Interpretable Modeling in Machine Learning

Cynthia Rudin, Massachusetts Institute of Technology

Arguably, the main stumbling block in getting machine learning algorithms used in medical and industrial applications is the fact that people do not trust them. One of the key reasons people do not trust them is because the models are not sufficiently transparent or interpretable. This lecture will be about interpretable machine learning with logic, lists and trees.

A machine learning model typically provides predictions, but the explanations for what it predicts arise out of a complicated formula that is difficult for a human to comprehend. Consider the case where the prediction of a machine learning system disagrees with a doctor's intuition on a high-stakes medical decision. If the machine learning model simply makes a prediction (e.g., patient will have a stroke next year), this may not be very helpful on its own. On the other hand, if it is known exactly which variables were important for the prediction (e.g., type of tumor, age, treatment), how they were combined, and similar patient's outcomes (e.g., 14,582/16,871 patients survived) this information can be very powerful in helping to determine whether to believe the prediction and how to make the right decision. Further, patients might not trust a black-box system with a classical evaluation measure that they do not understand (e.g., ROC curves). We do not want our predictive models to have the curse of the Greek mythological character Cassandra, who no one believes even though she tells the truth; on the other hand, we also do not want the problem of the stock market crash of 2008 where people blindly trusted a predictive model with incorrect assumptions. People need to know when to trust these models.

1) I will start by discussing knowledge-based systems that are used currently in medicine, criminology, and in other domains. In these cases, humans created the models without looking at data at all.

2) Then I will discuss the classic machine learning approach to decision trees, which are the algorithms CART (Classification and Regression Trees) and C4.5. CART is possibly the most widely used predictive modeling technique currently used in industry. CART is popular arguably because of the logical structure of its models: they are comprised of IF-THEN rules, which are similar to the ways humans naturally reason. However, a major flaw with CART and C4.5 is that they are greedy/myopic strategies, and often produce suboptimal models. C4.5 models tend to be more accurate than CART models, but they are much less interpretable. Neither CART nor C4.5 is optimized for either accuracy or interpretability.

3) I will discuss a more recent algorithm called Scalable Bayesian Rule Lists (SBRL), which builds machine learning models that have helpful decision-making properties. These models are optimized for both accuracy and interpretability. Because of this, they can be both more accurate and more interpretable than CART or C4.5. Because SBRL is not greedy, it is more computationally intensive. I will show how these models are applied to problems in healthcare and criminology.

4) Finally, I will pose an open question about what characteristics we might desire in a machine learning that a human can customize to be more interpretable, with several preliminary ideas. Because we now have the ability to optimize logical models, we no longer need to suffer only with the models that are easy to produce, for instance by CART or C4.5.