Variable importances in random forests

Erwan Scornet (CMAP, École Polytechnique, IP Paris)

Random forests are great!

Random forests are a class of algorithms created by Breiman [2001a] to solve regression and classification problems

- Among state-of-the-art methods for tabular data
- No need to precisely tune parameters
- Valuable in high-dimension settings
- Based on trees which are interpretable

Random forests are great!

Random forests are a class of algorithms created by Breiman [2001a] to solve regression and classification problems

- Among state-of-the-art methods for tabular data
- No need to precisely tune parameters
- Valuable in high-dimension settings
- Based on trees which are interpretable

- Difficult to analyze theoretically
- Difficult to interpret

General mathematical framework

Regression setting

We are given a training set $\mathcal{D}_n = \{(X_1, Y_1), ..., (X_n, Y_n)\}$ where the pairs $(X_i, Y_i) \in [0,1]^d \times \mathbb{R}$ are *i.i.d.* distributed as (X, Y) and

$$Y = m(\mathbf{X}) + \varepsilon,$$

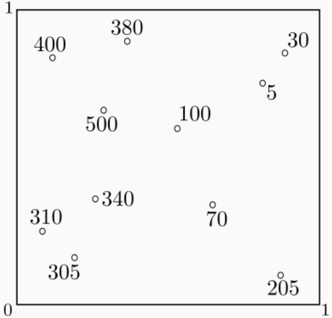
with $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$.

Aim: estimating the regression function m using random forests.

Random forests

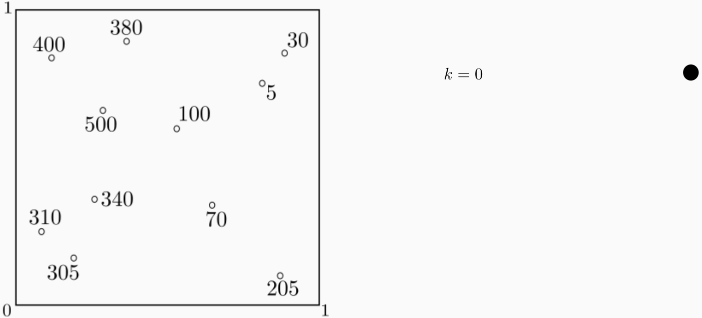
How to build a tree?

- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.



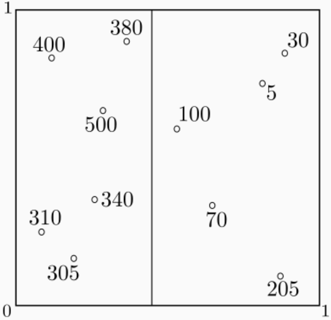
How to build a tree?

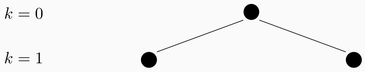
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.



How to build a tree?

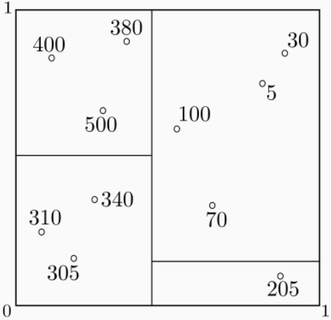
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

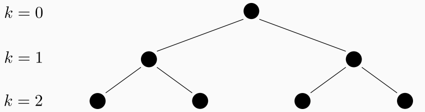




How to build a tree?

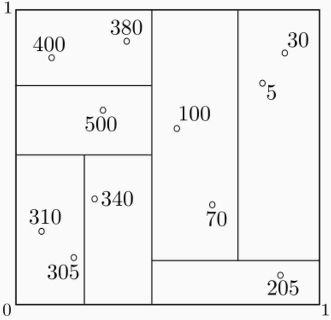
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

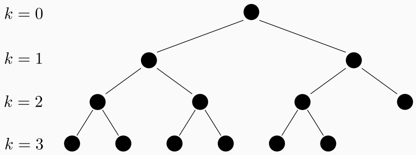




How to build a tree?

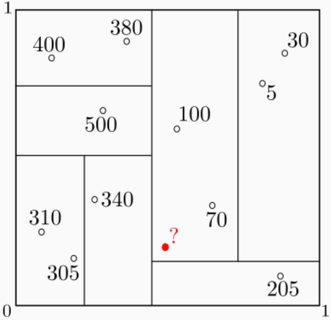
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

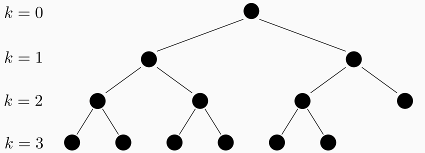




How to build a tree?

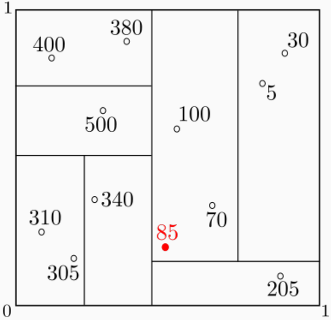
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.

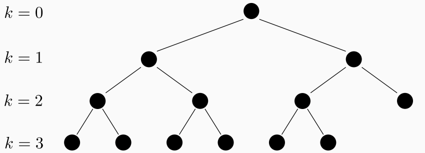




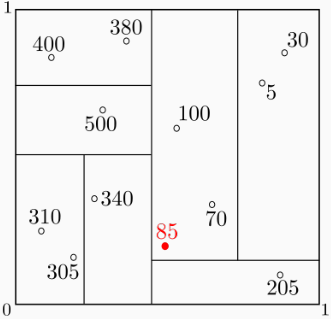
How to build a tree?

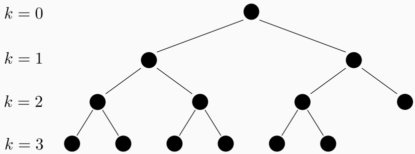
- Trees are built recursively by splitting the current cell into two children until some stopping criterion is satisfied.





How to build a tree?





Breiman Random forests are defined by

1. A splitting rule : minimize the variance within the resulting cells.
2. A stopping rule : stop when each cell contains less than `nodesize` $= 2$ observations.

How to perform splits?

For a split direction $j \in \{1, \ldots, d\}$ and a split position $z \in [0, 1]$, the criterion takes the form

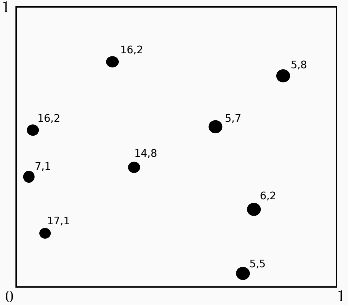
$$L_n(j, z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z} \right)^2,$$

where

- $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$ and $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$
- $\bar{Y}_A$ is the average of the Y_i's belonging to A.
- $N_n(A)$ is the number of points in A

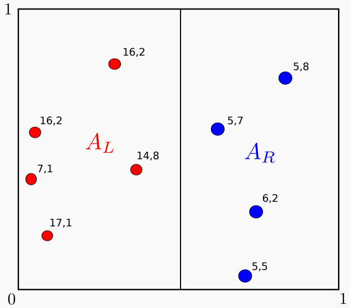
How to perform splits of Breiman's forests?

An example: $j = 1$ and $z = 0.5$.



How to perform splits of Breiman's forests?

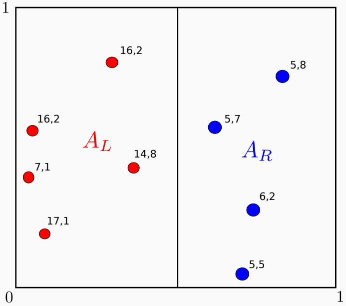
An example: $j = 1$ and $z = 0.5$.



$$L_n(1, 0.5) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \Big(Y_i - \underbrace{\bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(1)} < 0.5}}_{\text{Average on } A_L} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(1)} \geq 0.5} \Big)^2,$$

How to perform splits of Breiman's forests?

An example: $j = 1$ and $z = 0.5$.

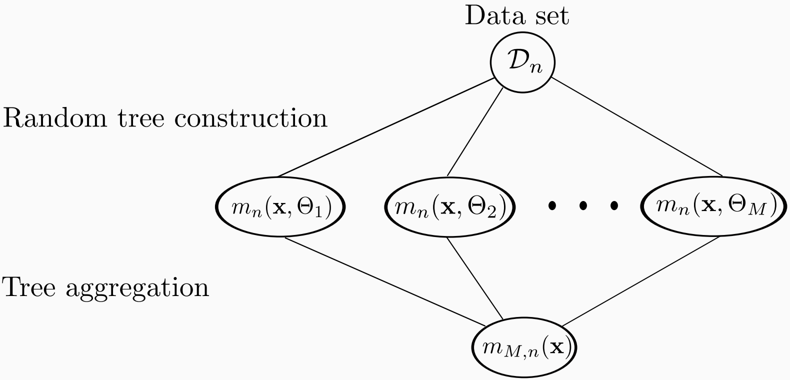


$$L_n(1, 0.5) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(1)} < 0.5} - \underbrace{\bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(1)} \geq 0.5}}_{\text{Average on } A_R} \right)^2,$$

Construction of random forests

Randomness in tree construction

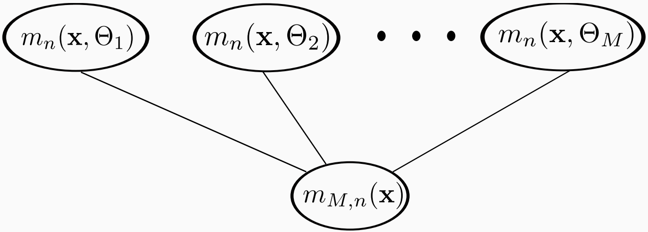
- Resampling the data set via bootstrap;
- For each cell:
    - Preselecting a subset of m_{try} variables, eligible for splitting.



Construction of Breiman forests

Breiman tree

- Select a_n observations with replacement among the original sample $\mathcal{D}_n$. Use only these observations to build the tree.
- For each cell,
  - Select randomly `mtry` coordinates among $\{1, \ldots, d\}$;
  - Choose the best split along previous direction, the one minimizing the CART criterion.
- Stop when each cell contains less than `nodesize` observations.



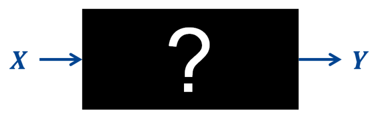
Literature

- Random forests were created by Breiman [2001a].
- Theoretical works started by focusing on convergence and upper bounds on the quadratic risk of a forest:
    - stylized forests, whose construction is independent of the dataset [Biau et al., 2008, Biau, 2012, Genuer, 2012, Zhu et al., 2015, Arlot and Genuer, 2014, Scornet, 2016b, Klusowski, 2018, Mourtada et al., 2020]
    - forest algorithms close to the original algorithm [Scornet et al., 2015, Scornet, 2016a, Wager and Walther, 2015]
- Another line of work: estimating the variance, proving the asymptotic normality of random forests:
  [Mentch and Hooker, 2016, Wager and Athey, 2017]
- Literature review on random forests:
    - Methodological review [Criminisi et al., 2011, Boulesteix et al., 2012],
    - Theoretical review [Biau and Scornet, 2016] with comments by S. Arlot, R. Genuer, P. Geurts, G. Hooker, L. Mentch, S. Wager, L. Wehenkel.

Interpretability

Existing Approaches

- Black-box models



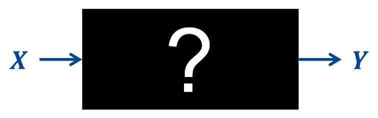
E.g. Neural networks, Random forests

Combined with post-processing
E.g. variable importance
sensitivity analysis
local linearization

Hard to operationalize

Existing Approaches

- Black-box models



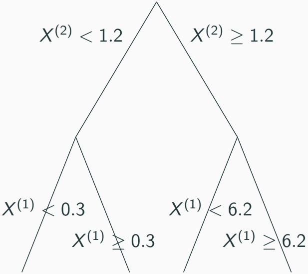
E.g. Neural networks, Random forests

Combined with post-processing
E.g. variable importance
sensitivity analysis
local linearization

Hard to operationalize

- Interpretable models

E.g. decision trees, decision rules



Unstable

Going beyong black-box nature of random forests

- Designing simple, interpretable and stable rules extracted from random forests: SIRUS
  - Interpretable Random Forests via Rule Extraction,
    by C. Bénard, G. Biau, S. Da Veiga, E. Scornet

- Variable importances in random forests:
  - Mean Decrease Accuracy (MDA) [Breiman, 2001a]

    MDA for random forests: inconsistency, and a practical solution via the Sobol-MDA,
    by C. Bénard, S. Da Veiga, E. Scornet
  - Mean Decrease Impurity (MDI) [Breiman, 2002]
    Subject of this talk

Variable importance in random forests: MDI

How to perform splits?

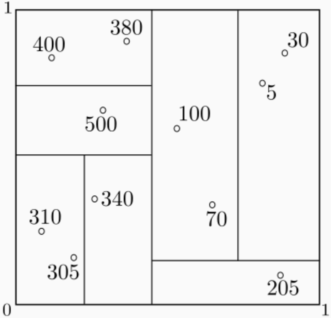
For a split direction $j \in \{1, \ldots, d\}$ and a split position $z \in [0, 1]$, the criterion takes the form

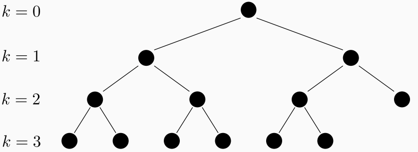
$$L_{n,A}(j, z) = \frac{1}{N_n(A)} \sum_{i=1}^{n} \left(Y_i - \bar{Y}_{A_L} \mathbb{1}_{\mathbf{X}_i^{(j)} < z} - \bar{Y}_{A_R} \mathbb{1}_{\mathbf{X}_i^{(j)} \geq z} \right)^2,$$

where

- $A_L = \{\mathbf{x} \in A : \mathbf{x}^{(j)} < z\}$ and $A_R = \{\mathbf{x} \in A : \mathbf{x}^{(j)} \geq z\}$
- $\bar{Y}_A$ is the average of the Y_i's belonging to A.
- $N_n(A)$ is the number of points in A

How to compute MDI?





The Mean Decrease in Impurity (MDI) for the variable $X^{(j)}$ computed via a tree $\mathcal{T}$ is defined by

$$\widehat{\mathrm{MDI}}_{\mathcal{T}}(X^{(j)}) = \sum_{\substack{A \in \mathcal{T} \\ j_{n,A}=j}} p_{n,A} L_{n,A}(j_{n,A}, z_{n,A}), \tag{1}$$

where the sum ranges over all cells A in $\mathcal{T}$ that are split along variable j and $p_{n,A}$ is the fraction of observations falling into A.

Literature

Empircally known flaws of MDI:

- favor variables with many categories [see, e.g., Strobl et al., 2007, Nicodemus, 2011]
- biased towards variables that possess a category having a high frequency [Nicodemus, 2011, Boulesteix et al., 2011]
- biased in presence of correlated features [Nicodemus and Malley, 2009]

Designing new tree building procedure: [Strobl et al., 2008, 2009].

Theory:

- Louppe et al. [2013]: study of theoretical MDI when all variables are categorical.
- Bias related to in-sample estimation [Li et al., 2019, Zhou and Hooker, 2019]

First result

Proposition [Scornet, 2020]

Let $\mathcal{T}_n$ be the CART tree, based on the data set $\mathcal{D}_n$. Then,

$$\widehat{\mathbb{V}[Y]} = \sum_{j=1}^{d} \widehat{\mathrm{MDI}}_{\mathcal{T}_n}(X^{(j)}) + R_n(\hat{m}_{\mathcal{T}_n}), \tag{2}$$

where $\hat{m}_{\mathcal{T}_n}$ is the estimate associated to $\mathcal{T}_n$.

- Valid for many tree building processes (telescopic sums)
- Relation between $\widehat{\mathrm{MDI}}$ and R^2:

$$R^2 = \frac{\sum_{j=1}^{d} \widehat{\mathrm{MDI}}_{\mathcal{T}_n}(X^{(j)})}{\widehat{\mathbb{V}[Y]}}$$

- MDI, computed with fully-grown trees is positively biased:

$$\lim_{n\to\infty} \sum_{j=1}^{d} \widehat{\mathrm{MDI}}_{\mathcal{T}_n}(X^{(j)}) = \mathbb{V}[m(\mathbf{X})] + \sigma^2.$$

Additive models

Definition: Additive model

The regression model writes $Y = \sum_{j=1}^{d} m_j(X^{(j)}) + \varepsilon$, where each m_j is continuous; ε is a Gaussian noise $\mathcal{N}(0, \sigma^2)$, independent of $\mathbf{X}$; and $\mathbf{X} \sim \mathcal{U}([0,1]^d)$.

Theorem [Additive model, Scornet, 2020]

Assume that the Additive Model holds. Let $\mathcal{T}_n$ be the empirical CART tree. Then, for all $\gamma > 0, \rho \in (0,1]$, there exists K such that, for all $k > K$, for all n large enough, with probability at least $1 - \rho$, for all j,

$$\left| \widehat{\mathrm{MDI}}_{\mathcal{T}_{n,k}}(X^{(j)}) - \mathbb{V}[m_j(X^{(j)})] \right| \leq \gamma.$$

Additive model - theoretical results

Theorem [Additive model, Scornet, 2020]

Assume that the Additive Model holds. Let $\mathcal{T}_n$ be the empirical CART tree. Then, for all $\gamma > 0, \rho \in (0,1]$, there exists K such that, for all $k > K$, for all n large enough, with probability at least $1 - \rho$, for all j,

$$\left| \widehat{\mathrm{MDI}}_{\mathcal{T}_{n,k}}(X^{(j)}) - \mathbb{V}[m_j(X^{(j)})] \right| \leq \gamma.$$

- MDI targets the same value as MDA (up to a constant 2).
- MDI targets the right quantity in an additive model with independent features.
  $\rightarrow$ MDI can be used to rank and select variables in this context
- MDI is consistent when computed with shallow trees.

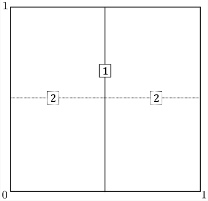
Moving beyond additivity

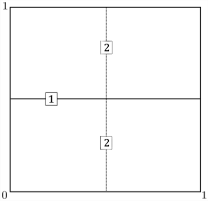
Model (Multiplicative model)

Let $\alpha \in \mathbb{R}$. The regression model writes $Y = 2^d \alpha \prod_{j=1}^{d} X^{(j)} + \varepsilon$, where $\alpha \in \mathbb{R}$; ε is a Gaussian noise $\mathcal{N}(0, \sigma^2)$, independent of $\mathbf{X}$; and $\mathbf{X} \sim \mathcal{U}([0,1]^d)$.

- This model contains interactions between all input variables
- There exists many theoretical trees

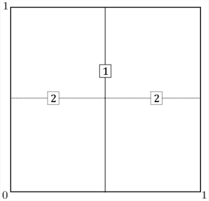
An example in dimension two:

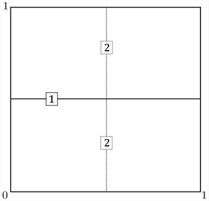




With interactions, important splits are not performed at the top of the tree

Two theoretical trees in the previous multiplicative model:

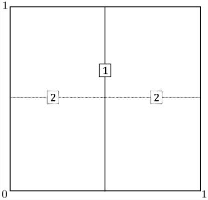


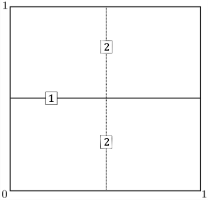


- In this example, the splits in the second level are associated with larger decreases in impurity/variance.
- In presence of interactions, the splits with the largest decreases in variance are not always in the first level of the tree!

A negative result in presence of interactions

Two theoretical trees in the previous multiplicative model:





Lemma [Scornet, 2020]

Assume that Model 1 holds. Then, there exists two theoretical trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that

$$\lim_{k\to\infty}\left(\mathrm{MDI}^{\star}_{\mathcal{T}_{2,k}}(X^{(1)}) - \mathrm{MDI}^{\star}_{\mathcal{T}_{1,k}}(X^{(1)})\right) = \alpha^2/16.$$

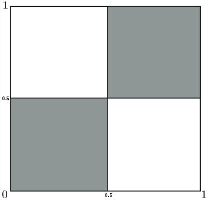
- MDI computed with a single tree is ill-defined

A correlation framework

Correlated Model

Let $\beta \in \mathbb{N}$. Assume that $Y = X^{(1)} + X^{(2)} + \alpha X^{(3)} + \varepsilon$, where $(X^{(1)}, X^{(2)}) \sim \mathcal{U}^{\otimes 2^\beta}$, $X^{(3)} \sim \mathcal{U}([0,1])$ is independent of $(X^{(1)}, X^{(2)})$, and ε is an independent noise distributed as $\mathcal{N}(0, \sigma^2)$.

The distribution $\mathcal{U}^{\otimes 2^\beta}$ is defined as $\mathcal{U}^{\otimes 2^\beta} = \mathcal{U}\left(\cup_{j=0}^{2^\beta - 1}\left[\frac{j}{2^\beta}, \frac{j+1}{2^\beta}\right)^2\right)$



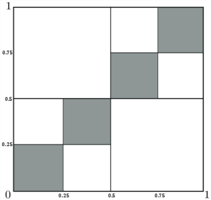


Figure 1: Illustration of $\mathcal{U}^{\otimes 2^\beta}$, with $\beta = 1$ (left) and $\beta = 2$ (right).

Correlated Model

Let $\beta \in \mathbb{N}$. Assume that $Y = X^{(1)} + X^{(2)} + \alpha X^{(3)} + \varepsilon$, where $(X^{(1)}, X^{(2)}) \sim \mathcal{U}^{\otimes 2^\beta}$, $X^{(3)} \sim \mathcal{U}([0,1])$ is independent of $(X^{(1)}, X^{(2)})$, and ε is an independent noise distributed as $\mathcal{N}(0, \sigma^2)$.

Lemma

Let $\beta \in \{0, \dots, 5\}$. Assume that the Correlated Model holds. Then, there exists two theoretical trees $\mathcal{T}_1$ and $\mathcal{T}_2$ such that

$$\lim_{k \to \infty} \left(MDI^{\star}_{\mathcal{T}_{2,k}}(X^{(1)}) - MDI^{\star}_{\mathcal{T}_{1,k}}(X^{(1)}) \right) = \frac{1}{3} - \frac{1}{3}\left(\frac{1}{4}\right)^{\beta}.$$

- Many theoretical trees exist.
- MDI computed with a single tree is ill-defined in this model (correlated design).

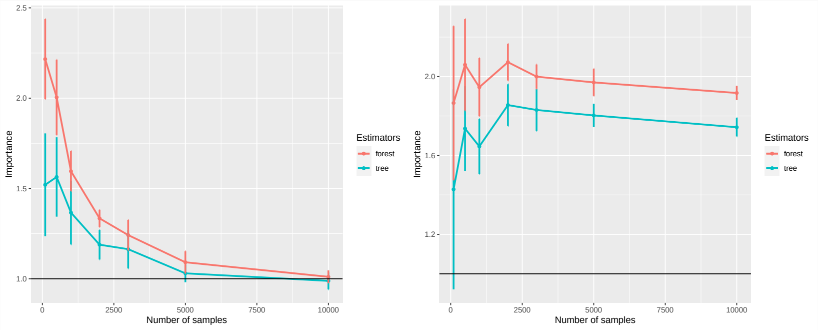
Experiments

We let $Y = \alpha_1 X^{(1)} + \alpha_2 X^{(2)} + \alpha_3 X^{(3)} + \varepsilon$, where ε is an independent noise, distributed as $\mathcal{N}(0, \sigma^2)$ and $(X^{(1)}, X^{(2)}, X^{(3)})$ is distributed as $\mathcal{N}(0, \Sigma)$ where

$$\Sigma = \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

For all j, we let $\alpha_j = \sqrt{j}$:

- The variable importance of the j-th component is j (for $\rho = 0$)
- Studying the impact of the noise σ^2 is easier in this setting.

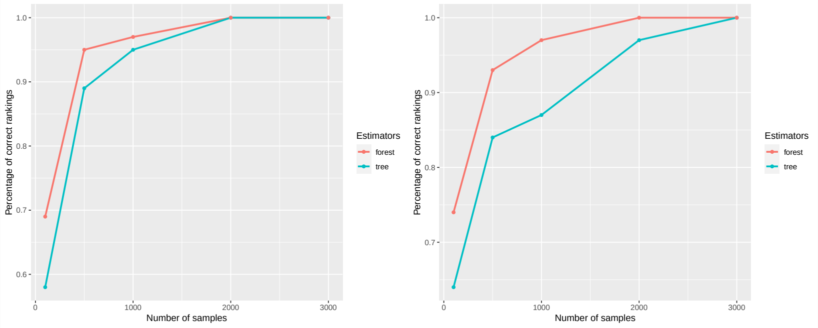


(a) **(b)**

Figure 2: Importance of the first variable in the previous simulated model, with $\sigma^2 = 3, \rho = 0$ and, from left to right $\texttt{maxnodes} = \lfloor n^{0.6} \rfloor, n$

In presence of noise, the MDI of the first variable is

- positively biased if computed with a fully-grown tree/forest.
- unibased if computed with an early-stopped tree/forest



(a) **(b)**

Figure 3: Percent of correct ranking in the previous simulated model, with $\sigma^2 = 12$ and, from left to right $\texttt{maxnodes} = \lfloor n^{0.6} \rfloor, n$

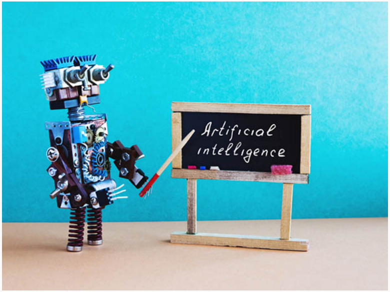
- Despite the fact that the MDIs are biased, the correct order is accurately retrieved.
- An early-stopped tree/forest produces more accurate rankings than a fully grown tree/forest.

Take-home messages

MDI analysis:

- If input variables are independent and in absence of interactions, using MDI to rank variable is ok
  - Proved for an early-stopped tree/forest
  - Empirically correct for fully-grown tree/forest
- In presence of correlation or interaction, the empirical MDI computed with a single tree does not converge, and therefore should not be used.
- In presence of correlation or interaction, the empirical MDI computed with a forest targets a quantity which is currently unknown.

Thank you!



References

S. Arlot and R. Genuer. Analysis of purely random forests bias. *arXiv:1407.3939*, 2014.

C. Bénard, G. Biau, S. Da Veiga, and E. Scornet. Sirus: making random forests interpretable. *arXiv preprint arXiv:1908.06852*, 2019.

G. Biau. Analysis of a random forests model. *Journal of Machine Learning Research*, 13:1063–1095, 2012.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.

G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9:2015–2033, 2008.

References II

A.-L. Boulesteix, A. Bender, J. Lorenzo Bermejo, and C. Strobl. Random forest gini importance favours snps with large minor allele frequency: impact, sources and recommendations. *Briefings in Bioinformatics*, 13:292–304, 2011.

A.-L. Boulesteix, S. Janitza, J. Kruppa, and I.R. König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2:493–507, 2012.

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

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16:199–231, 2001b.

L. Breiman. Manual on setting up, using, and understanding random forests v3. 1. *Statistics Department University of California Berkeley, CA, USA*, 1:58, 2002.

References III

A. Criminisi, J. Shotton, and E. Konukoglu. Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning. *Foundations and Trends in Computer Graphics and Vision*, 7:81–227, 2011.

F. Doshi-Velez and B. Kim. Towards a rigorous science of interpretable machine learning. *arXiv:1702.08608*, 2017.

R. Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24:543–562, 2012.

H. Ishwaran and U.B. Kogalur. *Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC)*, 2020. URL `https://cran.r-project.org/package=randomForestSRC`. R package version 2.9.3.

J.M. Klusowski. Sharp analysis of a simple model for random forests. *arXiv preprint arXiv:1805.02587*, 2018.

X. Li, Y. Wang, S. Basu, K. Kumbier, and B. Yu. A debiased mdi feature importance measure for random forests. In *Advances in Neural Information Processing Systems*, pages 8049–8059, New York, 2019.

References IV

A. Liaw and M. Wiener. Classification and regression by randomforest. *R News*, 2:18–22, 2002.

Z.C. Lipton. The mythos of model interpretability. *arXiv:1606.03490*, 2016.

G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*, pages 431–439, 2013.

T. A Mara, S. Tarantola, and P. Annoni. Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Environmental Modelling & Software*, 72:173–183, 2015.

Lucas Mentch and Giles Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *The Journal of Machine Learning Research*, 17(1):841–881, 2016.

J. Mourtada, S. Gaïffas, and E. Scornet. Minimax optimal rates for mondrian trees and forests. *Annals of Statistics*, 48(4):2253–2276, 2020.

References V

W.J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu. Interpretable machine learning: Definitions, methods, and applications. *arXiv:1901.04592*, 2019.

K. K. Nicodemus. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in bioinformatics*, 12(4):369–373, 2011.

K.K. Nicodemus and J.D. Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25:1884–1890, 2009.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

A. Saltelli. Making best use of model evaluations to compute sensitivity indices. *Computer Physics Communications*, 145:280–297, 2002.

References VI

E. Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146:72–83, 2016a.

E. Scornet. Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*, 2020.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43:1716–1741, 2015.

Erwan Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3):1485–1500, 2016b.

I.M. Sobol. Sensitivity estimates for nonlinear mathematical models. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC bioinformatics*, 8:25, 2007.

C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

References VII

C. Strobl, T. Hothorn, and A. Zeileis. Party on! 2009.

S. Wager and S. Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, (just-accepted), 2017.

S. Wager and G. Walther. Adaptive concentration of regression trees, with application to random forests. 2015.

M.N. Wright and A. Ziegler. ranger: a fast implementation of random forests for high dimensional data in c++ and r. *Journal of Statistical Software*, 77: 1–17, 2017.

B. Yu. Stability. *Bernoulli*, 19:1484–1500, 2013.

Z. Zhou and G. Hooker. Unbiased measurement of feature importance in tree-based methods. *arXiv preprint arXiv:1903.05179*, 2019.

R. Zhu, D. Zeng, and M.R. Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

SIRUS: Stable and Interpretable RUle Set

An example: SIRUS output on Titanic data set [Bénard et al., 2019]

Average survival rate $p_s = 39\%$.

if	`sex` is male	**then**	$p_s = 19\%$	**else**	$p_s = 74\%$
if	1^{st} or 2^{nd} `class`	**then**	$p_s = 56\%$	**else**	$p_s = 24\%$
if	1^{st} or 2^{nd} `class` **&** `sex` is female	**then**	$p_s = 95\%$	**else**	$p_s = 25\%$
if	`fare` < 10.5£	**then**	$p_s = 20\%$	**else**	$p_s = 50\%$
if	`no parents or children aboard`	**then**	$p_s = 35\%$	**else**	$p_s = 51\%$
if	2^{st} or 3^{nd} `class` **&** `sex` is male	**then**	$p_s = 14\%$	**else**	$p_s = 64\%$
if	`sex` is male **&** `age` ≥ 15	**then**	$p_s = 16\%$	**else**	$p_s = 72\%$

SIRUS

Principle

- Build a random forests and extract all decisions rules from all trees
- Select the rules that appear with a frequence larger than p_0
- Aggregate the rules to obtain the final estimator.

Principle

Frequent paths in random trees = strong and robust patterns in the data.

Technical detail

- Preprocessing: discretize features based on their quantiles
- Random forests: building trees of depth 2

Probability that a Θ-random tree contains a given path $\mathscr{P} \in \Pi$

$$p_n(\mathscr{P}) = \mathbb{P}(\mathscr{P} \in T(\Theta, \mathcal{D}_n) | \mathcal{D}_n)$$

Selected paths

$$\boxed{\hat{\mathscr{P}}_{M,n,p_0} = \{\mathscr{P} \in \Pi : \hat{p}_{M,n}(\mathscr{P}) > p_0\}}$$

where

$$\hat{p}_{M,n}(\mathscr{P}) = \frac{1}{M} \sum_{\ell=1}^{M} \mathbb{1}_{\mathscr{P} \in T(\Theta_\ell, \mathcal{D}_n)}$$

is the Monte-Carlo estimate, directly computed using the random forest with M trees parametrized by $\Theta_1, ..., \Theta_M$.

Stability - definition

Define

- $\mathcal{D}'_n$, Θ' independent copies of $\mathcal{D}_n$ and Θ
- $\hat{p}'_{M,n}(\mathscr{P})$, $\hat{\mathscr{P}}'_{M,n,p_0}$ built with $\mathcal{D}'_n$, Θ'

Dice-Sorensen index

$$\hat{S}_{M,n,p_0} = \frac{2\left|\hat{\mathscr{P}}_{M,n,p_0} \cap \hat{\mathscr{P}}'_{M,n,p_0}\right|}{\left|\hat{\mathscr{P}}_{M,n,p_0}\right| + \left|\hat{\mathscr{P}}'_{M,n,p_0}\right|}.$$

Stability - a theoretical result

(A1) The subsampling rate a_n satisfies $\lim_{n\to\infty} a_n = \infty$ and $\lim_{n\to\infty} \frac{a_n}{n} = 0$.

(A2) The number of trees M_n satisfies $\lim_{n\to\infty} M_n = \infty$.

(A3) $\mathbf{X}$ has a density f with respect to the Lebesgue measure, continuous, bounded, and strictly positive.

Let $\mathcal{U}^\star = \{p^\star(\mathscr{P}), \mathscr{P} \in \Pi\}$ be the set of all theoretical probabilities of appearance of all paths.

Proposition Bénard et al. [2019]

Assume that Assumptions (A1)-(A3) are satisfied. Then, provided $p_0 \in [0,1] \backslash \mathcal{U}^\star$, we have

$$\lim_{n\to\infty} \hat{S}_{M_n,n,p_0} = 1, \quad \text{in probability.}$$