Building Human-Machine Trust via Interpretability

Umang Bhatt, Pradeep Ravikumar, and José Moura
Carnegie Mellon University
umang@cmu.edu

Electrical & Computer ENGINEERING

Introduction

Developing human-machine trust is a prerequisite for adoption of machine learning systems in decision critical settings (e.g healthcare and governance). Users develop appropriate trust in these systems when they understand how the systems make their decisions. Interpretability not only helps users understand what a system learns but also helps users contest that system to align with their intuition. We propose an algorithm, AVA: Aggregate Valuation of Antecedents, that generates a consensus feature attribution, retrieving local explanations and capturing global patterns learned by a model. Our empirical results show that AVA rivals current benchmarks.

Feature-based Explanations

A feature-based explanation (ε), usually in the form of a feature attribution, saliency map, etc., denotes how much a feature, x_i, contributes to a trained model's output, $f(x)$.

There are two main approaches for feature-based model interpretability: *gradient based attribution techniques* and *perturbation based attribution techniques*. An example of a gradient-based attribution is Integrated Gradients, which accumulate gradients along a straight line path between x and $\bar{x}$ [1]:

$$\varepsilon_i = (x_i - \bar{x_i}) \int_{\alpha=0}^{1} \frac{\partial f(\bar{x} + \alpha(x - \bar{x}))}{\partial x_i} d\alpha$$

An example of a perturbation-based attribution is SHapley Additive exPlanations where $T = S \cup \{i\}$ and $R = (\frac{|S|!(|F|-|S|-1)!}{|F|!})$ [2]:

$$\varepsilon_i = \sum_{S \subseteq F \setminus \{i\}} R\left(f_T(x_T) - f_S(x_S)\right)$$

Current feature attribution approaches are impoverished, resulting in inconsistent attributions due to noisy gradients estimates or unrepresentative regions of perturbation: both of which decrease user trust.

Approach

We aim to explain which features in x_{test} contributed a prediction from model f locally whilst understanding global trends of f in the neighborhood of x_{test}. We propose the following stage-wise procedure, AVA: Aggregate Valuation of Antecedents.

1. Find the influence weight, ρ_j, of every training point, $x^{(j)}$, on a test point, x_{test}, using the approximation proposed in [3].

$$\rho_j = \frac{d}{d\epsilon} \mathcal{L}(f_{\epsilon, x^{(j)}}, x_{\text{test}})\big|_{\epsilon=0}$$

2. Select the local neighborhood, $\mathcal{N}_k$, of the k most influential training points on x_{test}.

$$\mathcal{N}_k(x_{\text{test}}, \mathcal{D}) = \underset{\mathcal{N} \subset \mathcal{D}, |\mathcal{N}|=k}{\arg\max} \sum_{x^{(j)} \in \mathcal{N}} \rho_j$$

3. Pick an explanation function, g, like SHAP or Integrated Gradients; then, find a consensus explanation of all k explanations using some aggregation mechanism, $\mathcal{A}$, based on centroids with respect to some distance $d : \mathcal{E} \times \mathcal{E} \mapsto R$.

$$\varepsilon_A = \mathcal{A}(g, \mathcal{N}_k) \in \arg\min_{\varepsilon \in \mathcal{E}} \sum_{x \in N_k} d(\varepsilon, g(x))$$

Aggregation Scheme

The simplest examples of distance, d, include: (a) ℓ_2 distance with real-valued attributions where $\mathcal{E} = R^d$, and (b) the Kendall-tau distance with rank-valued attributions where $\mathcal{E} = \mathcal{S}_d$, the set of permutations over d features. For rank valued attributions, any aggregation mechanism falls under rank aggregation from social choice, for which many practical "voting rules" exist.

- **Borda Count**: Gives weight to each position in a rank. The feature with the largest sum across all ranks is the most important in aggregate.
- **Markov Chains**: Uses markov chains to combine pair-wise rank comparisons.

Counterfactual Intuition

AVA gives rise to an intriguing counterfactual intuition as follows: what training point perturbations which when used to train the predictor would influence the test prediction the most? Suppose we perturb a training point as $z_\delta = z + \delta$. We denote the predictor obtained by upweighting substituting the training data point z by z_δ and moreover upweighting this by some constant ϵ as $f_{\epsilon, z_\delta, -z}$. Then the counterfactual mentioned above at a test point z_{test}, would compute:

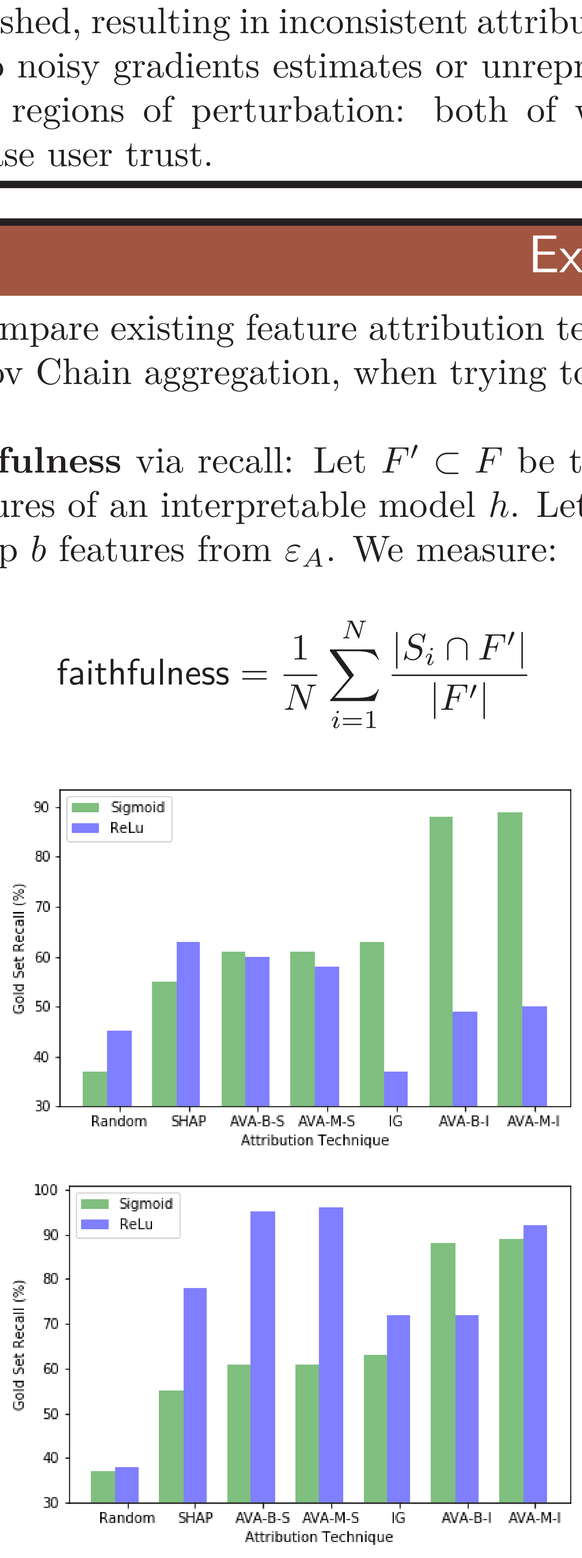
$$\nabla_\delta \frac{d}{d\epsilon} \mathcal{L}(f_{\epsilon, z_\delta, -z}, z_{\text{test}})\big|_{\epsilon=0}\big|_{\delta=0}.$$

Experimentation

We compare existing feature attribution techniques to AVA with Borda Count aggregation and with Markov Chain aggregation, when trying to explain the output of a multilayer perceptron.

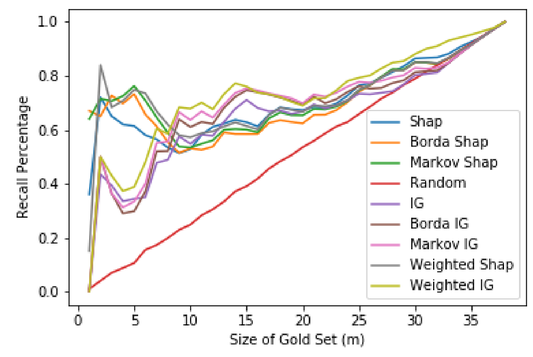
Faithfulness via recall: Let $F' \subset F$ be the top b features of an interpretable model h. Let S_i be the top b features from ε_A. We measure:

$$\text{faithfulness} = \frac{1}{N} \sum_{i=1}^{N} \frac{|S_i \cap F'|}{|F'|}$$



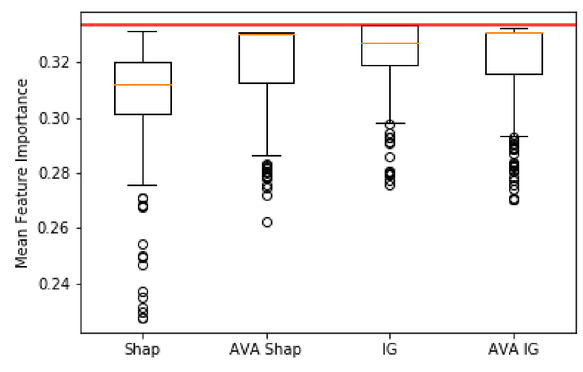
Top: Adult Dataset; Bottom: Titanic Dataset

K-sensitivity: We ask how the size of $\mathcal{N}_k$ affects faithfulness of ε_A.



Mean Feature Importance: We find the mean feature importance of the top m features in ε_A:

$$\text{mfi} = \frac{1}{m} \sum_{i=1}^{m} \varepsilon_{A\,i}$$



Conclusion

We explore how to combine sample-based and feature-based approaches to simultaneously extract local explanations and detect global patterns. With our stage-wise procedure AVA, we show that such combinations can outperform existing baselines with careful selection of k and the right aggregation mechanism.

References

[1] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 2017.

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

[3] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*, 2017.