Dawei Huang

Interpretability of ML Abstract

It is important to consider the interpretability of machine learning algorithms as the legitimacy of its implication is predicated upon our understanding of its working mechanisms. We will consider one avenue of ML interpretability that a model is interpretable if users could adequately discern the contribution of each feature (input variable) to the final output. Of course there exists model-agnostic methods like Partial Dependence Plots and Sharpley Values that would allow us to ascertain feature contributions with a fair degree of certainty. These model-agnostic methods, however, are computationally expensive especially for large datasets so a new model-agnostic model is developed in order to provide good approximations of feature weights in faster time.

The randomGen model-agnostic method accesses the contribution of each variable input to the final output by consecutively altering random feature data at a random point, running these new inputs through the algorithm, and then measuring the deviation of the output from the initial output. This process will be repeated for each feature and line graphs for each feature will be plotted on a graph of the number of random changes vs deviation from initial output. Reasonably, we expect to see a downwards sloping graph as more random changes to feature should coincide with greater deviation from the initial output. A steeper slope implies that for that given feature, small random changes to its data has a large impact on output and thus would have a greater weight and vice versa. To better gauge the accuracy of randomGen, we will be running it multiple times and compare it to Shapley Value analysis.

We ran randomGen analysis for both Interpretable Models (linear Reg, decision tree, etc) and black-box models (Random Forest) for mainly regression and classification ML tasks. In general we find that randomGen works particularly well with regression tasks and we were able to roughly reproduce weight rankings provided by Shapley Value analysis to about ~80% accuracy. Differences between the slopes of features in randomGen analysis do not reasonably correspond to differences between feature weights so we are not able to reproduce reliable weight approximations. Therefore, randomGen is useful for initial regression analysis to quickly discern a few important features in any ML model. In addition to working well with interpretable models, tests suggest that randomGen is also useful with black box models like Random Forest. Further tests show that our randomGen analysis is also effective for classification tasks for both interpretable and black box models. The accuracy of the randomGen compared to the Shapley Value analysis may be determined using the accuracyScore method.

The randomGen model is most successful for models that generate initial outputs that align closely with the desired outputs. Otherwise it would be trivial to measure the deviation after consecutive random changes from the initial output if the initial output was not accurate itself. Before randomGen is used, the data should be preprocessed so that relevant dummy variables are generated for categorical variables.