A New Model-Based Notion of Variable Importance

Author 1
Author 1 address

Author 2
Author 2 address

Abstract

Keep to 200 words or less and don't use references or mathematical markup.

1 Introduction

Talk about the improtance of extracting information from black-box models, blah, blah, blah.

1.1 Partial dependence plots

Harrison and Rubinfeld [1978] were among the first to analyze the well-known Boston housing data. One of their goals was to find a housing value equation using data on median home values from $n = 506$ census tracts in the suburbs of Boston from the 1970 census; see Harrison and Rubinfeld [1978, Table IV] for a description of each variable. The data violate many classical assumptions like linearity, normality, and constant variance. Nonetheless, Harrison and Rubinfeld—using a combination of transformations, significance testing, and grid searches—were able to find a reasonable fitting model ($R^2 = 0.81$). Part of the payoff for there time and efforts was an interpretable prediction equation which is reproduced in Equation (1).

$$\begin{aligned}\log\left(\widehat{MV}\right) = {} & 9.76 + 0.0063RM^2 + 8.98 \times 10^{-5} AGE - 0.19 \log\left(DIS\right) + 0.096 \log\left(RAD\right) \\ & - 4.20 \times 10^{-4} TAX - 0.031 PTRATIO + 0.36\left(B - 0.63\right)^2 - 0.37 \log\left(LSTAT\right) \\ & - 0.012 CRIM + 8.03 \times 10^{-5} ZN + 2.41 \times 10^{-4} INDUS + 0.088 CHAS \\ & - 0.0064 NOX^2.\end{aligned} \tag{1}$$

Nowadays, many supervised learning algorithms can fit the data automatically in seconds—typically with higher accuracy. The downfall, however, is some loss of interpretation since these algorithms typically do not produce simple prediction formulas like Equation (1). These models can still provide insight into the data, but it is not in the form of simple equations. For example, quantifying predictor importance has become an essential task in the analysis of "big data", and many supervised learning algorithms, like tree-based methods, can naturally assign variable importance scores to all of the predictors in the training data.

While determining predictor importance is a crucial task in any supervised learning problem, ranking variables is only part of the story and once a subset of "important" features is identified it is often necessary to assess the relationship between them (or subset thereof) and the response. This can be done in many ways, but in machine learning it is often accomplished by constructing *partial dependence plots* (PDPs); see Friedman [2001] for details. PDPs help visualize the relationship between a subset of the features (typically 1-3) and the response while accounting for the average effect of the other predictors in the model. They are particularly effective with black box models like random forests and support vector machines.

Let $\boldsymbol{x} = \{x_1, x_2, \ldots, x_p\}$ represent the predictors in a model whose prediction function is $\widehat{f}(\boldsymbol{x})$. If we partition $\boldsymbol{x}$ into an interest set, $\boldsymbol{z}_s$, and its compliment, $\boldsymbol{z}_c = \boldsymbol{x} \setminus \boldsymbol{z}_s$, then the "partial dependence" of the response on $\boldsymbol{z}_s$ is defined as

$$f_s\left(\boldsymbol{z}_s\right) = E_{\boldsymbol{z}_c}\left[\widehat{f}\left(\boldsymbol{z}_s, \boldsymbol{z}_c\right)\right] = \int \widehat{f}\left(\boldsymbol{z}_s, \boldsymbol{z}_c\right) p_c\left(\boldsymbol{z}_c\right) d\boldsymbol{z}_c, \tag{2}$$

where $p_c(\boldsymbol{z}_c)$ is the marginal probability density of $\boldsymbol{z}_c$: $p_c(\boldsymbol{z}_c) = \int p(\boldsymbol{x})\, d\boldsymbol{z}_s$. Equation (2) can be estimated from a set of training data by

$$\bar{f}_s(\boldsymbol{z}_s) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}(\boldsymbol{z}_s, \boldsymbol{z}_{i,c}), \tag{3}$$

where $\boldsymbol{z}_{i,c}$ $(i = 1, 2, \ldots, n)$ are the values of $\boldsymbol{z}_c$ that occur in the training sample; that is, we average out the effects of all the other predictors in the model.

A random forest analysis...

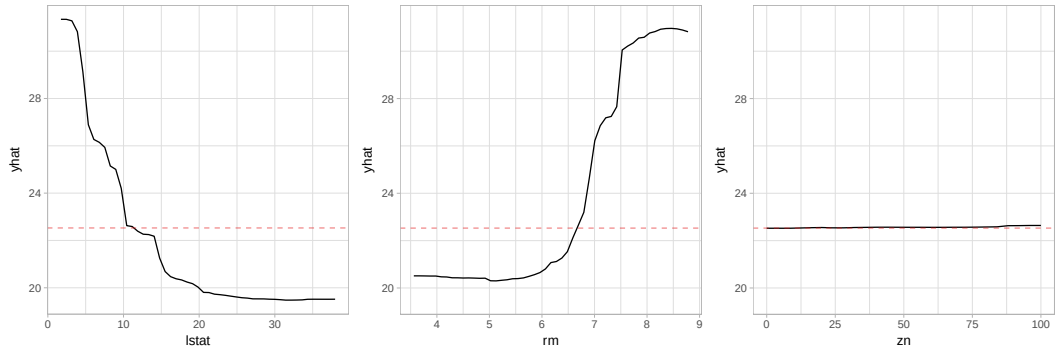

Figure 1: TBD.

2 Examples: a simple function of ten variables

For illustration, we use one of the regression problems described in Friedman (1991) and Breiman (1996). Inputs are 10 independent variables uniformly distributed on the interval $[0, 1]$; only 5 out of these 10 are actually used. Outputs are created according to the formula

$$\mathcal{Y} = 10\sin(\pi x_1 x_2) + 20(x_3 - 0.5)^2 + 10x_4 + 5x_5 + \epsilon, \tag{4}$$

where $\epsilon \sim \mathcal{N}(0, \sigma)$.

We fit a simple neural network with eight hidden units and a weight decay of 0.01. These parameters were chosen using 5-fold cross-validation to maximize the root mean squared error (RMSE). The network diagram is displayed in Figure **??** below.

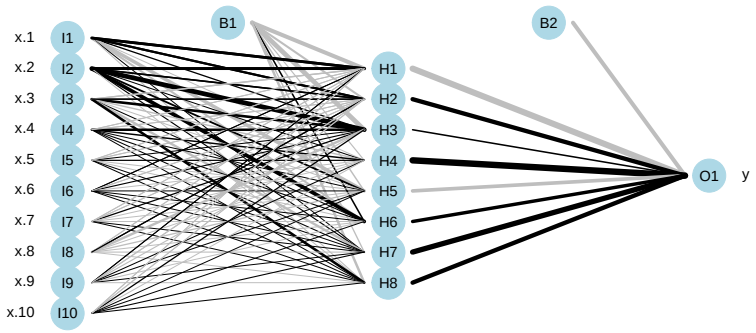

Figure 2: Diagram of the neural network fitted to the Friedman 1 data set.

Variable importance plots are displayed in Figure **??**. Notice how the Garson and Olden algorithms incorrectly label some of the features not in the true model as "important".

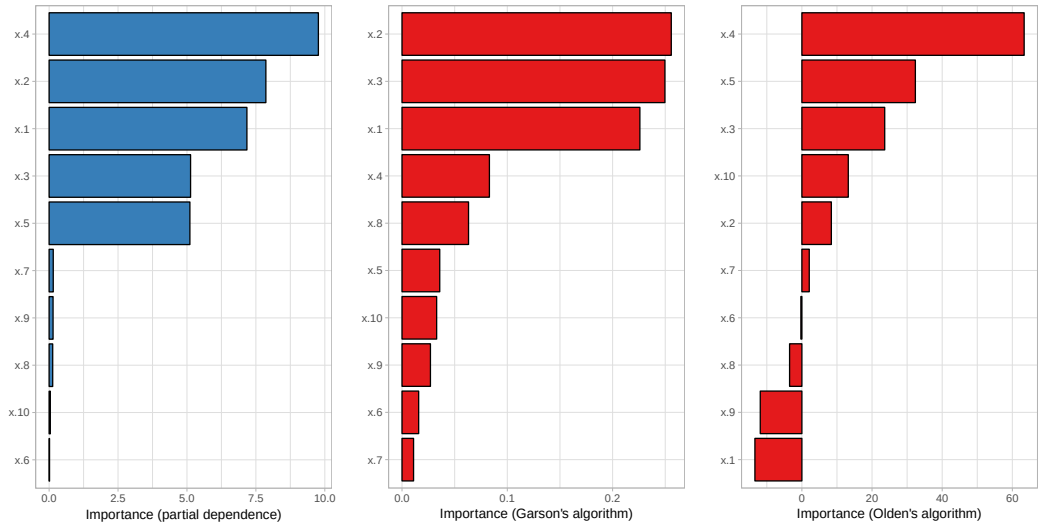

Figure 3: Variable importance plots for the neural network in Figure **??**. *Left*: partial dependence-based algorithm. *Middle*: Garson's algorithm. *Right*: Olden's algorithm.

For comparison, we fit a random forest [REFERENCE] with 1000 trees to the same data. Random forests (and other tree-based methods) have a natural way of denining variable importance. For details, see [REF]. Figure **??** displays two types of importance measures obtained from the fitted random forest.

3 Discussion

References

Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29:1189–1232, 2001. URL https://doi.org/10.1214/aos/1013203451.

David Harrison and Daniel L. Rubinfeld. Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1):81–102, 1978. URL https://doi.org/10.1016/0095-0696(78)90006-2.