldsymbol{z} \in \mathcal{D}_{\boldsymbol{x}} \mid \rho(\boldsymbol{x}, \boldsymbol{z}) \leq r, \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{z})\}$ be a neighborhood of datapoints within a radius r of $\boldsymbol{x}$.

Definition 1 (Max Sensitivity). Given a predictor $\boldsymbol{f}$, an explanation function $\boldsymbol{g}$, distance metrics D and ρ, a radius r, and a point $\boldsymbol{x}$, we define the max sensitivity of $\boldsymbol{g}$ at $\boldsymbol{x}$ as:

$$\mu_{\mathrm{M}}(\boldsymbol{f}, \boldsymbol{g}, r; \boldsymbol{x}) = \max_{\boldsymbol{z} \in \mathcal{N}_r} D(\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}), \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{z}))$$

Definition 2 (Average Sensitivity). Given a predictor $\boldsymbol{f}$, an explanation function $\boldsymbol{g}$, distance metrics D and ρ, a radius r, a distribution $\mathbb{P}_{\boldsymbol{x}}(\cdot)$ over the inputs centered at point $\boldsymbol{x}$, we define the average sensitivity of $\boldsymbol{g}$ at $\boldsymbol{x}$ as:

$$\mu_{\mathrm{A}}(\boldsymbol{f}, \boldsymbol{g}, r; \boldsymbol{x}) = \int_{\boldsymbol{z} \in \mathcal{N}_r} D(\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}), \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{z})) \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

3.2 Desideratum: High Faithfulness

Faithfulness has been defined in [Yeh *et al.*, 2019]. The feature importance scores from $\boldsymbol{g}$ should correspond to the important features of $\boldsymbol{x}$ for $\boldsymbol{f}$; as such, when we set particular features $\boldsymbol{x}_s$ to a baseline value $\bar{\boldsymbol{x}}_s$, the change in predictor's output should be proportional to the sum of attribution scores

of features in $\boldsymbol{x}_s$. We measure this as the correlation between the sum of the attributions of $\boldsymbol{x}_s$ and the difference in output when setting those features to a reference baseline. For a subset of indices $S \subseteq \{1, 2, \ldots d\}$, $\boldsymbol{x}_s = \{x_i, i \in S\}$ denotes a sub-vector of input features that partitions the input, $\boldsymbol{x} = \boldsymbol{x}_s \cup \boldsymbol{x}_c$. $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}$ denotes an input where $\boldsymbol{x}_s$ is set to a reference baseline while $\boldsymbol{x}_c$ remains unchanged: $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]} = \bar{\boldsymbol{x}}_s \cup \boldsymbol{x}_c$. When $|S| = d$, $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]} = \bar{\boldsymbol{x}}$.

Remark (Reference Baselines). Recent work has discussed how to pick a proper reference baseline $\bar{\boldsymbol{x}}$. [Sundararajan *et al.*, 2017] suggests using a baseline where $\boldsymbol{f}(\bar{\boldsymbol{x}}) \approx 0$, while others have proposed taking the baseline to be the mean of the training data. [Chang *et al.*, 2019] notes that the baseline can be learned using generative modeling.

Definition 3 (Faithfulness). Given a predictor $\boldsymbol{f}$, an explanation function $\boldsymbol{g}$, a point $\boldsymbol{x}$, and a subset size $|S|$, we define the faithfulness of $\boldsymbol{g}$ to $\boldsymbol{f}$ at $\boldsymbol{x}$ as:

$$\mu_{\mathrm{F}}(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x}) = \operatorname*{corr}_{S \in \binom{[d]}{|S|}} \left(\sum_{i \in S} \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i, \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}\left(\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]}\right) \right)$$

For our experiments, we fix $|S|$ then randomly sample subsets $\boldsymbol{x}_s$ of the fixed size from $\boldsymbol{x}$ to estimate correlation. Since we do not see all $\binom{[d]}{|S|}$ subsets in our calculation of faithfulness, we may not get an accurate estimate of the criterion. Though hard to codify and even harder to aggregate, faithfulness is desirable, as it demonstrates that an explanation captures which features the predictor uses to generate an output for a given input. Learning global feature importances that highlight, in expectation, which features a predictor relies on is a challenging problem left to future work.

3.3 Desideratum: Low Complexity

A complex explanation is one that uses all d features in its explanation of which features of $\boldsymbol{x}$ are important to $\boldsymbol{f}$. Though this explanation may be faithful to the model (as defined above), it may be too difficult for the user to understand (especially if d is large). We define a fractional contribution distribution, where $|\cdot|$ denotes absolute value:

$$\mathbb{P}_{\boldsymbol{g}}(i) = \frac{|\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_i|}{\sum_{j \in [d]} |\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x})_j|}; \ \mathbb{P}_{\boldsymbol{g}} = \{\mathbb{P}_{\boldsymbol{g}}(1), \ldots, \mathbb{P}_{\boldsymbol{g}}(d)\}$$

Note that $\mathbb{P}_{\boldsymbol{g}}$ is a valid probability distribution. Let $\mathbb{P}_{\boldsymbol{g}}(i)$ denote the fractional contribution of feature x_i to the total magnitude of the attribution. If every feature had equal attribution, the explanation would be complex (even if it is faithful). The simplest explanation would be concentrated on one feature. We define complexity as the entropy of $\mathbb{P}_{\boldsymbol{g}}$.

Definition 4 (Complexity). Given a predictor $\boldsymbol{f}$, explanation function $\boldsymbol{g}$, and a point $\boldsymbol{x}$, the complexity of $\boldsymbol{g}$ at $\boldsymbol{x}$ is:

$$\mu_{\mathrm{C}}(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x}) = \mathbb{E}_i\big[-\ln(\mathbb{P}_{\boldsymbol{g}})\big] = -\sum_{i=1}^d \mathbb{P}_{\boldsymbol{g}}(i) \ln(\mathbb{P}_{\boldsymbol{g}}(i))$$

4 Aggregating Explanations

Given a trained predictor $\boldsymbol{f}$, a set of explanation functions $\mathcal{G}_m = \{\boldsymbol{g}_1, \ldots, \boldsymbol{g}_m\}$, a criterion to optimize μ, and a set of

inputs $\mathcal{D}_x$, we want to find an aggregate explanation function g_{agg} that satisfies μ at least as well as any $g_i \in \mathcal{G}_m$. Let $h(\cdot)$ represent some function that combines m explanations into a consensus $g_{\text{agg}} = h(\mathcal{G}_m)$. We now explore different candidates for $h(\cdot)$.

4.1 Convex Combination

Suppose we have two different explanation functions g_1 and g_2 and have chosen a criterion μ to evaluate a g. Consider an aggregate explanation, $g_{\text{agg}} = h(g_1, g_2)$. A potential $h(\cdot)$ is a convex combination where $g_{\text{agg}} = h(g_1, g_2) = wg_1 + (1 - w)g_2 = \boldsymbol{w}^\intercal \mathcal{G}_m$.

Proposition 1. *If D is the ℓ_2 distance and $\mu = \mu_A$ (average sensitivity), the following holds:*

$$\mu_A(g_{\text{agg}}) \leq w\mu_A(g_1) + (1 - w)\mu_A(g_2)$$

Proof. Assuming $\mathbb{P}_x(z)$ is uniform, we can apply the triangle inequality and the convexity of D to arrive at the above. $\square$

A convex combination of explanation functions thus yields an aggregate explanation function that is at most as sensitive as any of the explanation functions taken alone. In order to learn w given g_1 and g_2, we set up an objective as follows.

$$w^* = \arg\min_w \mathbb{E}_{x \sim \mathcal{D}_x} \left[\mu_A(g_{agg}(f, x)) \right] \quad (1)$$

Assuming a uniform distribution around all $x \in \mathcal{D}_x$, we can rewrite this as:

$$w^* = \arg\min_w \int_{x \sim \mathcal{D}_x} \int_{z \in \mathcal{N}_r} D(g_{agg}(x), g_{agg}(z)) \mathbb{P}_x(z) dz dx$$

By Cauchy-Schwartz, we get the following:

$$w^* \leq \arg\min_w \int_{x \sim \mathcal{D}_x} \int_{z \in \mathcal{N}_r} D(a, b) \, dz dx$$

where $a = wg_1(f, x) + (1 - w)g_2(f, x)$ and $b = wg_1(f, z) + (1 - w)g_2(f, z)$. This implies that w^* will be minimal when one element of w^* is 0 and the other is 1. Therefore, a convex combination of two explanation functions, found by solving Equation (1), will be at most as sensitive as the least sensitive explanation function.

4.2 Centroid Aggregation

Another sensible candidate for $h(\cdot)$ to combine m explanation functions is based on centroids with respect to some distance function $D : \mathcal{G} \times \mathcal{G} \mapsto \mathbb{R}$, so that:

$$g_{\text{agg}} \in \arg\min_{g \in \mathcal{G}} \mathbb{E}_{g_i \in \mathcal{G}_m} \left[D(g, g_i)^p \right] = \arg\min_{g \in \mathcal{G}} \sum_{i=1}^m D(g, g_i)^p$$

where p is a positive constant. The simplest examples of distances are the ℓ_2 and ℓ_1 distances with real-valued attributions where $\mathcal{G} \subseteq \mathbb{R}^d$.

Proposition 2. *When D is the ℓ_2 distance and $p = 2$, the aggregate explanation is the feature-wise sample mean.*

$$g_{agg}(f, x) = g_{avg}(f, x) = \frac{1}{m} \sum_{i=1}^m g_i(f, x) \quad (2)$$

Proposition 3. *When D is the ℓ_1 distance and $p = 1$, the aggregate explanation is the feature-wise sample median.*

$$g_{agg}(f, x) = med\{\mathcal{G}_m\}$$

Propositions 2 and 3 follow from standard results in statistics that the mean minimizes the sum of squared differences and the median minimizes the sum of absolute deviations [Berger, 2013].

We could obtain rank-valued attributions by taking any quantitative vector-valued attributions and ranking features according to their values. If D is the Kendall-tau distance with rank-valued attributions where $\mathcal{G} \subseteq \mathcal{S}_d$ (the set of permutations over d features), then the resulting aggregation mechanism via computing the centroid is called the Kemeny-Young rule. For rank-valued attributions, any aggregation mechanism falls under the rank aggregation problem in social choice theory for which many practical "voting rules" exist [Bhatt *et al.*, 2019a].

We analyze the error of a candidate g_{agg}. Suppose the optimal explanation for x using f is $g^*(f, x)$ and suppose g_{agg} is the mean explanation for x in Equation (2). Let $\epsilon_{i,x} = ||g^*(f, x) - g_i(f, x)||$ be the error between the optimal explanation and the ith explanation function.

Proposition 4. *The error between the aggregate explanation $g_{agg}(f, x)$ and the optimal explanation $g^*(f, x)$ satisfies:*

$$\epsilon_{agg} \leq \frac{\sum_{i=1}^n \sum_{j=1}^m \epsilon_{j, x^i}}{mn}$$

Proof. For a fixed x, we have:

$$\epsilon_{\text{agg}, x} = ||g^*(f, x) - g_{\text{agg}}(f, x)||$$

$$= ||\frac{mg^*(f, x)}{m} - \frac{1}{m} \sum_{i=1}^m g_i(f, x)||$$

$$\leq \frac{1}{m} \sum_{i=1}^m ||g^*(f, x) - g_i(f, x)|| = \frac{\sum_{i=1}^m \epsilon_{i,x}}{m}$$

Averaging across $\mathcal{D}_x$, we obtain the result. $\square$

Hence, by aggregating, we do better than when using one explanation function alone. Many gradient-based explanation functions fit to noise [Hooker *et al.*, 2019]. One way to reduce noise would be to aggregate by ensembling or averaging. As proven in Proposition 4, the typical error of the aggregate is less than the expected error of each function alone.

5 Lowering Complexity Via Aggregation

In this section, we describe iterative algorithms for aggregating explanation functions to obtain $g_{\text{agg}}(f, x)$ with lower complexity whilst combining m candidate explanation functions $\mathcal{G}_m = \{g_1, \ldots, g_m\}$. We desire a $g_{\text{agg}}(f, x)$ that contains information from all candidate explanations $g_i(f, x)$ yet has entropy less than or equal to that of each explanation $g_i(f, x)$. As discussed, a reasonable candidate for an aggregate explanation function is the sample mean given by Equation (2). We may want $g_{\text{agg}}(f, x)$ to approach the sample mean, $g_{\text{avg}}(f, x)$; however, the sample mean may have greater complexity than that of each $g_i(f, x)$.

For example, let $\boldsymbol{g}_1(\boldsymbol{f}, \boldsymbol{x}) = [-1, 0]^T$ and $\boldsymbol{g}_2(\boldsymbol{f}, \boldsymbol{x}) = [0, 1]^T$. The sample mean is $\boldsymbol{g}_{\text{avg}}(\boldsymbol{f}, \boldsymbol{x}) = [-0.5, 0.5]^T$. Both $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ have the minimum possible complexity of 0, while $\boldsymbol{g}_{\text{avg}}$ has the maximum possible complexity, $\log(2)$. Our aggregation technique must ensure that $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ approaches $\boldsymbol{g}_{\text{avg}}(\boldsymbol{f}, \boldsymbol{x})$ while guaranteeing $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ has complexity less than or equal to that of each $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$. We now present two approaches for learning a lower complexity explanation, visually represented in Figure 1.

5.1 Gradient-Descent Style Method

Our first approach is similar to gradient descent. Starting from each $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$, we iteratively move towards $\boldsymbol{g}_{\text{avg}}(\boldsymbol{f}, \boldsymbol{x})$ in each of the d directions (i.e., changing the kth feature by a small amount) if the complexity decreases with that move. We stop moving when the complexity no longer decreases or $\boldsymbol{g}_{\text{avg}}(\boldsymbol{f}, \boldsymbol{x})$ is reached. Simultaneously, we start from $\boldsymbol{g}_{\text{avg}}(\boldsymbol{f}, \boldsymbol{x})$ and iteratively move towards each $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$ in each of the d directions if the complexity decreases. We stop moving when the complexity no longer decreases or any of the $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$ are reached. The final $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ is the location that has the smallest complexity from these $2d$ different walks. Since we only move if the complexity decreases and start from each $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$, the entropy of $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ is guaranteed to be less than or equal to the entropy of all $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$.

5.2 Region Shrinking Method

In our second approach, we consider the closed region, $\mathbf{R}$, which is the convex hull of all the explanation functions, $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$. Notice region $\mathbf{R}$ initially contains $\boldsymbol{g}_{\text{avg}}$. We consider an iterative approach to find the global minimum in the region $\mathbf{R}$. As before, we consider the convex combination formed by two explanation functions, $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$. Using convex optimization, we find the value on the line segment between $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ that has the minimum complexity; essentially, we iteratively shrink the region. For the region shrinking method, the convex combination formed by $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ is:

$$w(\boldsymbol{g}_i) + (1 - w)(\boldsymbol{g}_j), w \in [0, 1]$$

For every pair of functions in $\mathcal{G}_m$, we find the functions that produces the minimum complexity in the convex combination of the functions, producing a new set of candidates $\mathcal{G}'_m$. $\boldsymbol{g}_{\text{agg}}$ is the element in set $\mathcal{G}'_m$ with minimal complexity after K iterations. In each iteration, a function is chosen if it has the minimum complexity of all the functions in a convex combination. Thus, the minimum complexity of the set $\mathcal{G}'_m$ decreases or remains constant with each iteration.

6 Lowering Sensitivity Via Aggregation

To construct an aggregate explanation function $\boldsymbol{g}$ that minimizes sensitivity, we would need to ensure that a test point's explanation is a function of the explanations of its nearest neighbors under ρ. This is a natural analog for how humans reason: we use past similar events (training data) and facts about the present (individual features) to make decisions [Bhatt *et al.*, 2019b]. We now contribute a new explanation function $\boldsymbol{g}_{\text{AVA}}$ that combines the Shapley value explanations of a test point's nearest neighbors to explain the test point.
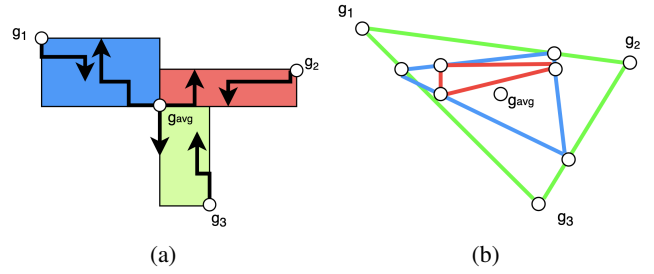

(a)   (b)

Figure 1: Visual examples of the two complexity lowering aggregation algorithms: gradient-descent style (a) and region shrinking (b) methods using explanation functions $\boldsymbol{g}_1, \boldsymbol{g}_2, \boldsymbol{g}_3$

6.1 Shapley Value Review

Borrowing from game theory, Shapley values denote the marginal contributions of a player to the payoff of a coalitional game. Let T be the number of players and let $v : 2^T \to \mathbb{R}$ be the characteristic function, where $v(S)$ denotes the worth (contribution) of the players in $S \subseteq T$. The Shapley value of player i's contribution (averaging player i's marginal contributions to all possible subsets S) is:

$$\phi_i(v) = \frac{1}{|T|} \sum_{S \subseteq T \setminus \{i\}} \binom{T-1}{S}^{-1} (v(S \cup \{i\}) - v(S))$$

Let $\Phi \in \mathbb{R}^T$ be a Shapley value contribution vector for all players in the game, where $\phi_i(v)$ is the ith element of Φ.

6.2 Shapley Values as Explanations

In the feature importance literature, we formulate a similar problem to where the game's payoff is the predictor's output $y = \boldsymbol{f}(\boldsymbol{x})$, the players are the d features of $\boldsymbol{x}$, and the ϕ_i values represent the contribution of x_i to the game $\boldsymbol{f}(\boldsymbol{x})$. Let the characteristic function be the importance score of a subset of features $\boldsymbol{x}_s$, where $\mathbb{E}_Y[\cdot | \boldsymbol{x}]$ is an expectation over $\mathbb{P}_{\boldsymbol{f}}(\cdot | \boldsymbol{x})$:

$$v_{\boldsymbol{x}}(S) = \mathbb{E}_Y \left[-\log \frac{1}{\mathbb{P}_{\boldsymbol{f}}(Y | \boldsymbol{x}_s)} \Big| \boldsymbol{x} \right]$$

This characteristic function denotes the negative of the expected number of bits required to encode the predictor's output based on the features in a subset S [Chen *et al.*, 2019]. Shapley value contributions can be approximated via Monte Carlo sampling [Štrumbelj and Kononenko, 2014] or via weighted least squares [Lundberg and Lee, 2017].

6.3 Aggregate Valuation of Antecedents

We now explore how to explain a test point in terms of the Shapley value explanations of its neighbors. Termed Aggregate Valuation of Antecedents (AVA), we derive an explanation function that explains a data point in terms of the explanations of its neighbors. We do the following: suppose we want to find an explanation function $\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}})$ for a point of interest $\boldsymbol{x}_{\text{test}}$. First we find the k nearest neighbors of $\boldsymbol{x}_{\text{test}}$ under ρ denoted by $\mathcal{N}_k(\boldsymbol{x}_{\text{test}}, \mathcal{D})$.

$$\mathcal{N}_k(\boldsymbol{x}_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N} \subset \mathcal{D}, |\mathcal{N}| = k}{\arg\min} \sum_{\boldsymbol{z} \in \mathcal{N}} \rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})$$

We define $\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}}) = \Phi_{\boldsymbol{x}_{\text{test}}}$ as the explanation function where:

$$\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}})_i = \phi_i(v_{\text{AVA}}) = \sum_{\boldsymbol{z} \in \mathcal{N}_k(\boldsymbol{x}_{\text{test}})} \frac{\boldsymbol{g}_{\text{SHAP}}(\boldsymbol{f}, \boldsymbol{z})_i}{\rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})}$$
$$= \sum_{\boldsymbol{z} \in \mathcal{N}_k(\boldsymbol{x}_{\text{test}})} \frac{\phi_i(v_{\boldsymbol{z}})}{\rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})}$$

In essence, we weight each neighbor's Shapley value contribution by the inverse distance from the neighbor to the test point. AVA is closely related to bootstrap aggregation from classical statistics, as we take an average of model outputs to improve explanation function stability.

Theorem 5. $\boldsymbol{g}_{AVA}(\boldsymbol{f}, \boldsymbol{x}_{test})$ *is a Shapley value explanation.*

Proof. We want to show that $\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}}) = \Phi_{\boldsymbol{x}_{\text{test}}}$ is indeed a vector of Shapley values. Let $\boldsymbol{g}_{\text{SHAP}}(\boldsymbol{f}, \boldsymbol{z}) = \Phi_{\boldsymbol{z}}$ be the vector of Shapley value contributions for a point $\boldsymbol{z} \in \mathcal{N}_k$. By [Lundberg and Lee, 2017], we know $\boldsymbol{g}_{\text{SHAP}}(\boldsymbol{f}, \boldsymbol{z})_i = \phi_i(v_{\boldsymbol{z}})$ is a unique Shapley value for the characteristic function $v_{\boldsymbol{z}}$. By linearity of Shapley values [Shapley, 1953], we know that:

$$\phi_i(v_{\boldsymbol{z}_1} + v_{\boldsymbol{z}_2}) = \phi_i(v_{\boldsymbol{z}_1}) + \phi_i(v_{\boldsymbol{z}_2}) \tag{3}$$

This means that the $\Phi_{\boldsymbol{z}_1} + \Phi_{\boldsymbol{z}_2}$ will yield a unique Shapley value contribution vector for the characteristic function $v_{\boldsymbol{z}_1} + v_{\boldsymbol{z}_2}$. By linearity (or additivity), we know for any scalar α:

$$\alpha \phi_i(v_{\boldsymbol{z}}) = \phi_i(\alpha v_{\boldsymbol{z}}) \tag{4}$$

This means $\alpha \Phi_{\boldsymbol{z}}$ will yield a unique Shapley value contribution vector for the characteristic function $\alpha v_{\boldsymbol{z}}$. Now define:

$$\Phi_{\boldsymbol{x}_{\text{test}}} = \sum_{\boldsymbol{z} \in \mathcal{N}_k(\boldsymbol{x}_{\text{test}})} \frac{\Phi_{\boldsymbol{z}}}{\rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})}$$

We can conclude that $\Phi_{\boldsymbol{x}_{\text{test}}}$ is a vector of Shapley values □

While [Sundararajan *et al.*, 2017] takes a path integral from a fixed reference baseline $\bar{\boldsymbol{x}}$ and [Lundberg and Lee, 2017] only considers attribution along the straight line path between $\bar{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{test}}$, AVA takes a weighted average of attributions along paths from training points in $\mathcal{N}_k$ to $\boldsymbol{x}_{\text{test}}$. AVA can similarly be thought of as a convex combination of explanation functions where the explanation functions are the explanations of the neighbors of $\boldsymbol{x}_{\text{test}}$ and the weights are $\rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})^{-1}$. Though the weights are guaranteed to be non-negative, we normalize the weights to sum to 1 and edit the AVA formulation to be: $\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}}) = \rho_{\text{tot}} \Phi_{\boldsymbol{x}_{\text{test}}}$ where $\rho_{\text{tot}} = \sum_{\boldsymbol{z} \in \mathcal{N}_k(\boldsymbol{x}_{\text{test}})} \rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})^{-1}$. Notice this formulation is a specific convex combination as described before; therefore, AVA will result in a lower sensitivity than $\boldsymbol{g}_{\text{SHAP}}(\boldsymbol{f}, \boldsymbol{x})$ alone.

6.4 Medical Connection

Similar to how a model uses input features to reach an output, medical professionals learn how to proactively search for risk predictors in a patient. Medical professionals not only use patient attributes (e.g., vital signs, personal information) to make a diagnosis but also leverage experiences with past patients; for example, if a doctor treated a rare disease over a decade ago, then that experience can be crucial when attributes alone are uninformative about how to diagnose [Goold and Lipkin Jr, 1999]. This is the analogous to "close" training points affecting a predictor's output. AVA combines the attributions of past training points (past patients) to explain an unseen test point (current patient). When using the MIMIC dataset [Johnson *et al.*, 2016], AVA models the aforementioned intuition.

7 Experiments

We now report some empirical results. We evaluate models trained on the following datasets: Adult, Iris [Dua and Graff, 2017], MIMIC [Johnson *et al.*, 2016], and MNIST [LeCun *et al.*, 1998]. We use the following explanation functions: SHAP [Lundberg and Lee, 2017], Shapley Sampling (SS) [Štrumbelj and Kononenko, 2014], Gradient Saliency (Grad) [Baehrens *et al.*, 2010], Grad*Input (G*I) [Shrikumar *et al.*, 2017], Integrated Gradients (IG) [Sundararajan *et al.*, 2017], and DeepLift (DL) [Shrikumar *et al.*, 2017].

For all tabular datasets, we train a multilayer perceptron (MLP) with leaky-ReLU activation using the ADAM optimizer. For Iris [Dua and Graff, 2017], we train our model to 96% test accuracy. For Adult [Dua and Graff, 2017], our model has 82% test accuracy. As motivated in Section 6.4, we use MIMIC (Medical Information Mart for Intensive Care III) [Johnson *et al.*, 2016]. We extract seventeen real-valued features deemed critical, per [Purushotham *et al.*, 2018], for sepsis prediction. Our model gets 91% test accuracy on the task. For MNIST [LeCun *et al.*, 1998], our model is a convolutional neural network and has 90% test accuracy.

For experiments with a baseline $\bar{\boldsymbol{x}}$, zero baseline implies that we set features to 0 and average baseline uses the average feature value in $\mathcal{D}$. Before doing aggregation, we unit norm all explanations. For the complexity criterion, we take the positive ℓ_1 norm. We set $D = \ell_2$ and $\rho = \ell_\infty$.

7.1 Faithfulness $\mu_{\mathbf{F}}$

In Table 2, we report results for faithfulness for various explanation functions. When evaluating, we take the average of multiple runs where, in each run, we see at least 50 datapoints; for each datapoint, we randomly select $|S|$ features and replace them with baseline values. We then calculate the Pearson's correlation coefficient between the predicted logits of each modified test point and the average explanation attribution for only the subset of features. We notice that, as subset size increases, faithfulness increases until the subset is large enough to contain all informative features. We find that Shapley values, approximated with weighted least squares, are the most faithful explanation function for smaller datasets.

7.2 Max and Avg Sensitivity $\mu_{\mathbf{M}}$ and $\mu_{\mathbf{A}}$

In Table 3, we report the max and average sensitivities for various explanation functions. To evaluate the sensitivity criterion, we sample a set of test points from $\mathcal{D}$ and an additional larger set of training points. We then find the training points that fall within a radius r neighborhood of each test point and

INPUT	BEST (DEEPLIFT)	CONVEX	GRADIENT-DESCENT	REGION-SHRINKING
	$\mu_C = 3.688$	$\mu_C = 3.685$	$\mu_C = 3.575$	$\mu_C = 3.208$

Table 1: Qualitative example of aggregation to lower complexity (μ_C): We show that it is possible to lower complexity slightly with both of our approaches; note that achieving lowest complexity on an image would imply that all attribution is placed on a single pixel.

METHOD SUBSET	ADULT 2	IRIS 2	MIMIC 10	MIMIC 20
SHAP	(**62**, **60**)	(**67**, **68**)	(31, 36)	(37, 47)
SS	(46, 27)	(32, 36)	(59, **58**)	(38, 45)
GRAD	(30, 53)	(14, 16)	(37, 41)	(28, **63**)
G*I	(38, 39)	(27, 30)	(54, 48)	(59, 43)
IG	(47, 33)	(60, 57)	(66, 51)	(**68**, 51)
DL	(58, 43)	(46, 48)	(**84**, 54)	(43, 45)

Table 2: Faithfulness μ_F averaged over a test set: (Zero Baseline, Training Average Baseline). Exact quantities can be obtained by dividing table entries by 10^2

METHOD RADIUS	ADULT 2	IRIS 0.2	MIMIC 4
SHAP	(60, 54)	(310, 287)	(**6**, **5**)
SS	(191, 168)	(477 , 345)	(83, 81)
GRAD	(60, 50)	(**68**, **66**)	(28, 28)
G*I	(86, 71)	(298, 279)	(77, 50)
IG	(**19**, **17**)	(495, 462)	(19, 15)
DL	(74, 74)	(850, 820)	(135, 111)

Table 3: Sensitivity: (Max μ_M, Avg μ_A). Exact quantities can be obtained by dividing table entries by 10^3

METHOD	ADULT	IRIS	MIMIC
$\mu_A(f, g_{\text{SHAP}})$	0.16 ± 0.11	0.22 ± 0.25	0.47 ± 0.12
$\mu_A(f, g_{\text{AVA}})$	$\mathbf{0.07 \pm 0.07}$	$\mathbf{0.13 \pm 0.18}$	$\mathbf{0.31 \pm 0.13}$
$\mu_M(f, g_{\text{SHAP}})$	0.68 ± 0.13	1.20 ± 0.36	0.83 ± 0.17
$\mu_M(f, g_{\text{AVA}})$	$\mathbf{0.52 \pm 0.11}$	$\mathbf{1.18 \pm 0.28}$	$\mathbf{0.72 \pm 0.22}$
$\mu_C(f, g_{\text{SHAP}})$	1.94 ± 0.26	1.36 ± 0.36	$\mathbf{2.33 \pm 0.23}$
$\mu_C(f, g_{\text{AVA}})$	$\mathbf{1.93 \pm 0.24}$	$\mathbf{1.24 \pm 0.32}$	2.61 ± 0.29

Table 4: AVA lowers the sensitivity of Shapley value explanations across all datasets. When d is small (fewer features), AVA explanations are slightly less complex.

in all experiments but also have less complex explanations (depending on the radius r and number of features d). After finding the average distance between pairs of points, we use $r = 1$ for Adult, $r = 0.3$ for Iris, and $r = 10$ for MIMIC.

8 Conclusion

Borrowing from earlier work in social science and the philosophy of science, we codify low sensitivity, high faithfulness, and low complexity as three desirable properties of explanation functions. We define these three properties for feature-based explanation functions, develop an aggregation scheme for learning combinations of various explanation functions, and devise schemes to learn explanations with lower complexity (iterative approaches) and lower sensitivity (AVA). We hope that this work will provide practitioners with a principled way to evaluate feature-based explanations and to learn an explanation which aggregates and optimizes for criteria desired by end users. Though we consider one criterion at a time, future work could further axiomatize our criteria, explore the interaction between different evaluation criteria, and devise a multi-objective optimization approach to finding a desirable explanation; for example, can we develop a procedure for learning a less sensitive and less complex explanation function simultaneously?

Acknowledgements

find the distance between each nearby training point explanation and the test point explanation to get a mean and max. We average over ten random runs of this procedure. Sensitivity is highly dependent on the dimensionality d and on the radius r. We find that as sensitivity decreases as r increases. Empirically, for MIMIC, Shapley values approximated by weighted least squares (SHAP) are the least sensitive.

7.3 MNIST Complexity μ_C

In Table 1, we provide a qualitative example for the gradient descent-style and region-shrinking methods for lowering complexity of explanations from a model trained on MNIST. We show an example with images since it illustrates the notion of lower complexity well; however, other data types (tabular) might be better suited for complexity optimization.

7.4 AVA

Our empirical findings support use of an AVA explanation if low sensitivity is desired. [Ghorbani *et al.*, 2019] note that perturbation-based explanations (like g_{SHAP}) are less sensitive than their gradient-based counterparts. In Table 4, we show that AVA explanations not only have lower sensitivities

References

[Baehrens *et al.*, 2010] David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Muller. How to explain individual classification decisions. *JMLR*, 11(Jun):1803–1831, 2010.

[Batterman and Rice, 2014] Robert W Batterman and Collin C Rice. Minimal model explanations. *Philosophy of Science*, 81(3):349–376, 2014.

[Berger, 2013] James O Berger. *Statistical decision theory and Bayesian analysis*. Springer Science & Business Media, 2013.

[Bhatt *et al.*, 2019a] Umang Bhatt, Pradeep Ravikumar, et al. Building human-machine trust via interpretability. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9919–9920, 2019.

[Bhatt *et al.*, 2019b] Umang Bhatt, Pradeep Ravikumar, and José M. F. Moura. Towards aggregating weighted feature attributions. *arXiv:1901.10040*, 2019.

[Bhatt *et al.*, 2020] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José M. F. Moura, and Peter Eckersley. Explainable machine learning in deployment. *ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, 2020.

[Camburu *et al.*, 2019] Oana-Maria Camburu, Eleonora Giunchiglia, Jakob Foerster, Thomas Lukasiewicz, and Phil Blunsom. Can I trust the explainer? Verifying post-hoc explanatory methods. *arXiv:1910.02065*, 2019.

[Chang *et al.*, 2019] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and David Duvenaud. Explaining image classifiers by counterfactual generation. In *International Conference on Learning Representations*, 2019.

[Chen *et al.*, 2019] Jianbo Chen, Le Song, Martin J Wainwright, and Michael I Jordan. L-shapley and C-shapley: Efficient model interpretation for structured data. *International Conference on Learning Representations*, 2019.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of neural networks is fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3681–3688, 2019.

[Goold and Lipkin Jr, 1999] Susan Dorr Goold and Mack Lipkin Jr. The doctor–patient relationship: challenges, opportunities, and strategies. *Journal of general internal medicine*, 14(Suppl 1):S26, 1999.

[Hooker *et al.*, 2019] Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, pages 9734–9745, 2019.

[Johnson *et al.*, 2016] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 2016.

[Lage *et al.*, 2019] Isaac Lage, Emily Chen, Jeffrey He, Menaka Narayanan, Been Kim, Samuel J Gershman, and Finale Doshi-Velez. Human evaluation of models built for interpretability. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 59–67, 2019.

[LeCun *et al.*, 1998] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[Lipton, 2003] Peter Lipton. *Inference to the best explanation*. Routledge, 2003.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, pages 4765–4774, 2017.

[Melis and Jaakkola, 2018] David Alvarez Melis and Tommi Jaakkola. Towards robust interpretability with self-explaining neural networks. In *Advances in Neural Information Processing Systems (NeurIPS 2018)*, 2018.

[Purushotham *et al.*, 2018] Sanjay Purushotham, Chuizheng Meng, Zhengping Che, and Yan Liu. Benchmarking deep learning models on large healthcare datasets. *Journal of Biomedical Informatics*, 83:112–134, 2018.

[Ruben, 2015] David-Hillel Ruben. *Explaining explanation*. Routledge, 2015.

[Shapley, 1953] Lloyd S Shapley. A value for n-person games. In *Contributions to the Theory of Games II*, pages 307–317, 1953.

[Shrikumar *et al.*, 2017] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70 (ICML 2017)*, pages 3145–3153. Journal of Machine Learning Research, 2017.

[Štrumbelj and Kononenko, 2014] Erik Štrumbelj and Igor Kononenko. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems*, 41(3):647–665, 2014.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70 (ICML 2017)*, pages 3319–3328. Journal of Machine Learning Research, 2017.

[Yang and Kim, 2019] Mengjiao Yang and Been Kim. BIM: Towards quantitative evaluation of interpretability methods with ground truth. *arXiv:1907.09701*, 2019.

[Yeh *et al.*, 2019] Chih-Kuan Yeh, Cheng-Yu Hsieh, Arun Suggala, David I Inouye, and Pradeep K Ravikumar. On the (in) fidelity and sensitivity of explanations. In *Advances in Neural Information Processing Systems*, pages 10965–10976, 2019.