Building Human-Machine Trust via Interpretability

Umang Bhatt
Carnegie Mellon University
umang@cmu.edu

Pradeep Ravikumar
Carnegie Mellon University
pradeepr@cs.cmu.edu

José M. F. Moura
Carnegie Mellon University
moura@cmu.edu

Abstract

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.

Introduction

As machine learning systems become pervasive, human-machine trust ought to become a potentially necessary objective. Currently, black-box systems beget powerful predictive power to the end user but come with a burden of opacity, creating space for distrust. Training interpretable models or coupling explainable models with black-box models demystifies the reasoning in these systems whilst maintaining respectable levels of accuracy (Lipton 2018). Such transparent machine learning systems that deliver post-hoc explanations with predictions have been extensively studied in the current machine learning literature.

We can explain a model’s output by looking at the training examples most influential to model prediction for an unseen test point (Koh and Liang 2017). We also can provide associations between input features and the target prediction, resulting in a feature attribution: a ranking of which features mattered most to the model. Feature attributions can be found via gradient-based methods, which find the partial derivative of the target with respect to every input feature (Sundararajan, Taly, and Yan 2017), or perturbation-based methods, which use parametric models to approximate the decision boundary in a region of interest (Lundberg and Lee 2017; Sundararajan, Taly, and Yan 2017) with a local neighborhood influence measure proposed in (Koh and Liang 2017). To first introduce notation, let $x \in \mathbb{R}^d$ be a datapoint’s feature vector where the $x_i \in R$ is a specific feature of that datapoint. Let $D = \{x^{(j)}\}_{j=1}^N$ represent the training datapoints, where $D \in \mathbb{R}^{d \times N}$. Let $\hat{f}$ be the learned predictor we wish to explain. Using the approximation in (Koh and Liang 2017), we define the influence weight, $\rho_j \in \mathbb{R}_{\geq 0}$, of training point, $x^{(j)}$, on a test point, $x_{test}$, as follows.

$$\rho_j = \mathcal{I}_{\text{op. loss}}(x^{(j)}, x_{test}) = \frac{d}{dx} \mathcal{L}(\hat{f}_{\epsilon,x^{(j)}}, x_{test})|_{\epsilon=0}$$

We then select the local neighborhood, $N_k$, of the $k$ most influential training points on $x_{test}$.

$$N_k(x_{test}, D) = \arg \max_{N \subseteq D, |N|=k} \sum_{x^{(j)} \in N} \rho_j$$

Using an attribution technique $g$, like (Lundberg and Lee 2017) or (Sundararajan, Taly, and Yan 2017), we obtain a value attribution for each of the $k$ points. Finally, once we have the set of value attributions $\{g(x^{(j)})\}_{x^{(j)} \in N_k} \in \mathbb{G}^k$, where $g(x^{(j)}) \in \mathbb{G}$, we can apply an aggregation scheme $A : \mathbb{G}^k \rightarrow \mathbb{G}$ to obtain a consensus feature attribution. The procedure is outlined in Algorithm 1.

Traditional Rank Aggregation

We leverage traditional aggregation techniques (i.e., Borda Count and Markov Chains) to combine the top $k$ attributions into a consensus attribution. A natural class of such aggregation mechanisms are based on centroids with respect to some distance $d : \mathbb{G} \times \mathbb{G} \rightarrow \mathbb{R}$, so that:

$$A(\{g(x)\}_{x \in N_k}) = \arg \min_{g \in \mathbb{G}} \sum_{j=1}^k d(g, g^j)$$
AVA aggregated with Borda Count as AVA-B and AVA aggregated with Markov Chains as AVA-M; the third letter S or I denotes which attribution technique was used SHAP or Integrated Gradients, respectively. In Figure 1(a), we report gold set recall for the Adult dataset over different attribution schemes to explain the same three layered MLP with the stated activation function trained with ADAM and do the same for the Titanic dataset in Figure 1(b). Evidently, AVA outperforms current benchmarks.

![Figure 1: Gold set recall with traditional rank aggregation schemes: (a) Adult; (b) Titanic](image)

**Conclusion**

We introduced AVA, Aggregate Valuation of Antecedents, as a new feature attribution technique. By calculating the top $k$ influences for a given test point, we aggregate those influences’ feature attributions to find a consensus feature attribution. We have shown that AVA’s consensus attribution outperforms current attribution benchmarks on tabular datasets. In future work, we hope to realize a medical use case of AVA, develop a more robust aggregation step that builds on counterfactual intuition, and adapt AVA for unstructured domains (i.e., images and natural language): all of which will continue to build human-machine trust via interpretability.

**References**


