Random forests: a sensitivity analysis perspective

C. Bénard
Sorbonne Université

Supervisor(s): E. Scornet (Ecole Polytechnique), G. Biau (Sorbonne Université), and S. Da Veiga (Safran Tech)

Ph.D. expected duration: Nov. 2018 - Nov. 2021

Adress: Safran Tech, Modeling & Simulation
Rue des Jeunes Bois, Châteaufort, 78114 Magny-Les-Hameaux, France

Email: clement.benard@safrangroup.com

Abstract:

Random forests. Random forests (Breiman, 2001) are an ensemble learning algorithm, which aggregates a large number of trees to perform regression and classification tasks, and achieve state-of-the-art accuracy on a wide range of problems. However, they suffer from a major drawback: a given prediction is generated through a large number of operations, typically ten thousands, which makes the interpretation of the prediction mechanism impossible. This black-box nature is a strong practical limitation, in particular for applications involving critical decisions: healthcare or the optimization of manufacturing processes are typical examples. The most popular way to interpret random forests is variable importance analysis: input variables are ranked by decreasing order of their importance in the prediction process. Thus, specific variable importance measures were developed along with random forests, and the MDA (Breiman, 2001, Mean Decrease Acccuracy) is the most widely used one. The MDA measures the decrease of accuracy when the values of a given input variable are permuted, thus breaking its relation to the output: a high value of the metric means that the variable is used in many important operations of the prediction mechanism of the forest. Unfortunately, there is no precise and rigorous interpretation since the MDA definition is purely empirical. Our objective is to use sensitivity analysis to unveil the theoretical properties of the MDA, and introduce the Sobol-MDA algorithm which fixes the flaws of the original MDA.

Sensitivity analysis. The goal of sensitivity analysis is to apportion the uncertainty of a system output to the uncertainty of the different inputs. In particular, GSA introduces well-defined importance measures of input contributions to the output variance, Sobol indices, especially widely used to analyze computer code experiments. However, the literature about variable importance in machine learning rarely mentions sensitivity analysis. In the last years, Gregorutti (2015) first established a link between GSA and the MDA for random forests: in the case of independent inputs, the theoretical counterpart of the MDA is the unnormalized total Sobol index. We bring the connection between random forests and sensitivity analysis one step further: we prove the convergence of Breiman's MDA in the general case of dependent inputs and break down the obtained limit as a sum of Sobol indices.

MDA definitions. A close inspection of the main random forest software reveals that several MDA implementations coexist without a clear mathematical formulation and are not equivalent: the Train-Test MDA (TT-MDA), the Breiman-Cutler MDA (Breiman, 2001, BC-MDA), and the Ishwaran-Kogalur MDA. To formalize the MDA definitions, we first consider a standard regression setting, where the response $Y \in \mathbb{R}$ follows $Y = m(\mathbf{X}) + \varepsilon$ with the input $\mathbf{X} = (X^{(1)}, \dots, X^{(p)}) \in [0,1]^p$, and a sample $\mathscr{D}_n = \{(\mathbf{X}_1, Y_1), \dots, (\mathbf{X}_n, Y_n)\}$ is available. Next, the random forest estimate $m_{M,n}(\mathbf{x}, \mathbf{\Theta}_M)$ aggregates M Θ-random CART denoted $m_n(\mathbf{x}, \Theta_\ell)$, each of which is randomized by a component of $\mathbf{\Theta}_M = (\Theta_1, \dots, \Theta_M)$ for the bootstrap sampling $(\Theta_\ell^{(S)})$ and the randomization of each node split $(\Theta_\ell^{(V)})$. In the original MDA definition from Breiman, the quadratic risk of each tree is estimated for both the out-of-bag sample and the permuted

	BC-MDA*	$\widehat{\text{BC-MDA}}$	IK-MDA*	$\widehat{\text{IK-MDA}}$	ST*	$\widehat{\text{S-MDA}}$
$\mathbf{X}^{(3)}$	0.47	0.37	0.47	0.43	0.47	0.45
$\mathbf{X}^{(4)}$	0.21	0.10	0.37	0.14	0.10	0.08
$\mathbf{X}^{(5)}$	0.21	0.09	0.37	0.13	0.10	0.08
$\mathbf{X}^{(1)}$	0.64	0.24	1.0	0.29	0.07	0.05
$\mathbf{X}^{(2)}$	0.64	0.24	1.0	0.28	0.07	0.05

Table 1: Normalized BC-MDA, normalized IK-MDA, and Sobol-MDA estimates.

out-of-bag sample. The difference between these two risks is averaged across all trees to define the BC-MDA. More precisely, for each Θ_ℓ-random tree, we randomly permute the j-th component of the out-of-bag dataset, and denote $\mathbf{X}_{i,\pi_{j\ell}}$ the i-th permuted sample for the ℓ-th tree. Then, with $N_{n,\ell} = \sum_{i=1}^n \mathbb{1}_{i \neq \Theta_\ell^{(S)}}$ the size of the out-of-bag sample of the ℓ-th tree, the BC-MDA is defined by

$$\widehat{\text{MDA}}_{M,n}^{(BC)}(X^{(j)}) = \frac{1}{M}\sum_{\ell=1}^{M}\frac{1}{N_{n,\ell}}\sum_{i=1}^{n}\left[(Y_i - m_n(\mathbf{X}_{i,\pi_{j\ell}}, \Theta_\ell))^2 - (Y_i - m_n(\mathbf{X}_i, \Theta_\ell))^2\right]\mathbb{1}_{i \notin \Theta_\ell^{(S)}}.$$

MDA theoretical limitations. To our knowledge, we provide the first convergence result of Breiman's MDA with Theorem 1 (under mild Assumptions (A1), (A2), and (A3)). Additionally, the MDA limits can be analyzed with a Sobol index decomposition. We define the total Sobol index $ST^{(j)}$, the full total Sobol index $ST_{full}^{(j)}$, and $\text{MDA}_3^{\star(j)} = \mathbb{E}[(\mathbb{E}[m(\mathbf{X})|\mathbf{X}^{(-j)}] - \mathbb{E}[m(\mathbf{X}_{\pi_j})|\mathbf{X}^{(-j)}])^2]$.

Theorem 1. *If Assumptions (A1), (A2) and (A3) are satisfied, then for all $M \in \mathbb{N}^\star$ and $j \in \{1, \dots, p\}$ we have*

$$(i) \quad \widehat{MDA}_{M,n}^{(TT)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)} + MDA_3^{\star(j)}$$

$$(ii) \quad \widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST_{full}^{(j)} + MDA_3^{\star(j)}.$$

If Assumption (A4) is additionally satisfied, then

$$(iii) \quad \widehat{MDA}_{M,n}^{(IK)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + MDA_3^{\star(j)}.$$

Importantly, the term $\text{MDA}_3^{\star(j)}$ is not related to an importance measure. In particular, $\text{MDA}_3^{\star(j)}$ is null when the regression function is additive, or inputs are independent. Otherwise, $\text{MDA}_3^{\star(j)}$ can considerably inflate the MDA without a clear meaning. Therefore, when inputs are dependent, the MDA can be misleading as already observed in many empirical studies.

Sobol-MDA. We propose the Sobol-MDA algorithm, a new importance measure for random forests which fixes the flaws of the original MDA. The Sobol-MDA is not permutation-based, but uses projections to eliminate a given variable from the forest predictions. We prove that the Sobol-MDA consistently estimates the total Sobol index even when inputs are dependent, as opposed to Breiman's MDA. Table 1 provides experiments for a simple regression function and a correlated Gaussian input of dimension 5. Only the Sobol-MDA ranks the variables in the appropriate order of the theoretical total Sobol index. In the specific setting where the sample size is small and the regression function is nonlinear but smooth, Gaussian processes often strongly outperforms random forests. In this case, the efficient black-box model may be used to generate a large sample to estimate the total Sobol index via the Sobol-MDA in a second step.

References

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

B. Gregorutti. *Random forests and variable selection : analysis of the flight data recorders for aviation safety*. Theses, Université Pierre et Marie Curie - Paris VI, March 2015.

Short biography – Clément Bénard is a research engineer at Safran Tech in the Modeling & Simulation team. His PhD is funded by Safran, and led in collaboration with Sorbonne Université.