Building Human-Machine Trust via Interpretability

Umang Bhatt
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
umang@cmu.edu

Abstract

Developing human-machine trust is a prerequisite for adoption of AI systems in decision critical settings (e.g healthcare, finance, and governance). People develop appropriate trust in an AI system when they understand how a system makes its decisions. My research focuses on model interpretability. I aim to develop techniques for extracting explanations and explaining predictions from complex AI systems. Such interpretability not only helps a user understand what a system learns but also helps a user audit that system to align with their intuition.

Introduction

Human-human trust has been extensively studied in the fields of psychology, philosophy, and management (Rousseau et al. 1998; Lewicki, Tomlinson, and Gillespie 2006; Roff and Danks 2018). This trust can be defined as a "confident relationship with the unknown" (Botsman 2017). As AI 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. Training interpretable models can demystify the reasoning in these systems whilst maintaining respectable levels of accuracy (Lipton 2016; Bhatt 2018). People can develop appropriate trust in AI systems when these systems prioritize transparency.

Transparent AI systems that deliver explanations to their predictions have been extensively studied in the current machine learning literature. Bespoke model architectures can concurrently generate real time explanations and predictions. The more common explanation technique looks to generate a post-hoc explanation given an already learned black-box model. We can explain a model's output by looking at the training examples most influential to the instance being classified (Koh and Liang 2017). On the other hand, we also can provide associations between input variables and the target prediction, resulting in a feature attribution: a ranking of which features mattered most to the target prediction. Feature attributions can be found via gradient-based methods, which find the partial derivative of the output with respect to every input feature (Sundararajan, Taly, and Yan 2017; Ancona et al. 2018), or perturbation-based methods, which use simple parametric models to approximate the decision boundary in the region of interest (Ribeiro, Singh, and Guestrin 2016; Lundberg and Lee 2017). Current feature attribution approaches are impoverished, providing inconsistent attributions due to noisy gradients estimates or unrepresentative regions of perturbation: both of which decreases a user's trust. My research looks to expose the consensus attribution a model uses to discriminate and predict on an unseen test point.

Aggregate Valuation of Antecedents

Collaborating with my advisor José Moura (Electrical and Computer Engineering Department at CMU) and Pradeep Ravikumar (Machine Learning Department at CMU), I am interested in developing a new class of explanations that aggregates feature attributions of the most influential points to a given test point. Since a model is pretrained on a given dataset, a new test point's attribution (and thus class) depends on the attribution and class of the k points "nearest" to the newly seen point in the feature space. I am working on stage-wise procedure, called Aggregate Valuation of Antecedents (AVA), that results in a consensus feature attribution that describes which features a model used.

1. Find the top k influences for a given test point using a technique like k-nearest neighbors or influence functions
2. Find the feature attributions for each influential point using a gradient-based or perturbation-based approach
3. Aggregate the k feature attributions into a consensus feature attribution

The aggregation mechanism would depend on the type of feature attributions we use. If we are provided with some distance metric over the domain of attribution vectors, then we can simply compute the centroid with respect to the distance. For example, for real-valued attribution vectors and ℓ_2 distance, the aggregation would be a simple average. Or when the attribution vectors consists of feature rankings (which we could compute even if simply provided real-valued attribution vectors) and we use the Kendall-tau distance, the aggregation rule would consist of the corresponding centroid, which is also called Kemeny-Young Rule in rank aggregation. In the case of feature rank based value attributions, we could also leverage other rank aggregation

schemes from social choice to provide a consensus value attribution. The overall meta-algorithm is outlined in Algorithm 1.

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

Input: test point X_{test}, model f, feature attribution technique g, aggregation technique $\mathcal{A}$
Find the top k most influential training w.r.t f using influence functions: $I_{k,X_{\text{test}}} \subseteq \{X_1, \ldots, X_N\}$
for data point $x \in I_{k,X_{\text{test}}}$ **do**
  Compute the feature attribution $g(x)$ of data point x
end for
Output: Consensus attribution $\mathcal{A}$: $\mathcal{A}\left(\{g(x)\}_{x \in I_{k,X_{\text{test}}}}\right)$

Future Plans

AVA looks to better capture the decision boundary used by a machine learning model to increase human-machine trust in the system, since we provide the actual correlation learned by the model via the consensus feature attribution. In the coming months, my plan is to ensure AVA and existing feature attribution techniques align with an information-theoretic approach to interpretability (Shwartz-Ziv and Tishby 2017) and to explore and extend how well these attribution techniques align with human concepts and people's subjective priors (Kim et al. 2017; Lage et al. 2018). Moreover, I hope to develop a metric of trust that can accompany any model to convey to user's how well the model represents the process they want automated (Jiang, Kim, and Gupta 2018). Thoughtful interpretability can build human-machine trust and ignite the adoption of AI systems.

Previous Research

My past work was two fold. Under the supervision of Zico Kolter (Computer Science Department in SCS at CMU), my oldest research focused on developing distributed machine learning techniques for detecting potholes and assessing road conditions via the accelerometer and gyroscope in smartphones across a city: a case study of how people trust governance actioning based on personal data from citizens (Bhatt et al. 2017). Furthermore, my research has also focused on human-machine trust in a game theoretic setting, in collaboration with Aaron Roth (Robotics Institute at CMU), Tamara Amin (Psychology at CMU), Fei Fang (Institute of Software Research at CMU), Afsaneh Doyrab (Institute of Software Research at CMU), and Manuela Veloso (Machine Learning Department at CMU). We were interested in understanding the effect of humanoid affect expression on humans. We found that, when a humanoid and human were playing a competitive game, a humanoid expressing negative affect (via demeaning language) led humans to use a more optimal strategy: that is, their quantal response increases, which suggests more rational decision-making (Roth et al. 2018). Extending previous work done in a cooperative setting with an affect-aware architecture (Scheutz, Schermerhorn, and Kramer 2006), we explored the limit of human-machine trust by using affect expression to sway human behavior.

References

Ancona, M.; Ceolini, E.; Oztireli, C.; and Gross, M. 2018. Towards better understanding of gradient-based attribution methods for deep neural networks. In *International Conference on Learning Representations*.

Bhatt, U.; Mani, S.; Xi, E.; and Kolter, J. Z. 2017. Intelligent pothole detection and road condition assessment. *Bloomberg Data for Good Exchange* abs/1710.02595.

Bhatt, U. 2018. Maintaining the humanity of our models. In *AAAI Spring Symposium Series*.

Botsman, R. 2017. *Who Can You Trust?: How Technology Brought Us Together and Why It Might Drive Us Apart*.

Jiang, H.; Kim, B.; and Gupta, M. 2018. To trust or not to trust a classifier.

Kim, B.; Wattenberg, M.; Gilmer, J.; Cai, C.; Wexler, J.; Viegas, F.; and Sayres, R. 2017. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav).

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

Lage, I.; Ross, A. S.; Kim, B.; Gershman, S. J.; and Doshi-Velez, F. 2018. Human-in-the-loop interpretability prior.

Lewicki, R. J.; Tomlinson, E. C.; and Gillespie, N. 2006. Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. *Journal of Management* 32(6):991–1022.

Lipton, Z. C. 2016. The mythos of model interpretability. *CoRR* abs/1606.03490.

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.

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.

Roff, H. M., and Danks, D. 2018. "trust but verify": The difficulty of trusting autonomous weapons systems. *Journal of Military Ethics* 17(1):2–20.

Roth, A. M.; Bhatt, U.; Amin, T.; Doryab, A.; Fang, F.; and Veloso, M. M. 2018. The impact of humanoid affect expression on human behavior in a game-theoretic setting. *1st Workshop on Humanizing AI (HAI) at IJCAI'18*.

Rousseau, D. M.; Sitkin, S. B.; Burt, R. S.; and Camerer, C. 1998. Introduction to special topic forum: Not so different after all: A cross-discipline view of trust. *The Academy of Management Review* 23(3):393–404.

Scheutz, M.; Schermerhorn, P.; and Kramer, J. 2006. The utility of affect expression in natural language interactions in joint human-robot tasks. *Human-Robot Interaction '06*.

Shwartz-Ziv, R., and Tishby, N. 2017. Opening the black box of deep neural networks via information. *CoRR* abs/1703.00810.

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.