A first interpretable approach: SIRUS

Why do we need interpretability?

Machine learning is used for **decision support**.

Predicting is not enough

Understanding predictions is vital

- for Machine learning to be accepted (sensible applications in health, justice, defense)
- To improve algorithms (e.g., detect unfairness and try to correct it)

Keywords: trust, transparency, accountability, fairness, ethics.

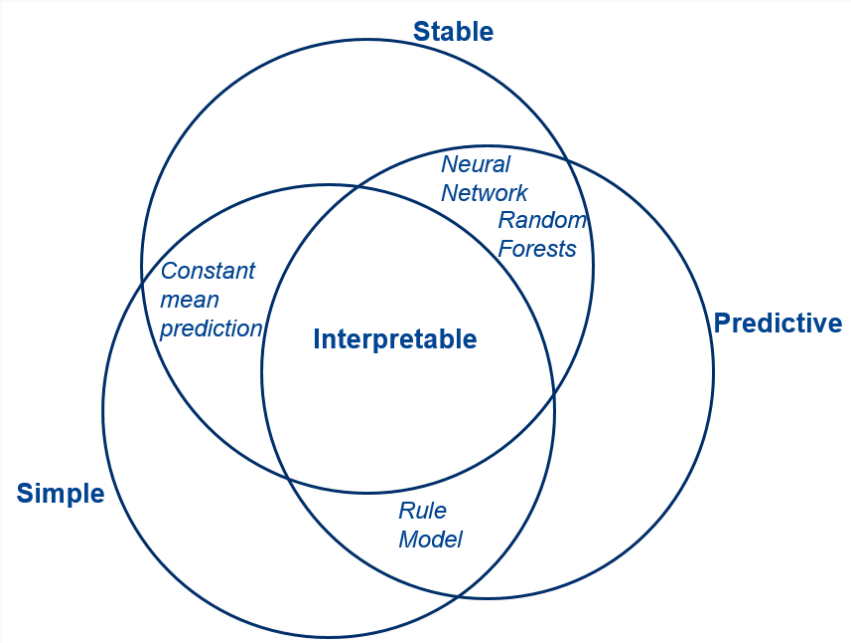
NIPS2017 debate: Interpretability is necessary for Machine learning

`https://www.youtube.com/watch?v=93Xv8vJ2acI`

Interpretable Models

- No agreement about a rigorous definition of interpretability [Lipton, 2016, Doshi-Velez and Kim, 2017, Murdoch et al., 2019]

- Minimum requirements for interpretability
  1. Simplicity [Murdoch et al., 2019]
  2. Stability [Yu, 2013]
  3. Predictivity [Breiman, 2001b]



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

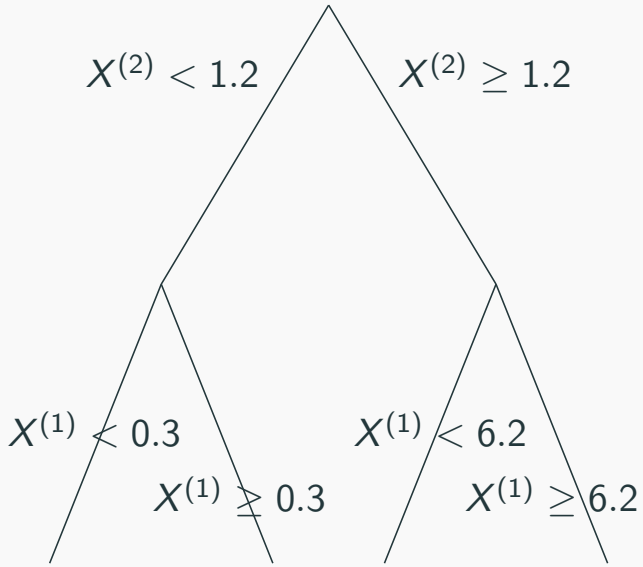
E.g. Neural networks, Random forests

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

Hard to operationalize

- Interpretable models

E.g. decision trees, decision rules



Unstable

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

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

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

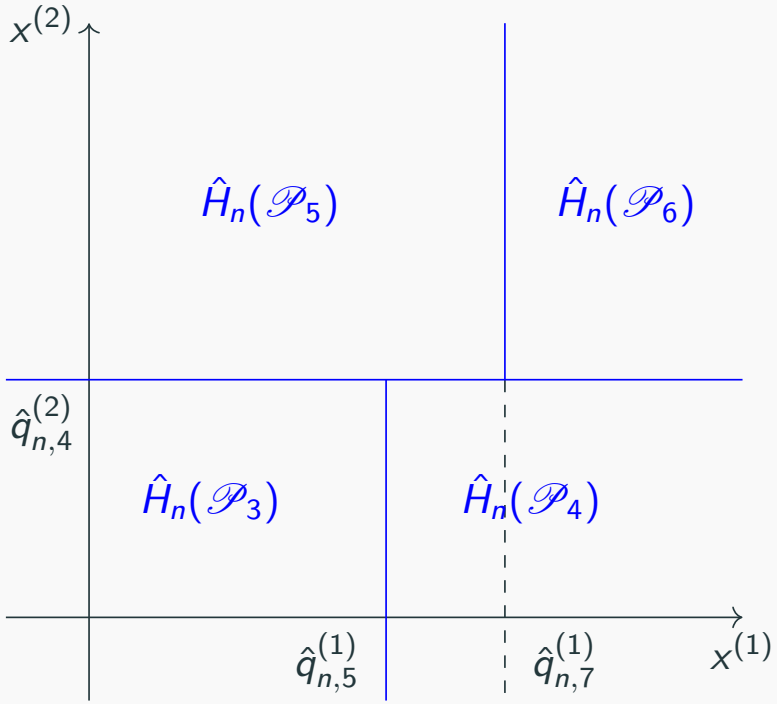
where

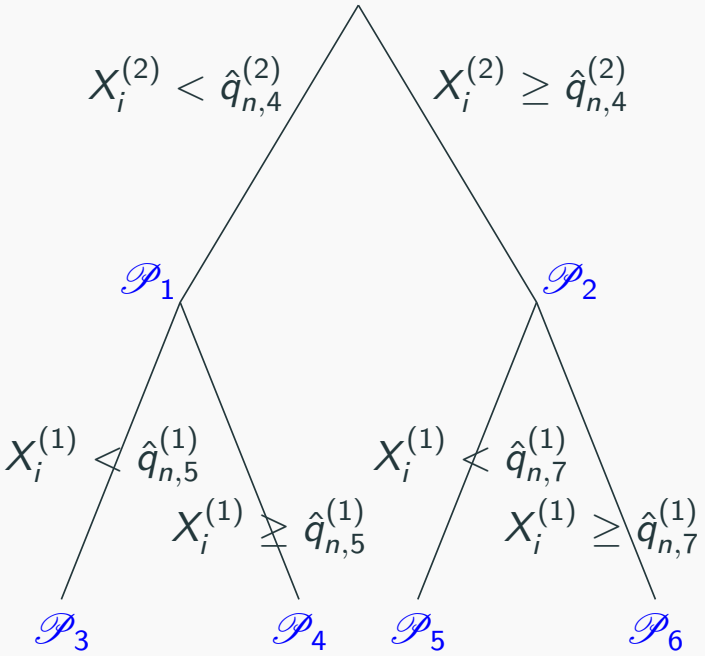
$$\hat{p}_{M,n}(\mathscr{P}) = \frac{1}{M} \sum_{\ell=1}^{M} \mathbb{1}_{\mathscr{P} \in T(\Theta_\ell, \mathcal{D}_n)}$$

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

SIRUS - Rule

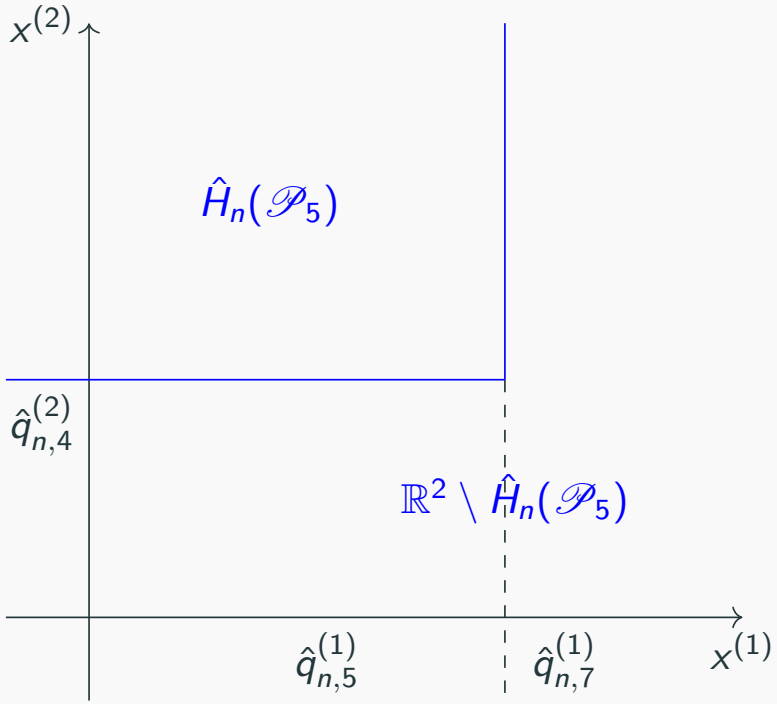
How to recover a rule from a path ?

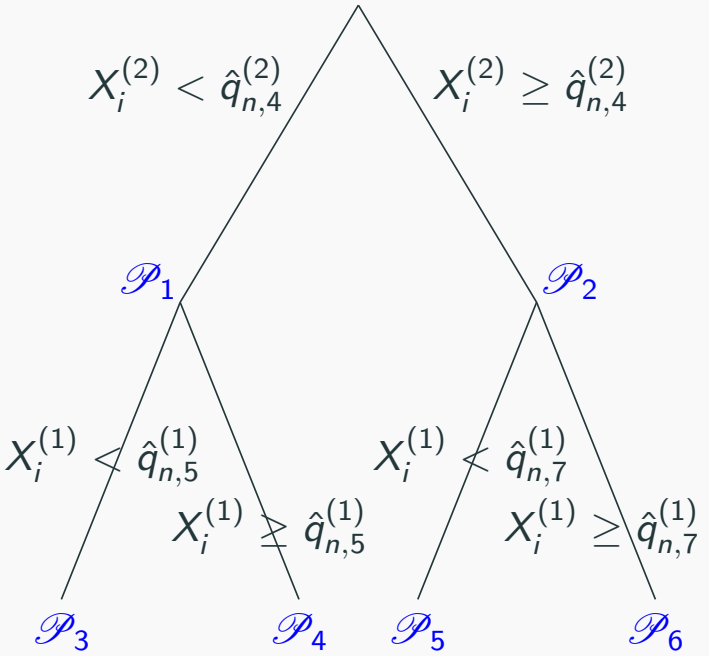




SIRUS - Rule

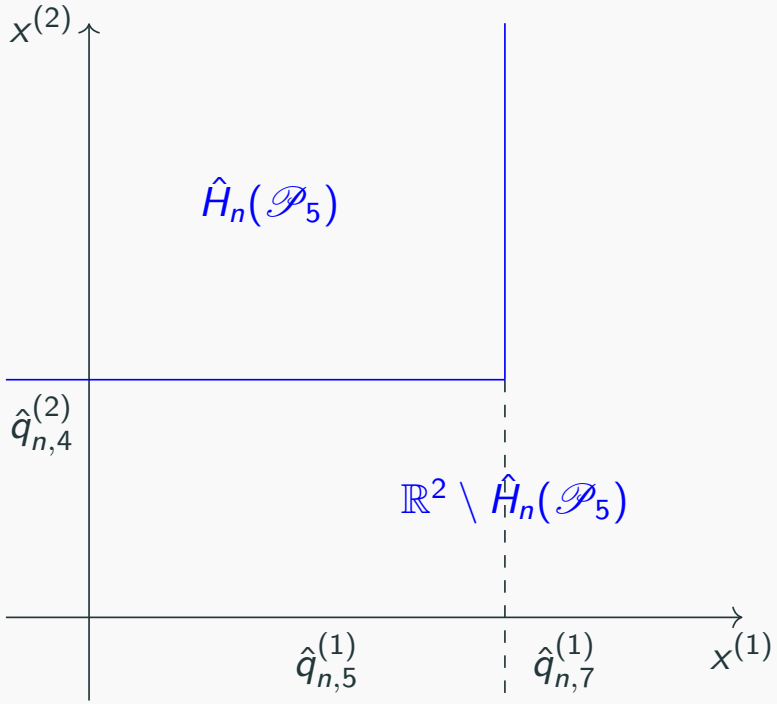
How to recover a rule from a path ?

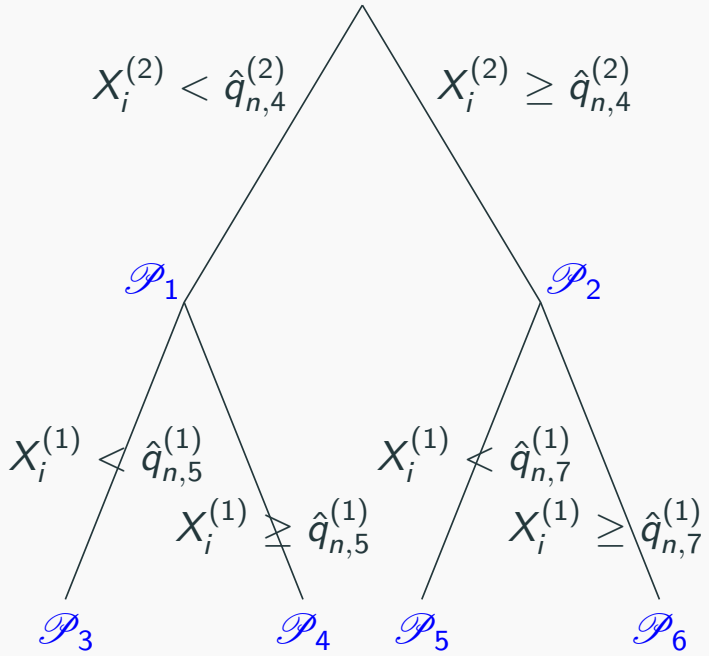




SIRUS - Rule

How to recover a rule from a path ?

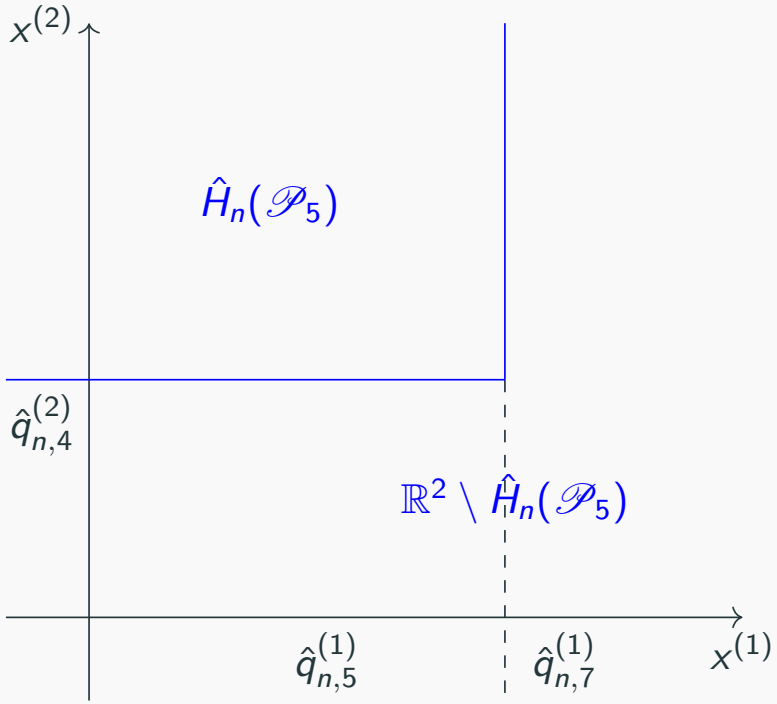


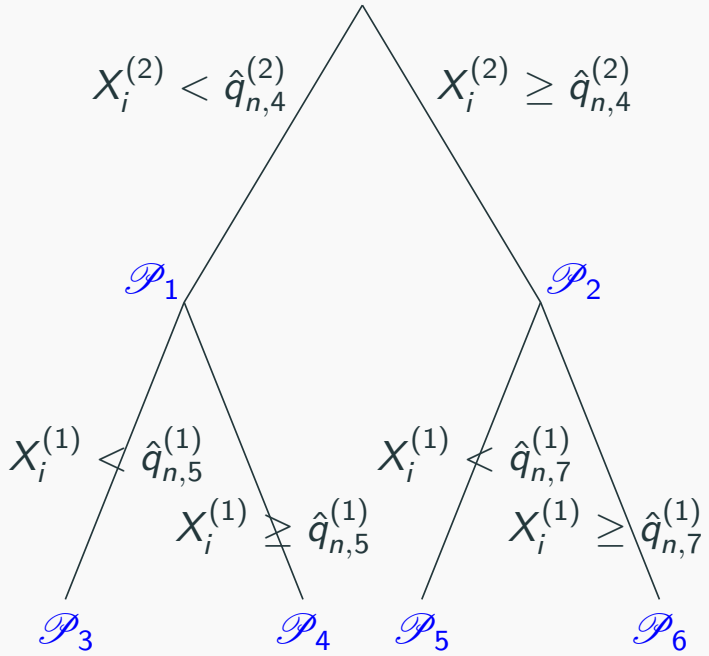


$$\forall \mathrm{x} \in \mathbb{R}^p, \quad \hat{g}_{n,\mathscr{P}}(\mathrm{x}) = \begin{cases} \frac{1}{N_n(\hat{H}_n(\mathscr{P}))} \sum_{i=1}^n Y_i \mathbb{1}_{\mathrm{X}_i \in \hat{H}_n(\mathscr{P})} & \text{if } \mathrm{x} \in \hat{H}_n(\mathscr{P}) \\ \frac{1}{n - N_n(\hat{H}_n(\mathscr{P}))} \sum_{i=1}^n Y_i \mathbb{1}_{\mathrm{X}_i \notin \hat{H}_n(\mathscr{P})} & \text{otherwise.} \end{cases}$$

SIRUS - Rule

How to recover a rule from a path ?





$$\forall \mathrm{x} \in \mathbb{R}^p, \quad \hat{g}_{n,\mathscr{P}}(\mathrm{x}) = \begin{cases} \frac{1}{N_n(\hat{H}_n(\mathscr{P}))} \sum_{i=1}^n Y_i \mathbb{1}_{\mathrm{X}_i \in \hat{H}_n(\mathscr{P})} & \text{if } \mathrm{x} \in \hat{H}_n(\mathscr{P}) \\ \frac{1}{n - N_n(\hat{H}_n(\mathscr{P}))} \sum_{i=1}^n Y_i \mathbb{1}_{\mathrm{X}_i \notin \hat{H}_n(\mathscr{P})} & \text{otherwise.} \end{cases}$$

The final classifier corresponds to the averaging of all selected rules.

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) 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

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

Sketch of proof

The asymptotic stability of SIRUS comes from the two following points:

1. The bias of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.

2. The variance of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.

Sketch of proof

The asymptotic stability of SIRUS comes from the two following points:

1. The bias of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.
    - Prove that **CART-splitting criterion** is **consistent** and **asymptotically normal** when cuts are limited to empirical quantiles and the number of trees grows with n (A3).
2. The variance of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.

Sketch of proof

The asymptotic stability of SIRUS comes from the two following points:

1. The bias of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.

   - Prove that **CART-splitting criterion** is **consistent** and **asymptotically normal** when cuts are limited to empirical quantiles and the number of trees grows with n (A3).

2. The variance of $\hat{p}_{M_n,n}(\mathscr{P})$ tends to zero.

   The variance can be decomposed into two terms:

   - the variance generated by the Monte-Carlo randomization, which goes to 0 as the number of trees increases (A2).
   - the variance of $p_n(\mathscr{P})$, which is a bagged estimate and thus an infinite-order U-statistic. The result comes from Mentch and Hooker [2016] since $\lim_{n\to\infty} a_n/n = 0$ (A1).

Numerical experiments

Competitors:

- CART [Breiman et al., 1984]
- Classical rule learning: RIPPER [Cohen, 1995]
- Frequent pattern mining: CBA [Classification Based on Association Rules, Liu et al., 1998], BRL [Bayesian Rule List, Letham et al., 2015]
- Tree ensemble: RuleFit [Friedman and Popescu, 2008], Node Harvest [Meinshausen, 2010].

Metrics:

- Accuracy/Error: 1-AUC
- Stability: Dice-Sorensen index
- Simplicity: Number of rules output by the procedure

Accuracy

	Black box	Decision tree	Classical rule learning	Frequent pattern mining		Tree ensemble		
Dataset	**Random Forest**	**CART**	**RIPPER**	**CBA**	**BRL**	**RuleFit**	**Node harvest**	**SIRUS**
Authentification	10^{-4}	0.02	0.02	0.14	0.009	**9.10^{-4}**	0.02	0.03
Breast Wisconsin	0.009	0.06	0.07	0.05	0.02	**0.01**	**0.01**	**0.01**
Credit Approval	0.07	0.14	0.15	0.14	0.11	**0.08**	**0.07**	0.09
Credit German	0.20	0.29	0.38	0.40	0.33	**0.23**	**0.26**	**0.25**
Diabetes	0.17	0.25	0.29	0.30	0.25	**0.18**	**0.19**	**0.19**
Haberman	0.31	0.48	0.39	0.50	0.43	**0.37**	**0.34**	**0.35**
Heart C2	0.10	0.19	0.23	0.17	0.23	0.12	0.12	**0.10**
Heart H2	0.11	0.23	0.24	0.24	0.16	**0.11**	**0.11**	**0.12**
Heart Statlog	0.10	0.20	0.21	0.17	0.22	0.12	0.12	**0.10**
Hepatitis	0.12	0.48	0.39	0.36	0.33	0.20	0.23	**0.17**
Ionosphere	0.02	0.11	0.12	0.13	0.10	**0.04**	0.07	0.07
Kr vs Kp	9.10^{-4}	0.02	**0.009**	0.05	0.01	**0.009**	0.04	0.04
Liver Disorders	0.23	0.33	0.35	0.48	0.44	**0.27**	0.30	0.35
Mushrooms	0	0.007	3.10^{-5}	5.10^{-4}	**2.10^{-5}**	5.10^{-4}	0.002	6.10^{-4}
Sonar	0.07	0.27	0.26	0.25	0.44	**0.12**	0.16	0.2
Spambase	0.01	0.11	0.08	0.12	0.05	**0.02**	0.04	0.07

Figure 1: Model error (1-AUC) over a 10-fold cross-validation for UCI datasets. Results are averaged over 10 repetitions of the cross-validation. Values within 10% of the minimum are displayed in bold, random forest is put aside.

Simplicity

	Decision tree	Classical rule learning	Frequent pattern mining		Tree ensemble		
Dataset	**CART**	**RIPPER**	**CBA**	**BRL**	**RuleFit**	**Node harvest**	**SIRUS**
Authentification	21	7	7	17	49	30	13
Breast Wisconsin	7	12	24	7	24	32	24
Credit Approval	5	4	55	4	15	27	16
Credit German	18	3	69	4	33	33	20
Diabetes	13	3	17	6	26	31	8
Haberman	2	1	2	2	3	17	5
Heart C2	10	3	34	4	23	36	20
Heart H2	5	2	29	3	12	24	12
Heart Statlog	10	3	27	4	22	35	16
Hepatitis	2	2	14	2	8	14	12
Ionosphere	4	4	38	4	20	35	15
Kr vs Kp	16	15	29	9	18	13	24
Liver Disorders	15	3	2	3	19	33	17
Mushrooms	4	8	25	11	10	22	23
Sonar	6	4	33	2	32	83	19
Spambase	14	16	126	16	68	60	21

Figure 2: Mean model size over a 10-fold cross-validation for UCI datasets. Results are averaged over 10 repetitions of the cross-validation.

Stability

	Decision tree	Classical rule learning	Frequent pattern mining		Tree ensemble		
Dataset	**CART**	**RIPPER**	**CBA**	**BRL**	**RuleFit**	**Node harvest**	**SIRUS**
Authentification	0.41	0.36	**0.87**	**0.86**	0.48	0.59	**0.81**
Breast Wisconsin	0.21	0.55	**0.80**	**0.78**	0.34	0.71	0.70
Credit Approval	0.52	0.32	0.43	0.52	0.25	0.23	**0.75**
Credit German	0.46	0.22	0.51	0.41	0.24	0.48	**0.66**
Diabetes	0.29	0.21	0.46	**0.73**	0.39	0.45	**0.81**
Haberman	**0.83**	0.09	**0.79**	0.50	0.46	0.52	0.65
Heart C2	0.25	0.35	0.38	0.60	0.39	0.49	**0.71**
Heart H2	0.46	0.27	0.52	**0.73**	0.29	0.29	**0.65**
Heart Statlog	0.30	0.41	0.41	**0.75**	0.35	0.48	**0.83**
Hepatitis	0.26	0.16	0.24	0.34	0.26	0.49	**0.68**
Ionosphere	**0.96**	0.39	0.13	0.70	0.17	0.33	0.69
Kr vs Kp	0.71	0.74	**0.84**	**0.80**	0.19	0.27	**0.87**
Liver Disorders	0.23	0.10	**0.91**	0.50	0.24	0.31	0.58
Mushrooms	**1**	0.84	**0.98**	0.80	0.69	0.48	0.86
Sonar	0.34	0.04	0.09	0.19	0.09	0.20	**0.55**
Spambase	0.49	0.10	0.46	**0.86**	0.28	0.66	**0.78**

Figure 3: Mean stability over a 10-fold cross-validation for UCI datasets. Results are averaged over 10 repetitions of the cross-validation. Values within 10% of the maximum are displayed in bold.

Conclusion on SIRUS

- Output a small, stable, and predictive set of rules
  - Predictive performances are on par with RF
  - Stability and number of rules improved over state-of-the-art algorithms
- Theoretical guarantees of stability for SIRUS
- Relies heavily on quantile discretization and a limited tree depth

Post-hoc methods: Sobol indices and Shapley effects

MDA

- MDA [Breiman, 2001a]: built-in variable importance algorithm for random forests

MDA

- MDA [Breiman, 2001a]: built-in variable importance algorithm for random forests

- MDA is used intensively

MDA

- MDA [Breiman, 2001a]: built-in variable importance algorithm for random forests

- MDA is used intensively

- MDA has flaws
  - Poor understanding of the MDA: what is estimated ?
  - Empirical studies show that the MDA is biased for dependent inputs [Strobl et al., 2007, Gregorutti et al., 2017, Hooker and Mentch, 2019]

MDA

- MDA [Breiman, 2001a]: built-in variable importance algorithm for random forests

- MDA is used intensively

- MDA has flaws
  - Poor understanding of the MDA: what is estimated ?
  - Empirical studies show that the MDA is biased for dependent inputs [Strobl et al., 2007, Gregorutti et al., 2017, Hooker and Mentch, 2019]

- Our objective [Bénard et al., 2021]
  - Theoretical analysis of the MDA
    - First convergence result for the original MDA [Ishwaran, 2007, Zhu et al., 2015]
    - Theoretical understanding of MDA bias
  - Design of Sobol-MDA algorithm to fix the MDA flaws

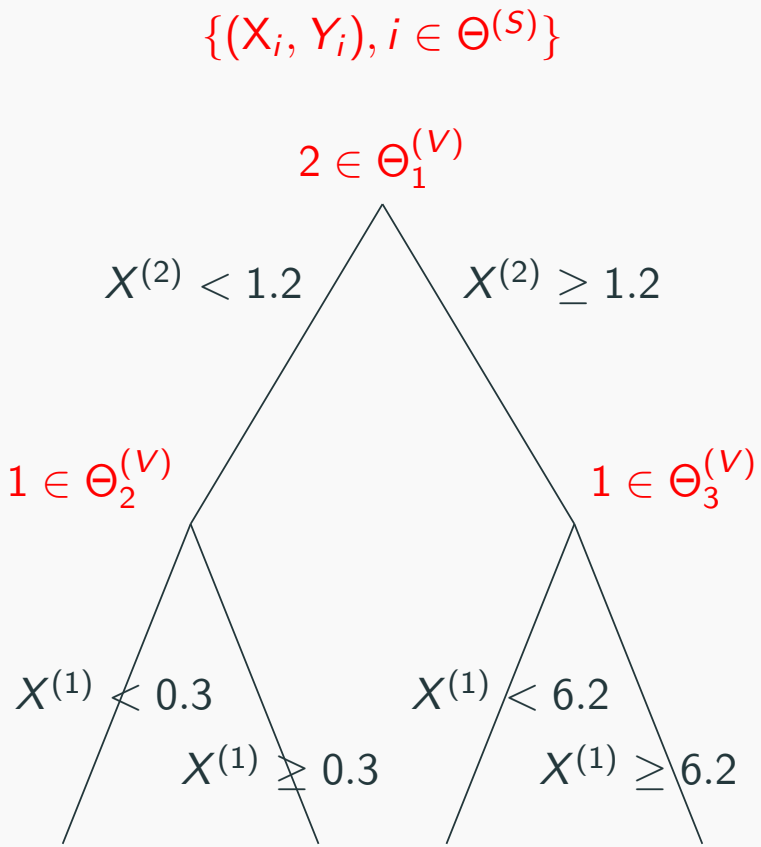
Random forests

- Regression setting
  - input vector $\mathsf{X} = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$
  - output $Y \in \mathbb{R}$
  - dataset $\mathcal{D}_n = \{(\mathsf{X}_i, Y_i), i = 1, \ldots, n\}$,
    where $(\mathsf{X}_i, Y_i) \sim \mathbb{P}_{\mathsf{X},Y}$.

Random forests

- Regression setting
  - input vector $\mathsf{X} = (X^{(1)}, \ldots, X^{(p)}) \in \mathbb{R}^p$
  - output $Y \in \mathbb{R}$
  - dataset $\mathcal{D}_n = \{(\mathsf{X}_i, Y_i), i = 1, \ldots, n\}$, where $(\mathsf{X}_i, Y_i) \sim \mathbb{P}_{\mathsf{X},Y}$.

- Random forest algorithm
  - Aggregation of Θ-random trees $\Theta = (\Theta^{(S)}, \Theta^{(V)})$
  - M: number of trees
  - $m_{M,n}(\mathsf{X}, \Theta_M)$: the forest estimate at X



MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest
3. permute randomly the values of a given input variable $X^{(j)}$:
   break the dependence between $X^{(j)}$ and Y

MDA principle

MDA principle:

decrease of accuracy of the forest when a variable is noised up

1. fit a random forest with $\mathcal{D}_n$
2. compute the accuracy of the forest
3. permute randomly the values of a given input variable $X^{(j)}$:
   break the dependence between $X^{(j)}$ and Y
4. compute the decrease of accuracy of the forest with the permuted data

MDA illustration

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

Table 1: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

Table 1: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table 1: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table 1: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

Explained variance of Y $= 16.4$

Explained variance of Y $= 13.7$

$$\mathrm{MDA}(X^{(j)}) = 16.4 - 13.7 = 2.7$$

MDA illustration

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	0.1	...	2.6	2.3
1.7	4.1	...	9.2	...	3.8	0.4
3.4	9.2	...	3.2	...	3.6	10.2
5.6	1.2	...	8.2	...	4.2	9.1
8.9	6.8	...	6.7	...	2.9	4.5

$X^{(1)}$	$X^{(2)}$	...	$X^{(j)}$	...	$X^{(p)}$	Y
2.1	4.3	...	6.7	...	2.6	2.3
1.7	4.1	...	3.2	...	3.8	0.4
3.4	9.2	...	9.2	...	3.6	10.2
5.6	1.2	...	0.1	...	4.2	9.1
8.9	6.8	...	8.2	...	2.9	4.5

Table 1: Example of the permutation of a dataset $\mathcal{D}_n$ for $n = 5$.

Question: Can I use $\mathcal{D}_n$ to both fit the forest and compute accuracy ?

No: overfitting and inflated accuracy.

How to handle this in practice?

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

Selected samples: $\Theta_\ell^{(S)} = \{1, 3, 4\}$

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

OOB samples: $\{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$

MDA versions

The explained variance estimate of MDA algorithms differ across implementations

Train-Test MDA: train data to fit the forest, and test data for accuracy

Out-of-bag (OOB) samples: $\mathcal{D}_n$ is bootstrap prior to the construction of each tree, leaving aside a portion of $\mathcal{D}_n$, which is not involved in the tree growing and defines the "out-of-bag" sample.

MDA Version	Package	Error	Data
Train-Test	`scikit-learn` `randomForestSRC`	Forest	Testing dataset
Breiman-Cutler	`randomForest` (normalized) `ranger` / `randomForestSRC`	Tree	OOB sample
Ishwaran-Kogalur	`randomForestSRC`	Forest	OOB sample

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathsf{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	9.2	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	6.7	$\ldots$	2.9	4.5

X_i

$X^{(1)}$	$X^{(2)}$	$\ldots$	$X^{(j)}$	$\ldots$	$X^{(p)}$	Y
2.1	4.3	$\ldots$	0.1	$\ldots$	2.6	2.3
1.7	4.1	$\ldots$	6.7	$\ldots$	3.8	0.4
3.4	9.2	$\ldots$	3.2	$\ldots$	3.6	10.2
5.6	1.2	$\ldots$	8.2	$\ldots$	4.2	9.1
8.9	6.8	$\ldots$	9.2	$\ldots$	2.9	4.5

$\mathsf{X}_{i,\pi_{j\ell}}$

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathrm{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

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

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathrm{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

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

Quadratic risk of the ℓ-th tree

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathrm{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

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

Inflated quadratic risk of the ℓ-th tree where $X^{(j)}$ is permuted

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathrm{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

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

Risks are computed over the OOB sample of each tree

Breiman-Cutler MDA

- $i \in \{1, \ldots, n\} \setminus \Theta_\ell^{(S)} = \{2, 5\}$: OOB sample of the ℓ-th tree
- $N_{n,\ell} = \sum_{i=1}^{n} \mathbb{1}_{i \neq \Theta_\ell^{(S)}} = 2$: size of the OOB sample of the ℓ-th tree
- $\mathrm{X}_{i,\pi_{j\ell}}$: i-th observation where the j-th component is permuted across the OOB sample of the ℓ-th tree

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

Average over all trees

Assumptions

(A1)

The response $Y \in \mathbb{R}$ *follows*

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

where

- $\mathrm{X} = (X^{(1)}, \ldots, X^{(p)}) \in [0,1]^p$
- X *admits a density* f *such that* $c_1 < f(\mathrm{x}) < c_2$, *with constants* $c_1, c_2 > 0$
- m *is continuous*
- *the noise* ε *is sub-Gaussian and centered*

Assumptions

(A2): the theoretical tree is consistent

(always true with slight modifications of the forest algorithm)

Assumptions

(A2): the theoretical tree is consistent

(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of (X, Y) *is consistent, that is, for all* $x \in [0,1]^p$*, almost surely,*

$$\lim_{k \to \infty} \Delta(m, A_k^\star(x, \Theta)) = 0.$$

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of (X, Y) *is consistent, that is, for all* $x \in [0,1]^p$*, almost surely,*

$$\lim_{k \to \infty} \Delta(m, A_k^\star(x, \Theta)) = 0.$$

(A3): tree partition is not too complex with respect to n

Assumptions

(A2): the theoretical tree is consistent
(always true with slight modifications of the forest algorithm)

(A2)

The randomized theoretical CART tree built with the distribution of (X, Y) *is consistent, that is, for all* $x \in [0,1]^p$*, almost surely,*

$$\lim_{k \to \infty} \Delta(m, A_k^\star(x, \Theta)) = 0.$$

(A3): tree partition is not too complex with respect to n

(A3)

The asymptotic regime of a_n*, the size of the subsampling without replacement, and the number of terminal leaves* t_n *is such that* $a_n \leq n-2$, $a_n/n < 1-\kappa$ *for a fixed* $\kappa > 0$, $\lim_{n\to\infty} a_n = \infty$, $\lim_{n\to\infty} t_n = \infty$, *and* $\lim_{n\to\infty} t_n \frac{(\log(a_n))^9}{a_n} = 0$.

MDA Convergence

Theorem (Bénard et al. [2021])

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

$$\widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{E}[(m(\mathsf{X}) - m(\mathsf{X}_{\pi_j}))^2]$$

X_{π_j}: X where the j-th component is replaced by an independent copy, i.e. $\mathsf{X}_{\pi_j} = (X^{(1)}, \dots, X'^{(j)}, \dots, X^{(p)})$

Limit interpretation?

Sensitivity analysis

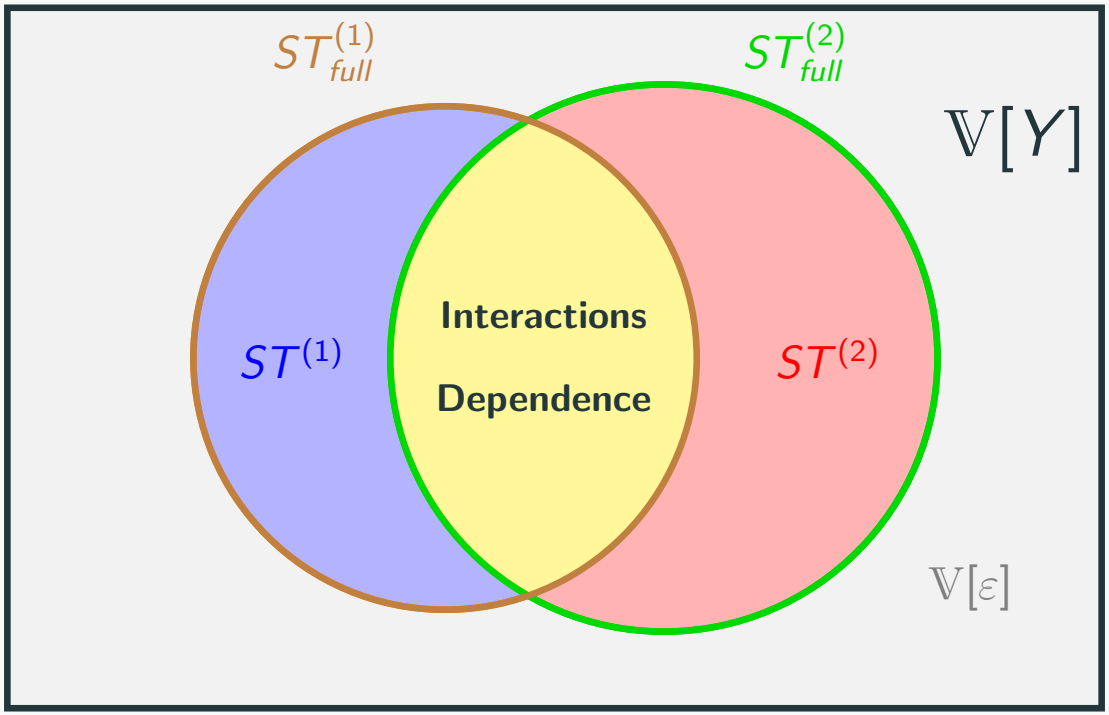


Figure 4: Standard and full total Sobol indices for $Y = m(X^{(1)}, X^{(2)}) + \varepsilon$.

Total Sobol index [Sobol, 1993]

$$ST^{(1)} = \frac{\mathbb{E}[\mathbb{V}(m(\mathrm{X})|\mathrm{X}^{(-1)})]}{\mathbb{V}(Y)}$$

Full total Sobol index [Mara et al., 2015, Benoumechiara, 2019]

$$ST^{(1)}_{full} = \frac{\mathbb{E}[\mathbb{V}(m(\mathrm{X}_{\pi_j})|\mathrm{X}^{(-1)})]}{\mathbb{V}(Y)}$$

MDA Decomposition

Proposition (Bénard et al. [2021])

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

$$\widehat{MDA}^{(BC)}_{M,n}(X^{(j)}) \xrightarrow{\mathbb{L}^1} \mathbb{V}[Y] \times ST^{(j)} + \mathbb{V}[Y] \times ST^{(j)}_{full} + MDA^{\star(j)}_3.$$

The term $\mathrm{MDA}^{\star(j)}_3$ is not an importance measure and is defined by

$$\mathrm{MDA}^{\star(j)}_3 = \mathbb{E}[(\mathbb{E}[m(\mathrm{X})|\mathrm{X}^{(-j)}] - \mathbb{E}[m(\mathrm{X}_{\pi_j})|\mathrm{X}^{(-j)}])^2].$$

MDA Decomposition

Proposition (Bénard et al. [2021])

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

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

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

If additionally $M \longrightarrow \infty$, *then*

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

Independent inputs

If inputs X are independent: $\mathrm{MDA}_3^{\star(j)} = 0$ and $ST^{(j)} = ST_{full}^{(j)}$.

Corollary (Bénard et al. [2021])

If X has independent components, and if Assumptions (A1)-(A3) are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \ldots, p\}$ we have

$$\widehat{MDA}_{M,n}^{(TT)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} 2\mathbb{V}[Y] \times ST^{(j)}$$
$$\widehat{MDA}_{M,n}^{(BC)}(X^{(j)}) \xrightarrow{\mathbb{L}^1} 2\mathbb{V}[Y] \times ST^{(j)}.$$

If additionally $M \longrightarrow \infty$, then

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

This Corollary completes the result from [Gregorutti, 2015].

Additive regression function

If m is additive: $\mathrm{MDA}_3^{\star(j)} = 0$.

Corollary (Bénard et al. [2021])

If the regression function m is additive, and if Assumptions (A1)-(A3) are satisfied, for all $M \in \mathbb{N}^\star$ and $j \in \{1, \ldots, p\}$ we have

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

If additionally $M \longrightarrow \infty$, then

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

MDA summary

- When inputs X are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.

MDA summary

- When inputs X are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !

MDA summary

- When inputs X are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !
- For variable selection, the total Sobol index is the relevant component

$$\mathbb{V}[Y] \times ST^{(j)} + \cancel{\mathbb{V}[Y] \times ST^{(j)}_{full}} + \cancel{\mathrm{MDA}_3^{\star(j)}}$$

MDA summary

- When inputs X are dependent and have interactions, the MDA is artificially inflated by the term MDA_3 and is therefore misleading.
- MDA versions have different theoretical counterparts, and thus different meanings: be careful when using forest packages !
- For variable selection, the total Sobol index is the relevant component

$$\mathbb{V}[Y] \times ST^{(j)} + \cancel{\mathbb{V}[Y] \times ST^{(j)}_{full}} + \cancel{\mathrm{MDA}_3^{\star(j)}}$$

- We develop the Sobol-MDA: a fast and consistent estimate of $ST^{(j)}$ for random forests

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m_{M,n}^{(-j,OOB)}(\mathrm{X}_i^{(-j)}, \Theta_M)\right]^2 - \left[Y_i - m_{M,n}^{(OOB)}(\mathrm{X}_i, \Theta_M)\right]^2$$

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

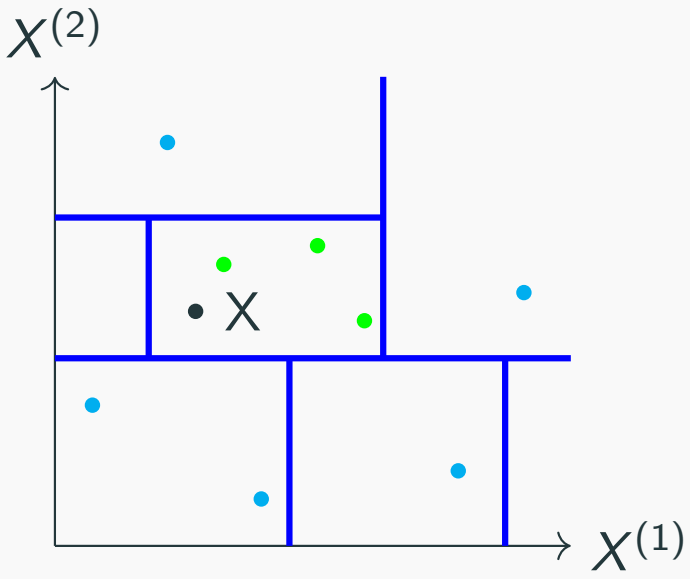


Figure 5: Partition of $[0,1]^2$ by a random tree (left side) projected on the subspace span by $\mathrm{X}^{(-2)} = X^{(1)}$ (right side), for $p = 2$ and $j = 2$.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m_{M,n}^{(-j,OOB)}(\mathrm{X}_i^{(-j)}, \Theta_M)\right]^2 - \left[Y_i - m_{M,n}^{(OOB)}(\mathrm{X}_i, \Theta_M)\right]^2$$

Sobol-MDA

Principle: **project** the partition of each tree along the j-th direction to remove $X^{(j)}$ from the prediction process.

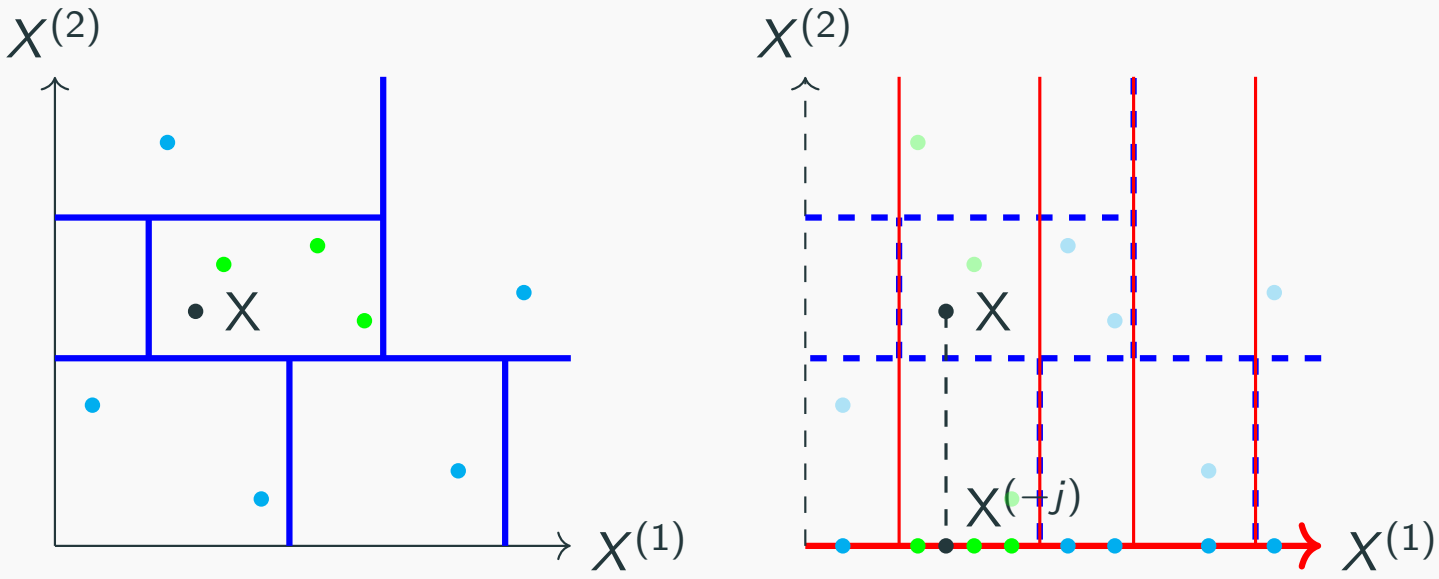


Figure 5: Partition of $[0,1]^2$ by a random tree (left side) projected on the subspace span by $\mathrm{X}^{(-2)} = X^{(1)}$ (right side), for $p = 2$ and $j = 2$.

$$\widehat{\text{S-MDA}}_{M,n}(X^{(j)}) = \frac{1}{\hat{\sigma}_Y^2}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i - m_{M,n}^{(-j,OOB)}(\mathrm{X}_i^{(-j)}, \Theta_M)\right]^2 - \left[Y_i - m_{M,n}^{(OOB)}(\mathrm{X}_i, \Theta_M)\right]^2$$

Consistency of the Sobol-MDA

The Sobol-MDA recovers the appropriate theoretical counterpart for variable selection: the total Sobol index

Theorem (Bénard et al. [2021])

If Assumptions (A1), (A2'), and (A3') are satisfied, for all $M \in \mathbb{N}^\star$ *and* $j \in \{1, \ldots, p\}$

$$\widehat{S\text{-}MDA}_{M,n}(X^{(j)}) \xrightarrow{p} ST^{(j)}.$$

Sobol-MDA Experiments

Settings [Archer and Kimes, 2008, Gregorutti et al., 2017]

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated

Sobol-MDA Experiments

Settings [Archer and Kimes, 2008, Gregorutti et al., 2017]

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated
- 1 variable from each group involved in m

$$m(\mathsf{X}) = 2X^{(1)} + X^{(41)} + X^{(81)} + X^{(121)} + X^{(161)}.$$

- independent Gaussian noise with $\mathbb{V}[\varepsilon] = 10\%\mathbb{V}[Y]$

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

Sobol-MDA Experiments

Settings [Archer and Kimes, 2008, Gregorutti et al., 2017]

- $p = 200$ input variables
- 5 independent groups of 40 variables
- each group is a Gaussian vector, strongly correlated
- 1 variable from each group involved in m

$$m(\mathsf{X}) = 2X^{(1)} + X^{(41)} + X^{(81)} + X^{(121)} + X^{(161)}.$$

- independent Gaussian noise with $\mathbb{V}[\varepsilon] = 10\%\mathbb{V}[Y]$

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

- $n = 1000$ observations
- $M = 300$ trees

Sobol-MDA Experiments

$\widehat{\text{S-MDA}}$		$\widehat{\text{BC-MDA}}/2\mathbb{V}[Y]$		$\widehat{\text{IK-MDA}}/\mathbb{V}[Y]$	
$X^{(1)}$	0.035	$X^{(1)}$	0.048	$X^{(1)}$	0.056
$X^{(161)}$	0.005	$X^{(25)}$	0.010	$X^{(5)}$	0.009
$X^{(81)}$	0.004	$X^{(31)}$	0.008	$X^{(81)}$	0.007
$X^{(121)}$	0.004	$X^{(14)}$	0.008	$X^{(41)}$	0.005
$X^{(41)}$	0.002	$X^{(40)}$	0.007	$X^{(161)}$	0.005
$X^{(179)}$	0.002	$X^{(3)}$	0.007	$X^{(15)}$	0.005
$X^{(13)}$	0.001	$X^{(17)}$	0.006	$X^{(121)}$	0.005
$X^{(25)}$	0.001	$X^{(26)}$	0.006	$X^{(7)}$	0.005
$X^{(73)}$	0.001	$X^{(41)}$	0.006	$X^{(4)}$	0.004
$X^{(155)}$	0.001	$X^{(121)}$	0.006	$X^{(28)}$	0.004

Table 3: Sobol-MDA, normalized BC-MDA, and normalized IK-MDA estimates with influential variables in blue.

Additional Experiments

Additional experiments are available in Bénard et al. [2021] (non-linear data with interactions and dependence)

- analytical example
- backward variable selection with real data

Extension

Sobol-MDA can be associated with any black-box algorithm

- fit a black box $\hat{f}$ on $\mathcal{D}_n$
- generate a large sample $\mathcal{D}'_N$ with $\hat{f}$
- run the Sobol-MDA with $\mathcal{D}'_N$

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]
- Attribute the value produced by a joint team to its individual members

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]
- Attribute the value produced by a joint team to its individual members
- Difference of produced value between a subset of the team and the same subteam with an additional member (averaged over all possible subteams).

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]
- Attribute the value produced by a joint team to its individual members
- Difference of produced value between a subset of the team and the same subteam with an additional member (averaged over all possible subteams).
- Adapted by Owen [2014] to variable importance in machine learning:

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]
- Attribute the value produced by a joint team to its individual members
- Difference of produced value between a subset of the team and the same subteam with an additional member (averaged over all possible subteams).
- Adapted by Owen [2014] to variable importance in machine learning:
  - member of the team = input variable

Definition of Shapley effects

- Originally defined in economics and game theory [Shapley, 1953]
- Attribute the value produced by a joint team to its individual members
- Difference of produced value between a subset of the team and the same subteam with an additional member (averaged over all possible subteams).
- Adapted by Owen [2014] to variable importance in machine learning:
  - member of the team = input variable
  - value function = explained output variance

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^{\star}(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]}{\mathbb{V}[Y]}.$$

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^{\star}(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]}{\mathbb{V}[Y]}.$$

Main property: equitably allocate contributions due to dependence and interactions across input variables

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^\star(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]}{\mathbb{V}[Y]}.$$

Main property: equitably allocate contributions due to dependence and interactions across input variables

Two obstacles arise to estimate Shapley effects:

1. the computational complexity is exponential with the dimension p

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^\star(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]}{\mathbb{V}[Y]}.$$

Main property: equitably allocate contributions due to dependence and interactions across input variables

Two obstacles arise to estimate Shapley effects:

1. the computational complexity is exponential with the dimension p
2. $\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]$ requires a fast and accurate estimate for all variable subsets $U \subset \{1, \dots, p\}$

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^\star(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|X^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|X^{(U)}]]}{\mathbb{V}[Y]}.$$

Main property: equitably allocate contributions due to dependence and interactions across input variables

Two obstacles arise to estimate Shapley effects:

1. the computational complexity is exponential with the dimension p
   Literature: Monte-Carlo methods
2. $\mathbb{V}[\mathbb{E}[Y|X^{(U)}]]$ requires a fast and accurate estimate for all variable subsets $U \subset \{1, \dots, p\}$

Definition of Shapley effects

Formally, the Shapley effect of the j-th variable is defined by

$$Sh^\star(X^{(j)}) = \sum_{U \subset \{1,\dots,p\} \setminus \{j\}} \frac{1}{p} \binom{p-1}{|U|}^{-1} \frac{\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U \cup \{j\})}]] - \mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]}{\mathbb{V}[Y]}.$$

Main property: equitably allocate contributions due to dependence and interactions across input variables

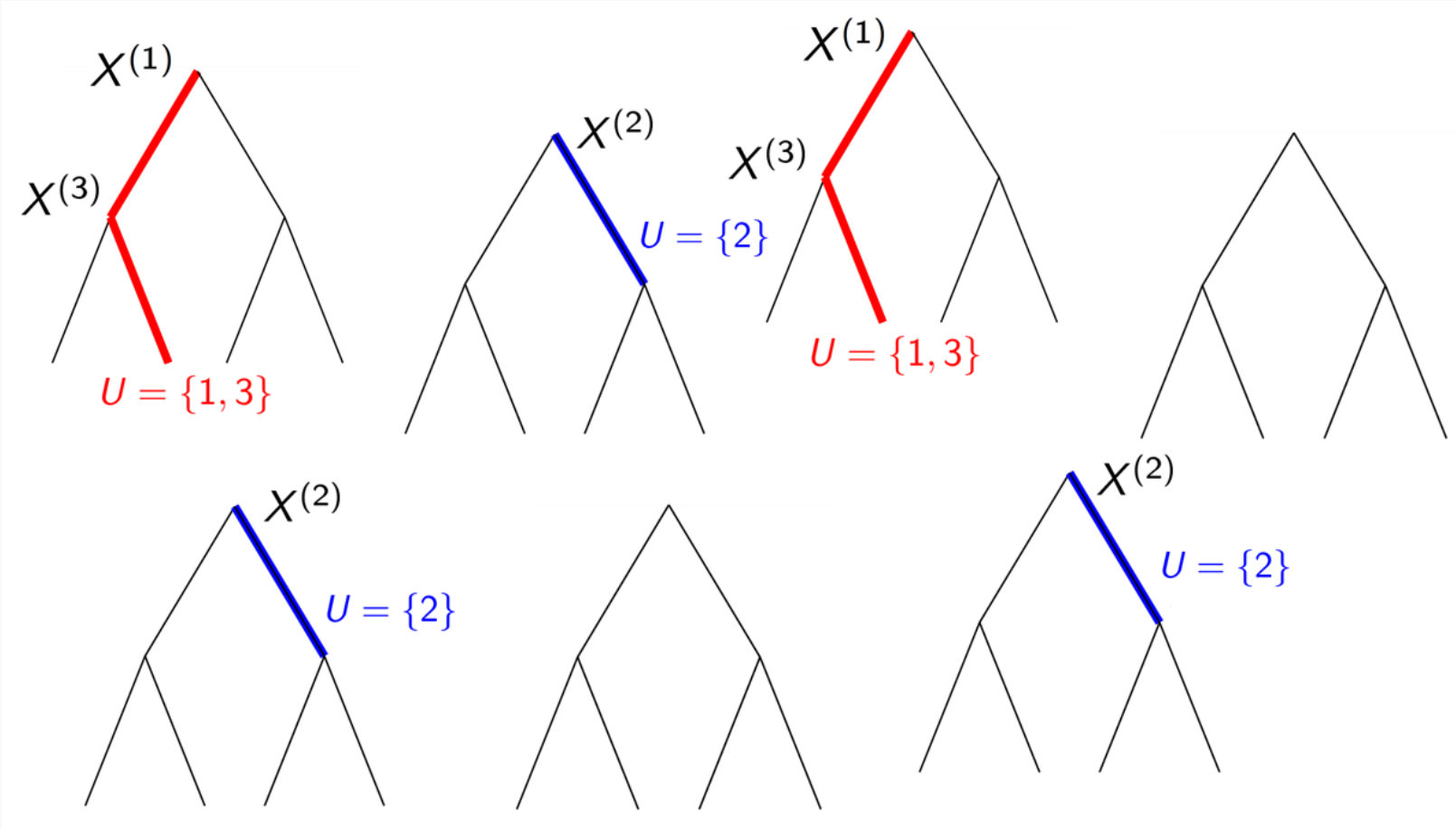
Two obstacles arise to estimate Shapley effects:

1. the computational complexity is exponential with the dimension p
   Literature: Monte-Carlo methods
2. $\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]$ requires a fast and accurate estimate for all variable subsets $U \subset \{1, \dots, p\}$
   Literature: strong approximation of the conditional distributions

SHAFF: SHApley efFects via random Forests

SHAFF proceeds in three steps:

1. sample many subsets U, typically a few hundreds, based on their occurrence frequency $\hat{p}_{M,n}(U)$ in the random forest



SHAFF: SHApley efFects via random Forests

SHAFF proceeds in three steps:

1. sample many subsets U, typically a few hundreds, based on their occurrence frequency $\hat{p}_{M,n}(U)$ in the random forest
2. estimate $\mathbb{V}[\mathbb{E}[Y|X^{(U)}]]$ with the projected forest algorithm for all selected U and their complementary sets $\{1,\dots,p\} \setminus U$: $\hat{v}_{M,n}(U)$

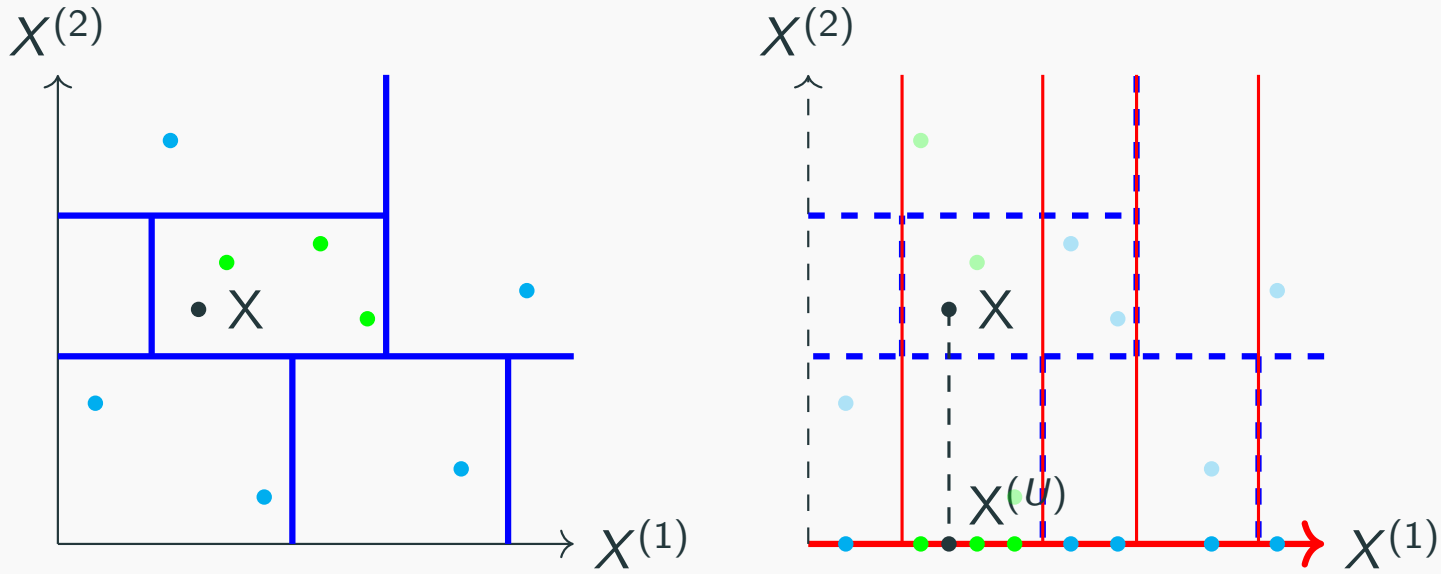


Figure 6: Partition of $[0,1]^2$ by a random tree (left side) projected on the subspace span by $X^{(U)} = X^{(1)}$ (right side), for $p = 2$ and $U = \{1\}$.

SHAFF: SHApley efFects via random Forests

SHAFF proceeds in three steps:

1. sample many subsets U, typically a few hundreds, based on their occurrence frequency $\hat{p}_{M,n}(U)$ in the random forest
2. estimate $\mathbb{V}[\mathbb{E}[Y|\mathrm{X}^{(U)}]]$ with the projected forest algorithm for all selected U and their complementary sets $\{1,\dots,p\} \setminus U$: $\hat{v}_{M,n}(U)$
3. solve a weighted linear regression problem to recover Shapley effects $\hat{\mathrm{Sh}}_{M_n,n}$ by minimizing in β

$$\ell_{M,n}(\beta) = \frac{1}{K} \sum_{U \in \mathcal{U}_{n,K}} \frac{w(U)}{\hat{p}_{M,n}(U)} (\hat{v}_{M,n}(U) - \beta^T I(U))^2,$$

where $w(U) = \frac{p-1}{\binom{p}{|U|}|U|(p-|U|)}$ and $I(U)$ is the binary vector of dimension p where the j-th component takes the value 1 if $j \in U$ and 0 otherwise.

SHAFF consistency

(A4)

The number of Monte-Carlo sampling K_n and the number of trees M_n grow with n, such that $M_n \longrightarrow \infty$ and $n.M_n/K_n \longrightarrow 0$.

Theorem

If Assumptions (A1), (A2'), (A3'), and (A4) are satisfied, then SHAFF *is consistent, that is*

$$\hat{\mathrm{Sh}}_{M_n,n} \xrightarrow{p} \mathrm{Sh}^{\star}.$$

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity
- Sobol-MDA fixes the flaws of original MDA
- R/C++ package `SobolMDA`, available online on Gitlab (https://gitlab.com/drti/sobolmda), and based on the package `ranger`

Conclusion

- Strong connections between the MDA and Sobol indices
- MDA does not target the appropriate quantity
- Sobol-MDA fixes the flaws of original MDA
- R/C++ package `SobolMDA`, available online on Gitlab (https://gitlab.com/drti/sobolmda), and based on the package `ranger`
- SHAFF: generalization of projected random forests to Shapley effects
- R/C++ package `shaff`, available online on Gitlab (https://gitlab.com/drti/shaff), and based on the package `ranger`

Questions ?

References

K.J. Archer and R.V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics & Data Analysis*, 52:2249–2260, 2008.

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

C. Bénard, S. Da Veiga, and E. Scornet. Mda for random forests: inconsistency, and a practical solution via the sobol-mda. *arXiv preprint arXiv:2102.13347*, 2021.

N. Benoumechiara. *Treatment of dependency in sensitivity analysis for industrial reliability*. PhD thesis, Sorbonne Université ; EDF R&D, 2019.

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

References ii

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

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Chapman & Hall, New York, 1984.

William W Cohen. Fast effective rule induction. In *Machine learning proceedings 1995*, pages 115–123. Elsevier, 1995.

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

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

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

B. Gregorutti, B. Michel, and P. Saint-Pierre. Correlation and variable importance in random forests. *Statistics and Computing*, 27:659–678, 2017.

References iii

G. Hooker and L. Mentch. Please stop permuting features: an explanation and alternatives. *arXiv preprint arXiv:1905.03151*, 2019.

H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

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

Bing Liu, Wynne Hsu, Yiming Ma, et al. Integrating classification and association rule mining. In *Kdd*, volume 98, pages 80–86, 1998.

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.

References iv

Nicolai Meinshausen. Node harvest. *The Annals of Applied Statistics*, pages 2049–2072, 2010.

L. Mentch and G. Hooker. Quantifying uncertainty in random forests via confidence intervals and hypothesis tests. *Journal of Machine Learning Research*, 17:841–881, 2016.

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

A.B. Owen. Sobol'indices and shapley value. *SIAM/ASA Journal on Uncertainty Quantification*, 2:245–251, 2014.

L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2:307–317, 1953.

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

References v

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(1):25, 2007.

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

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

Sparsity / interaction order

Interpretable without sparsity:

- Philosophical question?
- RF do not work well in the additive nonsparse context
- Therefore SIRUS does not work either

Sparsity

- SIRUS inherits sparsity properties of RF
- Sparsity has an impact on the required number of rules (which is automatically chosen based on an accuracy value).

Interaction order:

- RF can miss complex signals
- Sirus can detect only interactions of order two (see IRF and signed iterative RF)
- Can we adapt SIRUS to handle high-level interactions (they are masked by low-level interactions in the current version)? Modify the probability to encourage high-level interactions.

RF modification:

- Can we adapt RF to predict only when the output is larger than a threshold?
- Can we adapt the splitting criterion to focus on these regions (by adapting the pinball loss used for conditional quantiles)?