Prediction Error Reduction Function as a Variable Importance Score

Ernest Fokoué[*,†]

[†]*School of Mathematical Sciences*
Rochester Institute of Technology
98 Lomb Memorial Drive, Rochester, NY 14623, USA
e-mail: epfeqa@rit.edu

Abstract: This paper introduces and develops a novel variable importance score function in the context of ensemble learning and demonstrates its appeal both theoretically and empirically. Our proposed score function is simple and more straightforward than its counterpart proposed in the context of random forest, and by avoiding permutations, it is by design computationally more efficient than the random forest variable importance function. Just like the random forest variable importance function, our score handles both regression and classification seamlessly. One of the distinct advantage of our proposed score is the fact that it offers a natural cut off at zero, with all the positive scores indicating importance and significance, while the negative scores are deemed indications of insignificance. An extra advantage of our proposed score lies in the fact it works very well beyond ensemble of trees and can seamlessly be used with any base learners in the random subspace learning context. Our examples, both simulated and real, demonstrate that our proposed score does compete mostly favorably with the random forest score.

1. Introduction

Consider a data set $\mathscr{D} = \{(\mathbf{x}_1, \mathbf{y}_1), \cdots, (\mathbf{x}_n, \mathbf{y}_n)\}$ where $\mathbf{x}_i$ is a p-dimensional vector of attributes of potentially different types observable on some input space denoted here by $\mathscr{X}$, and $\mathbf{y}_i$ are the responses taken from $\mathscr{Y}$. We shall consider various scenarios, but mainly the regression scenario with $\mathscr{Y} = \mathbb{R}$ and the classification scenario with $\mathscr{Y} = \{1, 2, \cdots, K\}$. We consider the task of building the estimator $\widehat{f}(\cdot)$ of the true but unknown underlying f, and seek to build $\widehat{f}(\cdot)$ such that the true error (generalization error) is as small as possible. In this context, we shall use the average test error $\mathtt{AVTE}(\cdot)$, as our measure of predictive performance, namely

$$\mathtt{AVTE}(\widehat{f}) = \frac{1}{R} \sum_{r=1}^{R} \left\{ \frac{1}{m} \sum_{j=1}^{m} \ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})) \right\}, \tag{1.1}$$

where $\left(\mathbf{x}_j^{(r)}, \mathbf{y}_j^{(r)}\right)$ is the jth observation from the test set at the rth random replication of the split of the data. Throughout this paper, we shall use the zero-one loss (1.2) for all our classification tasks.

$$\ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})) = 1_{\{\mathbf{y}_j^{(r)} \neq \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})\}} = \begin{cases} 1 & \text{if } \mathbf{y}_j^{(r)} \neq \widehat{f}^{(r)}(\mathbf{x}_j^{(r)}) \\ 0 & \text{otherwise.} \end{cases} \tag{1.2}$$

For regression tasks, we shall use the squared error loss (1.2), namely

$$\ell(\mathbf{y}_j^{(r)}, \widehat{f}^{(r)}(\mathbf{x}_j^{(r)})) = (\mathbf{y}_j^{(r)} - \widehat{f}^{(r)}(\mathbf{x}_j^{(r)}))^2. \tag{1.3}$$

Besides, seeking the optimal predictive estimator of f, we also seek to select the most important (useful) predictor variables as a byproduct of our overall learning scheme. Indeed, while accurate prediction is very important in and of itself, it's often desirable or even crucial in some cases, provide the added description of the importance of the variables involved in the prediction task. The statistical literature is filled with thousands of papers on variable selection and measurement of variable importance.

---

[*]Corresponding author

2. Main result

We consider a framework with a p-dimensional input space $\mathscr{X}$ with typical input vector $\mathbf{x} = (\mathbf{x}_1, \cdots, \mathbf{x}_p)^\top$. We also consider building different models with different subsets of the p original variables. Let $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_p)^\top$ denote the p-dimensional indicator such that

$$\gamma_j = \left\{ \begin{array}{ll} 1 & \text{if } \mathbf{x}_j \text{ is active in the current model indexed by } \boldsymbol{\gamma} \\ 0 & \text{otherwise.} \end{array} \right. \tag{2.1}$$

Assume that we are given an ensemble (collection or aggregation) of models, say

$$\mathscr{H} = \{h(\cdot, \boldsymbol{\gamma}^{(1)})), h(\cdot, \boldsymbol{\gamma}^{(2)})), \cdots, h(\cdot, \boldsymbol{\gamma}^{(B)}))\} \tag{2.2}$$

where $h(\cdot, \boldsymbol{\gamma}^{(b)}))$ denotes the function built with only those variables that are active in the bth model of the ensemble (aggregation), and $\boldsymbol{\gamma}^{(b)} = (\gamma_1^{(b)}, \cdots, \gamma_p^{(b)})$ with

$$\gamma_j^{(b)} = \left\{ \begin{array}{ll} 1 & \text{if } \mathbf{x}_j \text{ is active in the } b\text{-th model of the ensemble} \\ 0 & \text{otherwise.} \end{array} \right. \tag{2.3}$$

For instance, we may consider a homogeneous ensemble, i.e, an ensemble in which all the functions are of the same family, like the case where all the base learners are multiple linear regression (MLR) models differing by the variables upon which they are built. Consider a score function $\texttt{score}(h(\cdot, \boldsymbol{\gamma}^{(b)}))$ used to assess the performance of model indexed by the variables active in $\boldsymbol{\gamma}^{(b)}$. We propose a variable importance score in the form of a function that measures the importance of a variable $\mathbf{x}_j$ in terms of the reduction in average score

$$\texttt{PERF}(\mathbf{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \texttt{score}(h(\cdot, \boldsymbol{\gamma}^{(b)})) - \frac{1}{B_j} \sum_{b=1}^{B} \gamma_j^{(b)} \texttt{score}(h(\cdot, \boldsymbol{\gamma}^{(b)})) \tag{2.4}$$

where B_j is the number of models containing the variable $\mathbf{x}_j$, specifically $B_j = \sum_{b=1}^{B} 1_{\{\gamma_j^{(b)}=1\}}$. In words,

$$\texttt{PERF}(\mathbf{x}_j) = \texttt{Average score over all models} - \texttt{Average score over all models with } \mathbf{x}_j$$

Intuitively, $\texttt{PERF}(\mathbf{x}_j)$ somewhat measures the impact of variable $\mathbf{x}_j$. In the way similar to the approach used by sports writers to decide the MVP on a team or in a league, $\texttt{PERF}(\mathbf{x}_j)$ looks at the overall performance of the whole ensemble and then for each variable $\mathbf{x}_j$ computes the direction and magnitude of the change to that overall performance of the ensemble brought by its presence in models. *If a variable $\mathbf{x}_j$ is important, then its presence in any model will cause that model to perform better in the sense of having a lower than common average error (score). The average score of all models containing an important variable $\mathbf{x}_j$ should therefore be lower than the overall average score.*

- $|\texttt{PERF}(\mathbf{x}_j)|$ measures the magnitude of the importance/impact.
- $\texttt{sign}(\texttt{PERF}(\mathbf{x}_j))$ measures the direction of the impact.
- If $\texttt{sign}(\texttt{PERF}(\mathbf{x}_j)) = +1$ and $|\texttt{PERF}(\mathbf{x}_j)|$ is relatively large, then $\mathbf{x}_j$ is an important variable.

- Seamlessly applied to large p small n.
- All variables with $\texttt{PERF}(\mathbf{x}_j) \leq 0$ are unimportant and can be discarded.
- The $\texttt{PERF}(\cdot)$ score can be used whenever an ensemble $\mathscr{H}$ is available along with a suitable score function for each base learner.
- This works with any base learner and can be adapted to parametric, nonparametric and semi-parametric models and one can imagine ensembles with any base learners as its atoms.
- A great advantage over the traditional variable importance Breiman (2001a), Breiman (2001b) score functions is that the clear cut-off at zero, in the sense that all variables with $\texttt{PERF}(\mathbf{x}_j) > 0$ are kept and all those variables with $\texttt{PERF}(\mathbf{x}_j) \leq 0$ are thrown away.

2.1. PERF score via Random Subspace Learning

A natural implementation of $\texttt{PERF}(\cdot)$ can be done using the ubiquitous bootstrap along with the random subspace learning scheme. The $\texttt{Out-of-Bag (oob)}$ error in the bagging or random subspace learning context is a good

(in fact excellent) candidate score function, especially when the goal if the selection of variables that lead to the lowest prediction error. The advantage of using `oob` as the score lies in the fact that the score is obtained as part of building the ensemble in the random subspace learning framework. Consider the training set $\mathscr{D} = \{\mathbf{z}_i = (\mathbf{x}_i^\top, \mathrm{y}_i)^\top,\ i = 1, \cdots, n\}$, where $\mathbf{x}_i^\top = (\mathrm{x}_{i1}, \cdots, \mathrm{x}_{ip})$ and $\mathrm{y}_i \in \mathscr{Y}$ are realizations of two random variables X and Y respectively. Let $\mathbf{x}_{i,\boldsymbol{\pi}_j} = (\mathrm{x}_{i,1}, \cdots, \mathrm{x}_{i,\pi_j}, \cdots, \mathrm{x}_{i,d})$. The permutation $\boldsymbol{\pi}_j$ acts the $|\bar{\mathscr{D}}^{(b)}|$-dimensional jth column of the out-of-bag data matrix. Essentially, $\boldsymbol{\pi}_j$ simply permutes the $|\bar{\mathscr{D}}^{(b)}|$ elements of the jth column of the out-of-bag data matrix.

Algorithm 1 PERF Score Estimate via Random Subspace Learning

1: **procedure** PERF SCORE(B) ▷ Computing the PERF Score based on B base learners
2:    Choose a base learner $\widehat{h}(\cdot)$ ▷ e.g.: Trees, MLR
3:    Choose an estimation method ▷ e.g.: Recursive Partitioning or OLS
4:    Initialize all the $\widehat{\mathrm{PERF}}(\mathrm{x}_j)$ and $\widehat{\mathrm{VI}}(\mathrm{x}_j)$ at zero
5:    **for** $b = 1$ to B **do**
6:        Draw with replacement from $\mathscr{D}$ a bootstrap sample $\mathscr{D}^{(b)} = \{\mathbf{z}_1^{(b)}, \cdots, \mathbf{z}_n^{(b)}\}$
7:        Draw without replacement from $\{1, \cdots, p\}$ a subset $\mathscr{V}^{(b)} = \{j_1^{(b)}, \cdots, j_d^{(b)}\}$ of d variables.
8:        Form the indicator vector $\boldsymbol{\gamma}^{(b)} = (\gamma_j^{(b)}, \cdots, \gamma_p^{(b)})$ with

$$\gamma_j^{(b)} = \left\{ \begin{array}{ll} 1 & \text{if } j \in \{j_1^{(b)}, \cdots, j_d^{(b)}\} \\ 0 & \text{otherwise.} \end{array} \right.$$

9:        Drop unselected variables from $\mathscr{D}^{(b)}$ so that $\mathscr{D}_{\text{sub}}^{(b)}$ is d dimensional
10:        Build the bth base learner $\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})$ based on $\mathscr{D}_{\text{sub}}^{(b)}$
11:        Compute score of the bth base learner $\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})$ ▷ e.g. Out-of-bag error

$$\mathbf{s}^{(b)} = \text{score}(\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})) = \frac{1}{|\bar{\mathscr{D}}^{(b)}|} \sum_{\mathbf{z}_i \notin \mathscr{D}^{(b)}} \ell(\mathrm{y}_i, \widehat{h}(\mathbf{x}_i, \boldsymbol{\gamma}^{(b)}))$$

12:        **for** $j \in \mathscr{V}^{(b)}$ **do**
13:            Generate the permutation of the jth column of $\bar{\mathscr{D}}^{(b)}$, namely

$$\boldsymbol{\pi}_j$$

14:            Compute the permutation impacted score

$$\mathbf{s}_{\boldsymbol{\pi}_j}^{(b)} = \text{score}_{\boldsymbol{\pi}_j}(\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})) = \frac{1}{|\bar{\mathscr{D}}^{(b)}|} \sum_{\mathbf{z}_i \notin \mathscr{D}^{(b)}} \ell(\mathrm{y}_i, \widehat{h}(\mathbf{x}_{i,\boldsymbol{\pi}_j}, \boldsymbol{\gamma}^{(b)}))$$

15:            Compute the bth instance of the importance of x_j

$$\widehat{\mathrm{VI}}^{(b)}(\mathrm{x}_j) = \mathbf{s}^{(b)} - \mathbf{s}_{\boldsymbol{\pi}_j}^{(b)}$$

16:        **end for**
17:    **end for**
18:    Use the ensemble $\mathscr{H} = \left\{ \widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)}),\ b = 1, \cdots, B \right\}$ to form the estimator

$$\widehat{\mathrm{PERF}}(\mathrm{x}_j) = \frac{1}{B} \sum_{b=1}^{B} \text{score}(\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})) - \frac{1}{B_j} \sum_{b=1}^{B} \gamma_j^{(b)} \text{score}(\widehat{h}(\cdot, \boldsymbol{\gamma}^{(b)})) \tag{2.5}$$

$$\widehat{\mathrm{VI}}(\mathrm{x}_j) = \frac{1}{B_j} \sum_{b=1}^{B} \gamma_j^{(b)} \widehat{\mathrm{VI}}^{(b)}(\mathrm{x}_j) \tag{2.6}$$

19: **end procedure**

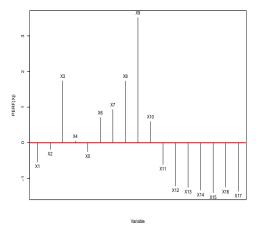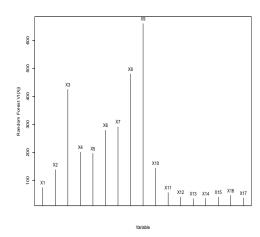## 3. Computational demonstrations

3.1. Simulated Example

The dataset in this example is simulated data with different scenarios on the level of correlation among the variables, and the ratio n and p. In this particular example, the true function is

$$f(\mathbf{x}) = 1 + 2\mathrm{x}_3 + \mathrm{x}_7 + 3\mathrm{x}_9$$

(a) Permutation-free Variable Importance.　　(b) Permutation-based Variable Importance.

FIG 1. *Variable score for simulated data with high correlation among the variables in low dimension high sample size setting*

with $\mathbf{x} \sim \mathtt{MVN}(\mathbf{1}_9, \Sigma_\rho)$ and $\epsilon \sim \mathbf{N}(0,1)$. The dataset in this example is simulated data with different scenarios on the level of correlation among the variables, and the ratio n and p. Specifically, we simulate data by defining $\rho \in [0,1)$, then we generate our predictor variables using a multivariate normal distribution. Throughout this paper, the multivariate Gaussian density will be denoted by $\phi_p(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$

$$\phi_p(\mathbf{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^p |\Sigma|}} \exp\left\{ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) \right\} \tag{3.1}$$

Furthermore, in order to study the effect of the correlation pattern, we simulate the data using a covariance matrix Σ parameterized by τ and ρ and defined by $\tau\Sigma$ where $\Sigma = (\sigma_{ij})$ with $\sigma_{ij} = \rho^{|i-j|}$.

$$\Sigma = \Sigma(\tau, \rho) = \tau \begin{pmatrix} 1 & \rho & \cdots & \rho^{p-2} & \rho^{p-1} \\ \rho & 1 & \rho & \cdots & \rho^{p-2} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ & & \ddots & & \\ \rho^{p-2} & \ddots & \rho & 1 & \rho \\ \rho^{p-1} & \rho^{p-2} & \cdots & \rho & 1 \end{pmatrix}$$
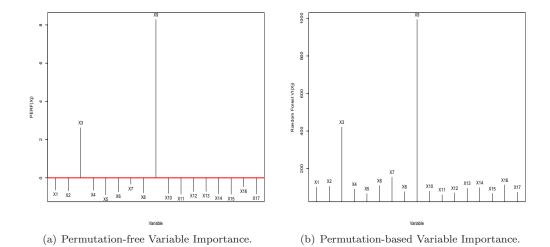
For simplicity however, we use the first Σ with $\tau = 1$ throughout this paper. For the remaining parameters, we use $\rho \in \{0, 0.25, 0.75\}$ and $p \in \{17, 250\}$, with the same $n = 200$.
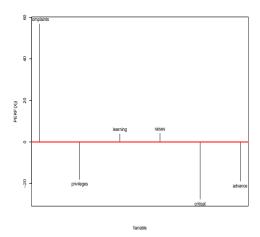
4. Conclusion and Discussion

We have presented a variable importance score function in the context of ensemble learning. Our proposed score function is simple and more straightforward than its counterpart proposed in the context of random forest, and by avoiding permutations, it is by design computationally more efficient than the random forest variable importance function. Just like the random forest variable importance function, our score handles both regression and classification seamlessly. One of the distinct advantage of our proposed score is the fact that it offers a natural cut off at zero, with all the positive scores indicating importance and significance, while the negative scores are deemed indications of insignificance. An extra advantage of our proposed score lies in the fact it works very well beyond ensemble of trees and can seamlessly be used with any base learners in the random subspace learning context. Our examples, both simulated and real, demonstrated that our proposed score does compete mostly favorably with the random forest score. In our future work, we present and compare the corresponding average test errors of the single models made up of the most important variables. We also provide in our future work theoretical proofs of the connection between our score function and the significance
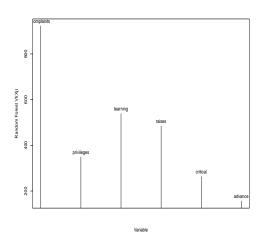
(a) Permutation-free Variable Importance.

(b) Permutation-based Variable Importance.

FIG 2. *Variable Importance Scores for simulated data with mild correlation among the variables in low dimension high sample size setting*
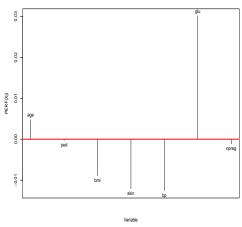

(a) Permutation-free Variable Importance.

(b) Permutation-based Variable Importance.

FIG 3. *Variable Importance Scores for simulated data with zero correlation among the variables in low dimension high sample size setting*
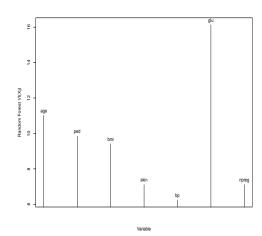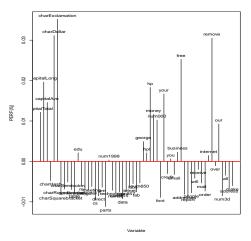
(a) Permutation-free Variable Importance.

(b) Permutation-based Variable Importance.

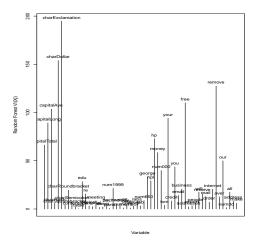FIG 4. *Variable Importance Scores for the Attitude Data Set, for which $n = 30$ and $p = 6$.*

(a) Permutation-free Variable Importance.

(b) Permutation-based Variable Importance.

FIG 5. *Variable Importance Scores for the Spam Detection Dataset where $n = 200$ and $p = 7$, and $K = 2$ classes.*

(a) Permutation-free Variable Importance.

(b) Permutation-based Variable Importance.

FIG 6. *Variable Importance Scores for the Spam Detection Dataset where $n = 4601$ and $p = 57$, and $K = 2$ classes.*

of variables selected using existing criteria. It is also our plan to address the fact that sometimes the correlation structure among the predictor variables obscures the ability of our proposed score to correctly identify some significant variables.

Acknowledgements

References

Breiman, L. (2001a). Random forests. *Machine Learning 45*, 5–32.
Breiman, L. (2001b, August). Statistical modeling: The two cultures. *Statistical Science 16*(3), 199–215.