Evaluating and Aggregating Feature-based Model Explanations

Umang Bhatt[1,2], Adrian Weller[2,3], and José M.F. Moura[1]
[1] *Carnegie Mellon University,* [2] *University of Cambridge,* [3] *The Alan Turing Institute*
usb20@cam.ac.uk

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.

Evaluation Criteria

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. 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 $\boldsymbol{f}$, $\boldsymbol{g}$, and $\boldsymbol{x}$, and outputs a scalar: $\mu(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x})$.

Desideratum: Low Sensitivity. If inputs are near each other and their model outputs are similar, then their explanations should be close to each other. 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 $\boldsymbol{x}$.

$$\text{Max Sensitivity: } \mu_{\text{M}}(\boldsymbol{f}, \boldsymbol{g}, r; \boldsymbol{x}) = \max_{\boldsymbol{z} \in \mathcal{N}_r} D(\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}), \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{z}))$$
$$\text{Average Sensitivity: } \mu_{\text{A}}(\boldsymbol{f}, \boldsymbol{g}, r; \boldsymbol{x}) = \int_{\boldsymbol{z} \in \mathcal{N}_r} D(\boldsymbol{g}(\boldsymbol{f}, \boldsymbol{x}), \boldsymbol{g}(\boldsymbol{f}, \boldsymbol{z})) \mathbb{P}_{\boldsymbol{x}}(\boldsymbol{z}) d\boldsymbol{z}$$

Desideratum: High Faithfulness. 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 $S \subseteq \{1, 2, \dots d\}$, $\boldsymbol{x}_s = \{x_i, i \in S\}$ 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 $|S| = d$, $\boldsymbol{x}_{[\boldsymbol{x}_s = \bar{\boldsymbol{x}}_s]} = \bar{\boldsymbol{x}}$.

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

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:

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

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

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

Other Aggregation Methods

Given $\boldsymbol{f}$, $\mathcal{G}_m = \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_m\}$, μ, and a set of inputs $\mathcal{D}_{\boldsymbol{x}}$, we want to find an aggregate explanation function $\boldsymbol{g}_{\text{agg}}$ that satisfies μ at least as well as any $\boldsymbol{g}_i \in \mathcal{G}_m$. Let $h(\cdot)$ represent some function that combines m explanations into a consensus $\boldsymbol{g}_{\text{agg}} = h(\mathcal{G}_m)$.

Convex Combination. Suppose we have two different explanation functions $\boldsymbol{g}_1$ and $\boldsymbol{g}_2$ and have chosen a criterion μ to evaluate a $\boldsymbol{g}$. Consider an aggregate explanation, $\boldsymbol{g}_{\text{agg}} = h(\boldsymbol{g}_1, \boldsymbol{g}_2)$. A potential $h(\cdot)$ is a convex combination where $\boldsymbol{g}_{\text{agg}} = h(\boldsymbol{g}_1, \boldsymbol{g}_2) = w\boldsymbol{g}_1 + (1 - w)\boldsymbol{g}_2 = \boldsymbol{w}^\intercal \mathcal{G}_m$.

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:

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

where p is a positive constant. Assuming real-valued attributions where $\mathcal{G} \subseteq \mathbb{R}^d$, when D is ℓ_2 and $p = 2$, the aggregate explanation is the feature-wise sample mean; when D is ℓ_1 and $p = 1$, the aggregate is the feature-wise sample median [1]. 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 is the Kemeny-Young rule.

Future Work. We could leverage multi-objective optimization to find $\boldsymbol{g}_{\text{agg}}$. For example, we could learn a less sensitive and less complex explanation function by optimizing:

$$\boldsymbol{g}_{\text{agg}} \in \underset{\boldsymbol{g} \in \mathcal{G}}{\arg\min} \underset{\boldsymbol{x} \in \mathcal{D}}{\mathbb{E}} \left[\mu_{\text{A}}(\boldsymbol{f}, \boldsymbol{g}, r; \boldsymbol{x}) + \lambda \mu_{\text{C}}(\boldsymbol{f}, \boldsymbol{g}; \boldsymbol{x}) \right]$$

Algorithms for Aggregation

We can aggregate explanations from various $\boldsymbol{g}_i$ to optimize a particular evaluation criterion.

Lowering Sensitivity. Termed Aggregate Valuation of Antecedents (AVA), we derive an explanation function that explains a data point in terms of the explanations of its neighbors. To obtain an explanation $\boldsymbol{g}_{\text{AVA}}(\boldsymbol{f}, \boldsymbol{x}_{\text{test}})$ for a point of interest $\boldsymbol{x}_{\text{test}}$, we first find the k nearest neighbors of $\boldsymbol{x}_{\text{test}}$ under ρ denoted by $\mathcal{N}_k(\boldsymbol{x}_{\text{test}}, \mathcal{D})$.

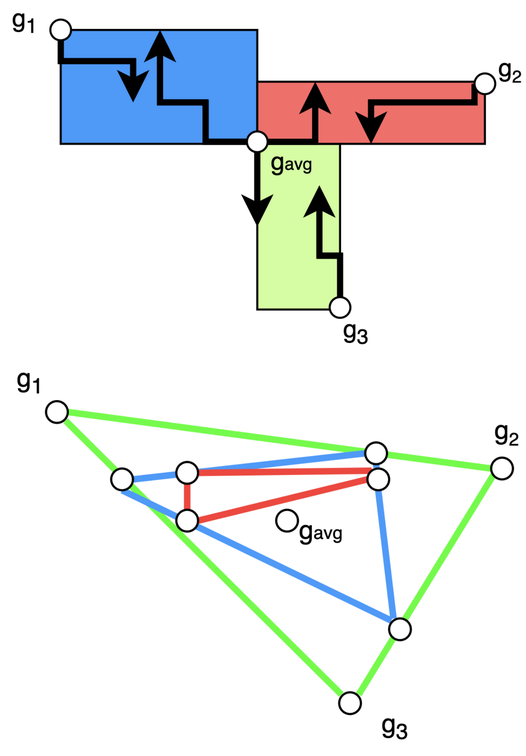
$$\mathcal{N}_k(\boldsymbol{x}_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N} \subset \mathcal{D}, |\mathcal{N}| = k}{\arg\min} \sum_{\boldsymbol{z} \in \mathcal{N}} \rho(\boldsymbol{x}_{\text{test}}, \boldsymbol{z})$$

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

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

Theorem 1. *$\boldsymbol{g}_{AVA}(\boldsymbol{f}, \boldsymbol{x}_{test})$ is a Shapley value explanation.*

Lowering Complexity. We devise iterative algorithms for aggregating explanation functions to obtain $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ with lower complexity whilst combining m candidate explanation functions $\mathcal{G}_m = \{\boldsymbol{g}_1, \dots, \boldsymbol{g}_m\}$. We desire a $\boldsymbol{g}_{\text{agg}}(\boldsymbol{f}, \boldsymbol{x})$ that contains information from all candidate explanations $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$ yet has entropy less than or equal to that of each explanation $\boldsymbol{g}_i(\boldsymbol{f}, \boldsymbol{x})$.



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

References

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