Understandig variable importances in forests of randomized trees [Sun88]
Gilles Louppe, Louis Wehenkel, Antonio Sutera and Pierre Geurts

**Abstract**

Despite growing interest and practical use in various scientific areas, variable importances derived from tree-based ensemble methods are not well understood from a theoretical point of view. In this work we characterize the Mean Decrease Impurity (MDI) variable importances as measured by an ensemble of totally randomized trees in asymptotic sample and ensemble size conditions. We derive a three-level decomposition of the information jointly provided by all input variables about the output in terms of:

i) the MDI importance of each input variable, 
ii) the degree of interaction of an input variable with the other input variables, 
iii) the different interaction terms of a given degree.

We then show that this MDI importance of a variable is equal to zero if and only if the variable is irrelevant and that the MDI importance of a relevant variable is invariant with respect to the removal or the addition of irrelevant variables. We illustrate these properties on a simple example and discuss how they may change in the case of non-totally randomized trees such as Random Forests and Extra-Trees.

**Theoretical results**

1. **Thm. 1 and 2**: Variable importances provide a three-level decomposition of the information jointly provided by all the input variables about the output, accounting for all interaction terms in a fair and exhaustive way.

2. **Thm. 3 and 5**: Variable importances depend only on the relevant variables.

**Theorem 1.** The MDI importance of $X_m \in V$ for $Y$ as computed with an infinite ensemble of fully developed totally randomized trees and an infinitely large training set is:

$$\text{Imp}(X_m) = \lim_{K \to \infty} \mathbb{E} \left[ \sum_{i=1}^{K} \sum_{B \in P_t(v_i)} I(X_m, Y | B) \right]$$

where $P_t(v_i)$ is the probability of the branch $B = \{X_i1, X_i2\}$ (in light blue) and $I(X_m, Y | B)$ is the conditional mutual information of $X_m$ and $Y$ given the variables in $B$.

**Proof.** (sketch)

(i) Using the Shannon entropy, $\Delta(t) = I(X_m, Y | t)$.
(ii) As $N \to \infty$, $\rho(t) \to p(B = b)$ and $I(X_m, Y | t) \to I(X_m, Y | B = b)$, where $B$ is the subset of $k$ variables in the branch leading to $t$ and $b$ is the vector of values of these variables; 
(iii) As $N \to \infty$, branches $B = b$ of size $k$ all appear with equal probability $\frac{1}{|B|}$ and $X_m$ is tested at the end of $\frac{1}{|B|}$ of them.

$\Rightarrow$ Equation (1) transforms into Equation (2).

**Theorem 2.** For any ensemble of fully developed trees in asymptotic learning sample size conditions, we have that

$$\sum_{I(X_m)} \text{Imp}(X_m) \leq \text{Inf}(X_m)$$

where $\text{Inf}(X_m)$ is the information provided by all input variables about the output.

**Theorem 3.** $X \in V$ is irrelevant to $Y$ if and only if its infinite sample size importance as computed with an infinite ensemble of fully developed totally randomized trees built on $V$ for $Y$ is 0.

**Theorem 5.** Let $V_b \subseteq V$ be the subset of all variables in $V$ that are relevant to $Y$ with respect to $V$. The infinite sample size importance of any variable $X_i \in V_b$ as computed with an infinite ensemble of fully developed totally randomized trees built on $V_b$ for $Y$ is the same as its importance computed in the same conditions by using all variables in $V$.

**Illustration**

**Task.** Let us consider a 7-segment indicator displaying numerals using lights in on-off combinations. Let $Y$ be a random variable taking its value in $\{0, 1, \ldots, 9\}$ and let $X_1, \ldots, X_8$ be binary variables corresponding to the light segments. We illustrate variable importances as computed with totally randomized trees built from training samples drawn from $P(X_1, \ldots, X_8, Y)$.

**Effect of randomization.** Let $K$ (aka. $\text{ntree}$ or max_features) be the number of variables drawn to maximize $\Delta$. Variable importances at $K = 1$ follow theoretical values of Theorem 1. However, as $K$ increases, importances diverge due to masking effects. In accordance with Theorem 2, their sum is also always equal to $I(X_1, \ldots, X_8, Y)$.

**Illustration.** Variable importances decompose along the degrees $k$ of interactions of one variable with the other ones. At $K = 1$ (left), all $I(X_m, Y | B)$ are accounted for in the total importance, while at $K = 7$ (right) only some of them are taken into account due to masking effects.

**Conclusions**

- First step towards understanding variable importances, as computed with a forest of totally randomized trees.
- Variable importances offer a three-level decomposition of the information provided by the inputs about the output. 
- MDI importances exhibit desirable properties for assessing the relevance of a variable:
  - it accounts for all interaction terms, in a fair and exhaustive way; 
  - it is null if and only if the variable is irrelevant; 
  - it depends only on the relevant variables; 
- Fully formalize variable importances of actual Random Forests and variants. 
- Characterize the distribution of variable importances in a finite setting.