Understanding variable importances in forests of randomized trees [Sun88]

Gilles Louppe, Louis Wehenkel, Antonio Sutera and Pierre Geurts

g.louppe@ulg.ac.be @glouppe https://github.com/glouppe/paper-variable-importances

LA LIBERTÉ DE CHERCHER

Université de Liège

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**

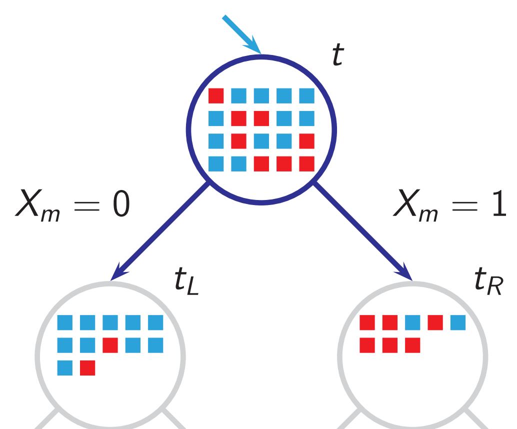
- **i) the MDI importance of each input variable,**
- **ii) the degree of interaction of an input variable with the other input variables,**
- **iii) the different interaction terms of a given degree.**

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.

Variable importances in trees

Notations. Let assume a set $V = \{X_1, ..., X_p\}$ of categorical input variables and a categorical output variable Y. Given a training sample $\mathcal{L}$ of N joint observations of $X_1, ..., X_p, Y$ drawn from $P(X_1, ..., X_p, Y)$, let us define for any internal node t of a decision tree built from $\mathcal{L}$:

- The number of training samples in t as N_t ;
- The proportion of training samples in t as $p(t) = \frac{N_t}{N}$;
- The impurity of node t as $i(t) = H(Y|t)$ (i.e., the Shannon entropy) ;
- The impurity decrease at node t as $\Delta i(t) = i(t) - \frac{N_{t_L}}{N_t} i(t_L) - \frac{N_{t_R}}{N_t} i(t_R)$.



$i(t) = 0.97$

$i(t_L) = 0.65$

$i(t_R) = 0.81$

$\Delta i(t) = i(t) - \frac{12}{20} i(t_L) - \frac{8}{20} i(t_R)$

$= 0.25$

Definition. In an ensemble of decision trees, the *Mean Decrease Impurity* (MDI) importance of an input variable X_m is the sum of the weighted impurity decreases $p(t)\Delta i(t)$, for all nodes t where X_m is used, averaged over all N_T trees in the ensemble :

$$Imp(X_m) = \frac{1}{N_T} \sum_T \sum_{t \in T : v(t) = X_m} p(t) \Delta i(t) \qquad (1)$$

where $v(t)$ is the variable used to split node t.

Definition. A *fully developed totally randomized tree is a decision tree* in which each node t is partitioned using a variable X_i picked uniformly at random (among those not yet used at the parent nodes) into $|\mathcal{X}_i|$ sub-trees (i.e., one for each possible value of $\mathcal{X}_i$) and where the recursive construction halts when all p variables have been used along the current branch.

Theoretical results

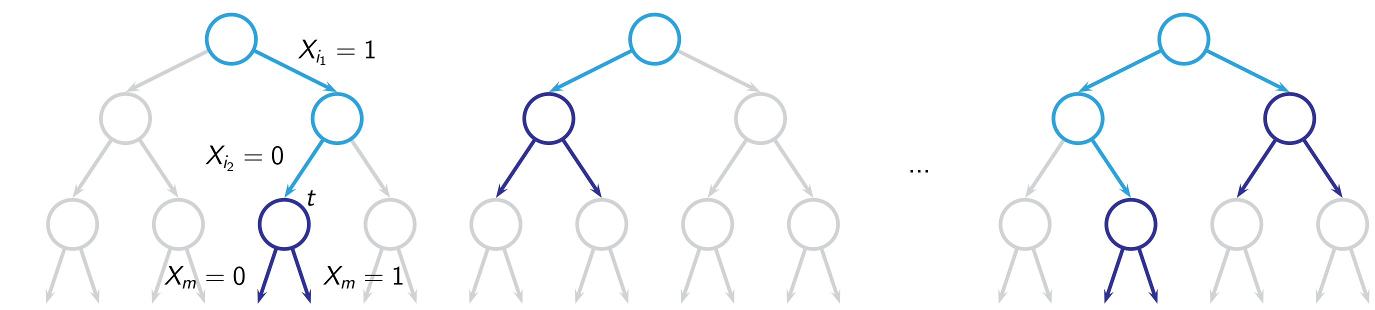
✓ Thm. 1 and 2 : **Variable importances provide a three-level decomposition of the information jointly provided by all the input variables about the output**, accounting for all interaction terms in a fair and exhaustive way.
✓ Thm. 3 and 5 : **Variable importances depend only on the relevant variables.**

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 :*

$$Imp(X_m) = \underbrace{\sum_{k=0}^{p-1} \frac{1}{C_p^k} \frac{1}{p-k}}_{\text{ii) Decomposition along the degrees } k \text{ of interaction with the other variables}} \underbrace{\sum_{B \in \mathcal{P}_k(V^{-m})} I(X_m; Y|B)}_{\text{iii) Decomposition along all interaction terms } B \text{ of a given degree } k}, \qquad (2)$$

where V^{-m} denotes the subset $V \setminus \{X_m\}$, $\mathcal{P}_k(V^{-m})$ is the set of subsets of V^{-m} of cardinality k, and $I(X_m; Y|B)$ is the conditional mutual information of X_m and Y given the variables in B.

Proof. (*sketch*)



$\frac{1}{C_p^k}$ is the probability of the branch $B = \{X_{i_1}, X_{i_2}\}$ (in light blue)

$\frac{1}{p-k}$ is the probability of drawing X_m (in blue) given B

(i) Using the Shannon entropy, $\Delta i(t) = I(X_m; Y|t)$;
(ii) As $N \to \infty$, $p(t) \to p(B = b)$ 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 the vector of values of these variables ;
(iii) As $N_T \to \infty$, branches $B = b$ of size k all appear with equal probability $\frac{1}{C_p^k}$ and X_m is tested at the end of $\frac{1}{p-k}$ of them.
⇒ Equation (1) transforms into Equation (2). □

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

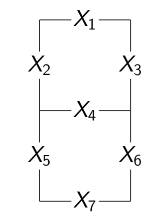
$$\underbrace{\sum_{m=1}^{p} Imp(X_m)}_{\text{i) Decomposition in terms of the MDI importance of each input variable}} = \underbrace{I(X_1, \ldots, X_p; Y)}_{\text{Information jointly provided by all input variables about the output}} \qquad (3)$$

Theorem 3. *$X_i \in V$ is irrelevant to Y with respect to V 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 $V_R \subseteq V$ 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 $X_m \in V_R$ as computed with an infinite ensemble of fully developed totally randomized trees built on V_R 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 $\{0, 1, ..., 9\}$ and let $X_1, ..., X_7$ be binary variables corresponding to the light segments. We illustrate variable importances as computed with totally randomized trees built from training samples drawn from $P(X_1, ..., X_7, Y)$.

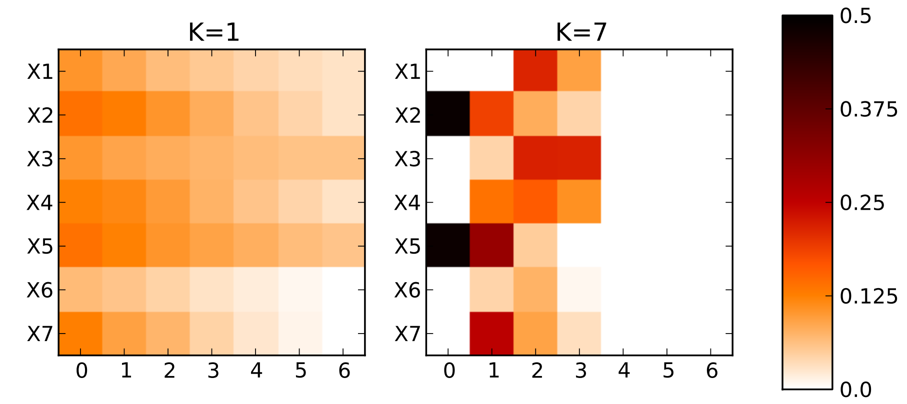


Effect of randomization. Let K (aka. `mtry` or `max_features`) be the number of variables drawn to maximize Δi. Variable importances at $K = 1$ follow theoretical values of Theorem 1. However, as K increases, importances diverge due to masking effects. In accordance with Theorem 2, their sum is also always equal to $I(X_1, \ldots, X_7; Y) = H(Y) = \log_2(10) = 3.321$ since inputs allow to perfectly predict the output.

	Thm.1	$K=1$	$K=2$	$K=3$	$K=4$	$K=5$	$K=6$	$K=7$
X_1	0.412	0.414	0.362	0.327	0.309	0.304	0.305	0.306
X_2	0.581	0.583	0.663	0.715	0.757	0.787	0.801	0.799
X_3	0.531	0.532	0.512	0.496	0.489	0.483	0.475	0.475
X_4	0.542	0.543	0.525	0.484	0.445	0.414	0.409	0.412
X_5	0.656	0.658	0.731	0.778	0.810	0.827	0.831	0.835
X_6	0.225	0.221	0.140	0.126	0.122	0.122	0.121	0.120
X_7	0.372	0.368	0.385	0.392	0.387	0.382	0.375	0.372
$\sum$	3.321	3.321	3.321	3.321	3.321	3.321	3.321	3.321

=

Decomposition. 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.



✗ Because of masking effects due to the non-totally random choices of split variables, **Theorems 1, 3 and 5 do not apply for Random Forests** and variants. Increasing K makes importance scores diverge from a fair and exhaustive exploration of all interaction terms.

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.