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). 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.

Aggregate Valuation of Antecedents

To maintain human-machine trust, we develop a new class of explanations that aggregates feature attributions of the most influential points to a given test point, exposing local explanations and global patterns simultaneously. Our proposed method, AVA: Aggregate Valuation of Antecedents, combines the idea of approximating a black-box model of interest to develop a feature attribution for a test point via (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 $x \in \mathbf{R}^d$ be a datapoint's feature vector where the $x_i \in \mathbf{R}$ is a specific feature of that datapoint. Let $\mathcal{D} = \{x^{(j)}\}_{j=1}^N$ represent the training datapoints, where $\mathcal{D} \in \mathbf{R}^{d \times N}$. Let $\widehat{f}$ be the learned predictor we wish to explain. Using the approximation in (Koh and Liang 2017), we define the influence weight, $\rho_j \in \mathbf{R}_{\geq 0}$, of training point, $x^{(j)}$, on a test point, x_{test}, as follows.

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

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

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

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 $\{g(x^{(j)})\}_{x^{(j)} \in \mathcal{N}_k} \in \mathcal{G}^*$, where $g(x^{(j)}) \in \mathcal{G}$, we can apply an aggregation scheme $\mathcal{A} : \mathcal{G}^* \mapsto \mathcal{G}$ 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 $d : \mathcal{G} \times \mathcal{G} \mapsto \mathbf{R}$, so that:

$$\mathcal{A}\left(\{g(x)\}_{x \in \mathcal{N}_k}\right) \in \arg\min_{g \in \mathcal{G}} \sum_{j=1}^{k} d(g, g^j)$$

Algorithm 1 AVA for a single test point, x_{test}

Input: test point x_{test}, training data $\mathcal{D}$, learnt predictor $\widehat{f}$, feature attribution technique g, aggregation technique $\mathcal{A}$
Find the top k most influential training points w.r.t. $\widehat{f}$ using influence functions: $\mathcal{N}_k(x_{\text{test}}, \mathcal{D})$
for data point $x \in \mathcal{N}_k$ **do**
  Compute the feature attribution $g(x)$ of a point x
end for
Output: Consensus attribution using $\mathcal{A}$: $\mathcal{A}\left(\{g(x)\}_{x \in \mathcal{N}_k}\right)$

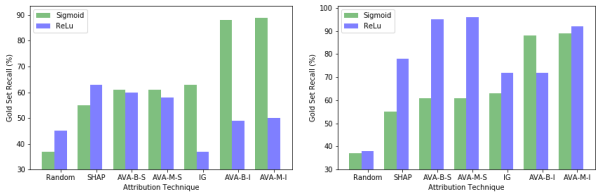
The simplest examples of distances include: (a) ℓ_2 distance with real-valued attributions where $\mathcal{G} = \mathbf{R}^d$, and (b) the Kendall-tau distance with rank-valued attributions where $\mathcal{G} = \mathcal{S}_d$, the set of permutations over d elements (in this case, features); the resulting aggregation mechanism via computing the centroid in this case is called the Kemeny-Young rule. We could obtain such rank-valued attributions by taking any quantitative vector-valued attributions, ranking the features according to these values, and thus obtaining a rank-valued attribution. For such rank valued attributions, any aggregation mechanism falls under the rank aggregation problem in social choice, for which many practical "voting rules" exist. In fact, the aforementioned Kemeny-Young rule is computationally intractable with $O(n!)$ complexity due to solving an optimization problem over the set of permutations over n elements. Accordingly, we leverage other rank aggregation schemes that are more computationally practical.

- **Borda Count** (Narodytska and Walsh 2014): This technique gives a weight to each position in the rank. The feature with the largest sum across all ranks is the most important in the aggregate rank.
- **Markov Chains** (Negahban, Oh, and Shah 2012): This technique uses Markov Chains to consolidate pair-wise comparisons.

Experiments

We present experiments to evaluate the consensus attribution given by AVA on tabular datasets. To explain an individual prediction via value attribution, we compare AVA with the attribution given by the feature attribution technique itself (SHAP or Integrated Gradients). We can quantify the faithfulness of a feature attribution through its recall on a *gold set* of m important features obtained from an interpretable model like in (Ribeiro, Singh, and Guestrin 2016). To obtain a *gold set*, we use a decision tree classifier that we prune to a maximum of m features, where m is picked by cross validation for each dataset to maximize accuracy of the known-interpretable classifier. As a sanity check, we also compare against a random procedure that randomly picks m features as an explanation.

We had two degrees of freedom in our experimentation: the explanation technique g and the aggregation technique $\mathcal{A}$. We selected two attribution techniques (SHAP and Integrated Gradients) and two aggregation techniques (Borda Count and Markov Chains) to fold into AVA. We denote 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.



(a) (b)

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

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

Koh, P. W., and Liang, P. 2017. Understanding black-box predictions via influence functions. In *Proceedings of the 34th International Conference on Machine Learning*.

Lipton, Z. C. 2018. The mythos of model interpretability. *Queue* 16(3):30:31–30:57.

Lundberg, S. M., and Lee, S.-I. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*. 4765–4774.

Narodytska, N., and Walsh, T. 2014. The computational impact of partial votes on strategic voting. In *ECAI*.

Negahban, S.; Oh, S.; and Shah, D. 2012. Iterative ranking from pair-wise comparisons. In *Advances in Neural Information Processing Systems 25*. 2474–2482.

Ribeiro, M. T.; Singh, S.; and Guestrin, C. 2016. Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD*, 1135–1144.

Sundararajan, M.; Taly, A.; and Yan, Q. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70.