Interpretable Approaches to Detect Bias in Black-Box Models

Sarah Tan
Cornell University
Ithaca, NY 14850
ht395@cornell.edu

Black-box machine learning models permeate our lives and are increasingly being deployed for high stakes decisions, such as credit scoring (Louzada, Ara, and Fernandes 2016), judicial bail decisions (Angwin et al. 2016), and hospital admissions. More complicated models are being trained for the promise of an increase in accuracy, sometimes at the expense of transparency or interpretability.

My dissertation research is grounded in the field of interpretability. I aim to develop methods to explain and interpret predictions from black-box machine learning models to help creators, as well as users, of machine learning models increase their trust and understanding of the models. Moreover, interpretability may be additionally valuable for bias detection when specific biases are not *a priori* known, as I will elaborate on below.

Previous Research

My past work has focused on interpreting predictions from tree-based black-box models, including random forests and gradient boosted trees. One line of work in interpretability centers on developing models that are sparse in features or model elements. Examples include training regression models with regularization to select less features, or post-training pruning of the weights of a neural network to reduce model complexity. I have been exploring sparsity in *observation* methods. The canonical example of this class of methods is prototype selection, where representative observations of a class are selected for presentation to a user. (Kim, Khanna, and Koyejo 2016).

One output from training a random forest that has received less attention is the proximity matrix, a n-by-n matrix (n is the number of observations) describing the proportion of trees in the forest where a pair of observations end up in the same terminal node. This similarity metric between observations is locally adaptive in tree space (Wager and Athey 2017) and reflects how the forest makes its predictions based on the observations' features. I utilized this similarity metric to develop a prototype selection method (Tan, Hooker, and Wells 2016), presenting an alternative to other tree ensemble interpretability methods such as seeking one tree that best represents the ensemble (Banerjee, Ding, and Noone 2012) or feature importance methods (Breiman 2001).

Current Research

Besides tree ensembles, I am interested in developing methods for interpretability of non-convolutional neural networks, an area of less research yet no less important than interpretability for convolutional neural networks. Convolutional neural networks (CNNs) have been applied with great success to structured data sets such as images (Krizhevsky, Sutskever, and Hinton 2012), text (Zhang and LeCun 2015), and speech (Mohamed, Dahl, and Hinton 2012). Correspondingly there has been much interest in interpreting the outputs of convolutional networks. However, data arising from critical domains such as healthcare is typically in the form of column-based features such as demographic variables, health information, etc., and if no spatial, temporal, or otherwise structured relationships are present[1], may be better modeled using non-convolutional neural networks, one example of which is multilayer perceptrons.

Interpreting Multilayer Perceptrons Using Model Distillation

Model distillation was originally introduced to distill knowledge from a large, complex model (the "teacher") to a simpler, faster model (the "student") (Hinton, Vinyals, and Dean 2015). Perhaps the first to explore the idea of model distillation for understanding were Craven and Shavit who distilled a multilayer perceptron into a decision tree (Craven and Shavlik 1995). I am interested whether modern neural networks that are deeper, have more complex architectures, and trained using modern techniques, including dropout, batch normalization, weight decay, etc. can still be distilled into model classes typically considered as transparent, such as decision trees, sparse regression models, etc. Preliminary results suggest that shallow (up to five layers) multilayer perceptrons teachers on small data sets and classification tasks can be distilled into student models such as gradient boosted trees and tree-based generalized additive models (Caruana et al. 2015). I am working on determining if the method works on larger data sets and regression tasks.

[1] Long Short-Term Memory networks, a type of recurrent neural network, have been compared to CNNs on longitudinal healthcare data. See (Suresh et al. 2017) for a summary.

Bias Detection Using Model Distillation

I am also applying the idea of transparent model distillation to black-box risk scoring models, and I will be presenting the paper "Detecting Bias in Black-Box Models Using Transparent Model Distillation" as an oral in the main track of the AI, Ethics, and Society conference. The paper is also available on arxiv (Tan et al. 2017). To summarize the approach, the black-box risk scoring model is treated as the teacher and distilled into a transparent student model in which each feature and its relationship to the risk score can be examined. We also train another model on the true outcome that the risk score is supposed to predict (i.e. default on a loan, for a credit score) which we use to compare against the student model of black-box risk score, to increase confidence that the student model is an accurate representation of the teacher model.

Casting bias as systematic differences between the black-box risk score and the true outcome, recalling that the risk score was designed to predict the true outcome, we find significant differences between certain feature groups other than race such as younger (age 18 and 19) and older (age above 70) age groups, as well as gender. Testing on the Chicago Police Department's (CPD) "Strategic Subject" risk score[2], our approach picks up the eight features that CPD claims were used to construct the risk score, and none of the other features the CPD claims were not used. While this work is further along than the multilayer perceptrons work, I am still working on more evaluations on simulated ground-truth data to validate that the model does not detect spurious differences.

Future Plans

The project on detecting bias using transparent model distillation has piqued my interest in exploring interpretability for bias detection. One compelling reason to investigate the use of transparent and interpretable models for bias detection is that specific biases need not be *a priori* known. Instead, a transparent model that reveals its inner workings could suggest areas of potential bias that did not previously come to mind but warrant more investigation.

For example, in my "Detecting Bias..." paper, the transparent model distillation approach suggested that COMPAS predicted recidivism risk for younger and older age groups (feature regions that we had not suspected of bias) to be significantly different than that for true recidivism outcomes. This then allowed us to go back to the data and attempt to generate possible explanations for this discrepancy that we could then further investigate. When deploying this approach initially on the UCI German credit data[3], after training a transparent student model on the true outcome, we found our error bars for the effect for native Germans much larger than that for foreign nationals. A quick examination of the data revealed that the data comprises mostly foreign nationals, with only a handful of German nationals, suggesting that this data is drawn from a very specific population that likely is not representative of the population one wishes to study when investigating possible bias in issuing loans.

Hence, interpretable methods for bias detection could be particularly useful there are likely many sources of biases – as is likely in modern data sets, with their size and complexity – that may be *a priori* not known. This motivates my dissertation research: to develop methods to explain and interpret predictions from black-box machine learning models.

[2] `https://data.cityofchicago.org/Public-Safety/Strategic-Subject-List/4aki-r3np`

[3] `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`

References

Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. Accessed May 26, 2017.

Banerjee, M.; Ding, Y.; and Noone, A.-M. 2012. Identifying representative trees from ensembles. *Statistics in Medicine*.

Breiman, L. 2001. Random forests. *Machine Learning*.

Caruana, R.; Lou, Y.; Gehrke, J.; Koch, P.; Sturm, M.; and Elhadad, N. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *KDD*.

Craven, M. W., and Shavlik, J. W. 1995. Extracting tree-structured representations of trained networks. In *NIPS*.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *NIPS Deep Learning Workshop*.

Kim, B.; Khanna, R.; and Koyejo, O. O. 2016. Examples are not enough, learn to criticize! criticism for interpretability. In *NIPS*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS*.

Louzada, F.; Ara, A.; and Fernandes, G. B. 2016. Classification methods applied to credit scoring: Systematic review and overall comparison. *Surveys in Operations Research and Management Science*.

Mohamed, A.-r.; Dahl, G. E.; and Hinton, G. 2012. Acoustic modeling using deep belief networks. *IEEE Transactions on Audio, Speech, and Language Processing*.

Suresh, H.; Hunt, N.; Johnson, A.; Celi, L. A.; Szolovits, P.; and Ghassemi, M. 2017. Clinical intervention prediction and understanding with deep neural networks. In *MLHC*.

Tan, S.; Caruana, R.; Hooker, G.; and Lou, Y. 2017. Detecting bias in black-box models using transparent model distillation. *arXiv:1710.06169*.

Tan, S.; Hooker, G.; and Wells, M. T. 2016. Tree space prototypes: Another look at making tree ensembles interpretable. *NIPS Interpretability Workshop*.

Wager, S., and Athey, S. 2017. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*.

Zhang, X., and LeCun, Y. 2015. Text understanding from scratch. In *NIPS*.