Title

Author 1[*]
Department of YYY, University of XXX
and
Author 2
Department of ZZZ, University of WWW

August 30, 2017

Abstract

The text of your abstract. 200 or fewer words.

Keywords: 3 to 6 keywords, that do not appear in the title

[*] The authors gratefully acknowledge *please remember to list all relevant funding sources in the unblinded version*

1 Introduction

Complex nonparametric models—like neural networks, random forests, and support vector machines—are more common than ever in predictive analytics, especially when dealing with large observational databases that don't adhere to the strict assumptions imposed by traditional statistical techniques (e.g., multiple linear regression which assumes linearity, homoscedasticity, and normality). Unfortunately, it can be challenging to understand the results of such models and explain them to management. Variable importance plots and partial dependence plots (PDPs) offer a simple solution. PDPs are low-dimensional graphical renderings of the prediction function $\widehat{f}\left(\boldsymbol{x}\right)$ so that the relationship between the outcome and predictors of interest can be more easily understood. These plots are especially useful in explaining the output from black box models.

While PDPs can be constructed for any predictor in a fitted model, variable importance scores are more difficult to define. Some methods—like random forests and other tree-based methods—have a natural way of defining variable importance. Unfortunately, this is not the case for other popular supervised learning algorithms like support vector machines. In this paper, we offer a solution by providing a partial dependence-based variable importance metric that can be used with any supervised learning algorithm.

2 Background

In the age of "big data", we are often confronted with the task of extracting knowledge from large databases. For this task we turn to various statistical learning algorithms which, when tuned correctly, can have state-of-the-art predictive performance. However, having a model that predicts well is only solving part of the problem. It is still necessary to extract information about the relationships uncovered by the learning algorithm. For instance, we often want to know which predictors, if any, are important by assigning some type of variable importance score to each feature. Once a set of "important" features has been identified, the next step would be to summarize the functional relationship between each feature, or subset thereof, and the outcome of interest. However, since most statistical learning algorithms are "black box" models, extracting this information is not always straightforward.

Fortunately, some learning algorithms have a natural way of defining variable importance.

In a binary decision tree, at each node t, a single predictor is used to partition the data into two homogeneous groups. The chosen predictor is the one that maximizes some measure of improvement $\widehat{i}_t$. The relative importance of predictor x is just the sum of the squared improvements over all internal nodes of the tree for which x was chosen as the partitioning variable; see Breiman et al. (1984) for details. This idea also extends to ensembles of decision trees like boosting and random forest. In ensembles the the improvement score for each predictor is averaged across all the trees in the ensemble. Fortunately, due to the stabilizing effect of averaging, the improvement-based variable importance metric is often more reliable in large ensembles. Random forests offer an additional way to compute variable importance scores. The idea is to use the left over out-of-bag (OOB) data to construct validation-set errors for each tree. Then each predictor is randomly shuffled in the OOB data and the error is computed again. The idea is that if variable X is important, then the validation error will go up when X is perturbed in the OOB data. The difference in the two errors is recorded for the OOB data for each predictor then averaged across all trees in the forest.

In multiple linear regression, the absolute value of the t statistic is commonly used as a measure of variable importance. The same idea also extends to generalized linear models and nonlinear least squares. Multivariate adaptive regression splines (MARS), which were introduced in Friedman (1991), is an automatic and adaptive regression technique which can be seen as a generalization of multiple linear regression and generalized linear models. In the MARS algorithm, the contribution (or variable importance score) for each predictor is determined using a generalized cross-validation (GCV) statistic.

For neural networks, two popular methods for constructing variable importance scores are the Garson algorithm (Garson, 1991), later modified by Goh (1995), and the Olden algorithm (Olden et al., 2004). Both algorithms use the network's connection weights to form the basis of the variable importance scores. The Garson algorithm determines variable importance by identifying all weighted connections between the nodes of interest. Olden's algorithm, on the other hand, uses the product of the raw input-hidden and hidden-output connection weights between each input and output neuron and sums the product across all hidden neurons. This has been shown to outperform the Garson method in various simulations.

3 Conclusion

SUPPLEMENTARY MATERIAL

Title: Brief description. (file type)

R-package for MYNEW routine: R-package ?MYNEW? containing code to perform the diagnostic methods described in the article. The package also contains all data sets used as examples in the article. (GNU zipped tar file)

HIV data set: Data set used in the illustration of MYNEW method in Section 3.2. (.txt file)

4 BibTeX

We hope you've chosen to use BibTeX! If you have, please feel free to use the package natbib with any bibliography style you're comfortable with. The .bst file Chicago was used here, and agsm.bst has been included here for your convenience.

References

Breiman, L., J. Friedman, and R. A. O. Charles J. Stone (1984). *Classification and Regression Trees*. The Wadsworth and Brooks-Cole statistics-probability series. Taylor & Francis.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics 19*(1), 1–67.

Garson, D. G. (1991). Interpreting neural-network connection weights. *Artificial Intelligence Expert 6*(4), 46–51.

Goh, A. (1995). Back-propagation neural networks for modeling complex systems. *Artificial Intelligence in Engineering 9*(3), 143–151.

Olden, J. D., M. K. Joy, and R. G. Death (2004). An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. *Ecological Modelling 178*(3), 389–397.