Adaptative Concentration of Regression Trees

Gaspard Beugnot, Célia Escribe, Céline Moucer

12 mars 2019

Plan

1. Theoretical framework and main theorem
   - State of the art
   - Theoretical Framework
   - Main result
2. Theoritical development
   - Proof's sketch
   - Leaves'approximation
3. Consistency guarantees for random forests in high-dimensionnal setting
   - Guess-and-check forest procedure
   - Point-wise consistency theorem
   - Proof's sketch

What we know about random forests

- Decorrelate multiple trees by adding randomness
- Two stages approach : model selection and model fitting
- Raises the problem of post-selection inference

Classical inference	Post Selection Inference
Model	Data
$\downarrow$	$\downarrow$
Data	Model
$\downarrow$	$\downarrow$
Inference	Inference

When the model is chosen after looking to the variables, it becomes random itself ! So far :

- Fixed dimensional asymptotic guarantee
- Simplified procedure using a holdout set

Some definitions

Recursive partitioning

Starting from a parent node $\nu = [0, 1]^d$, we select a currently unsplit node $\nu \subseteq \mathbb{R}^d$, a splitting variable $j \in \{1, \ldots, d\}$ and a threshold $\tau \in [0, 1]$, and then splitting ν into two children $\nu_- = \nu \cap \{x : x_j \leq \tau\}$ and $\nu_+ = \nu \cap \{x : x_j > \tau\}$. The final leaf nodes generated by this algorithm, denoted by L, form a partition Λ of $[0, 1]^d$.

Valid partition

A partition Λ is $\{\alpha, k\}$-*valid* if it can by generated by a recursive partitioning scheme in which :

- each node contains at least a fraction α of the data points in its parent node
- each leaf contains at least k training examples for some $k \in \mathbb{N}$

Given a dataset $\mathcal{X}$, the set of $\{\alpha, k\}$-valid partitions is $\mathcal{V}_{\alpha, k}(\mathcal{X})$.

Valid and partition optimal trees

A valid partition induces a *valid tree*

$$T_\Lambda : [0, 1]^d \to \mathbb{R}, \quad T_\Lambda(x) = \frac{1}{|\{X_i : X_i \in L(x)\}|} \sum_{\{i : X_i \in L(x)\}} Y_i. \tag{1}$$

The set of all $\{\alpha, k\}$-valid trees T_Λ with $\Lambda \in \mathcal{V}_{\alpha, k}(\mathcal{X})$ is $\mathcal{T}_{\alpha, k}(\mathcal{X})$. Given a partition Λ, we define the *partition-optimal tree* as

$$T^*_\Lambda : [0, 1]^d \to \mathbb{R}, \quad T^*_\Lambda(x) = \mathbb{E}\left[Y | X \in L(x)\right], \tag{2}$$

where (X, Y) is a new random sample from our data-generating distribution.

We turn trees into forests by averaging multiple trees. Variance reduction of the forest improves as the correlation between trees decreases.

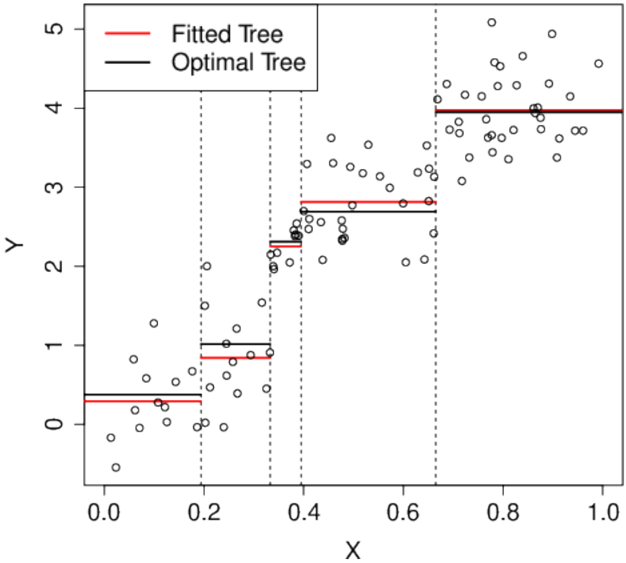

Figure – The theorem aims at providing a consistency guarantee between fitted decision tree and the optimal tree, *given* a certain partition Λ *(taken from the article)*.

Main result

Theorem

Suppose that we have n training examples $(X_i,\ Y_i) \in [0,\ 1]^d \times [-M,\ M]$ satisfying (A1), and that we have a sequence of problems with parameters $(n,\ d,\ k)$ satisfying (A2). Then, sample averages over all possible valid partitions concentrate around their expectations with high probability :

$$\lim_{n,\, d,\, k \to \infty} \mathbb{P}\left[\sup_{x \in [0,\, 1]^d,\, \Lambda \in \mathcal{V}_{\alpha,\, k}} \left| T_\Lambda(x) - T^*_\Lambda(x) \right| \leq 9M \sqrt{\frac{\log(n/k)\left(\log(dk) + 3\log\log(n)\right)}{\log\left((1-\alpha)^{-1}\right)}}\frac{1}{\sqrt{k}} \right] = 1. \tag{3}$$

Main result (2)

Theorem

In a moderately high-dimensional regime with $\liminf d/n > 0$*, the bound simplifies to*

$$\mathbb{P}\left[\sup_{x\in[0,1]^d,\,\Lambda\in\mathcal{V}_{\alpha,\,k}} |T_\Lambda(x) - T^*_\Lambda(x)| \leq 9M\sqrt{\frac{\log(n)\log(d)}{\log\left((1-\alpha)^{-1}\right)}\frac{1}{\sqrt{k}}}\right] \to 1. \tag{4}$$

Assumptions

Assumption 1 : Weakly dependant features

We have n independent and identically distributed training examples, whose features $X \in [0, 1]^d$ are distributed according to a density $f(\cdot)$ satisfying $\zeta^{-1} \leq f(x) \leq \zeta$ for all $x \in [0, 1]^d$, and some constant $\zeta \geq 1$.

Assumption 2 : Minimum leaf size

The minimum leaf-size k grows with n at a rate bounded from below by

$$\lim_{n \to \infty} \frac{\log(n) \max\{\log(d), \log\log(n)\}}{k} = 0. \tag{5}$$

Outline

A strong result ?

Given a single Tree $T \in \mathcal{T}_{\alpha,\, k}\left(\mathcal{X}\right)$ non-adaptively, i.e., without looking at the labels Y_i, a simple Hoeffding bound where we take n as a crude upper for the total number of leaves shows that

$$\mathbb{P}\left[\sup_{x \in [0,\,1]^d} \left|T\left(x\right) - T^*\left(x\right)\right| \leq M\sqrt{\frac{2.1\log\left(n\right)}{k}}\right] \rightarrow 1. \tag{6}$$

Uniforme concentration bound : a factor $\mathcal{O}(\sqrt{\log(d)})$ weaker

Post-selection inference interpretation

1 Creating a design matrix :

$$A \in \{0, 1\}^{n \times m}, \text{ where } A_{ij} = 1\left(\{X_i \in L_j\}\right),$$

2 Optimal regression vector :

$$\beta^*_A = \left(A^\top A\right)^{-1} A^\top \mu_i \text{ where } \mu_i = \mathbb{E}\left[Y_i \,\middle|\, A_i\right];$$

3 Berk and al. : PoSI constant between $\mathcal{O}_p(\sqrt{\log(d)})$ and $\mathcal{O}_p(\sqrt{d})$;

Proof's sketch

1 Leaves'approximation under Lebesgue Measure
2 Leave's approximation under the empirical measure
3 Proof of the Theorem

AIM : bounding large deviations of the process

$$\frac{1}{|\{i : X_i \in L\}|} \sum_{\{i:X_i \in L\}} Y_i - \mathbb{E}\left[Y \,\middle|\, X \in L\right], \tag{7}$$

Leaves approximation under Lebesgue measure

Theorem

Let $S \in \{1, \ldots, d\}$ be a set of size $|S| = s$, and let $w, \varepsilon \in (0, 1)$. Then, there exists a set of rectangles $\mathcal{R}_{S,w,\varepsilon}$ such that the following properties hold. Any rectangle R with support $S(R) \subseteq S$ and of volume $\lambda(R) \geq w$ can be well approximated by elements in $\mathcal{R}_{S,w,\varepsilon}$ from both above and below in terms of Lebesgue measure. Specifically, there exist rectangles $R_-, R_+ \in \mathcal{R}_{S,w,\varepsilon}$ such that

$$R_- \subseteq R \subseteq R_+, \quad \text{and} \quad e^{-\varepsilon}\lambda(R_+) \leq \lambda(R) \leq e^{\varepsilon}\lambda(R_-). \tag{8}$$

Moreover, the set $\mathcal{R}_{S,w,\varepsilon}$ has cardinality bounded by

$$|\mathcal{R}_{S,w,\varepsilon}| \leq \frac{1}{w}\left(\frac{8s^2}{\varepsilon^2}\left(1 + \log_2\left\lfloor\frac{1}{w}\right\rfloor\right)\right)^s \cdot (1 + \mathcal{O}(\varepsilon)). \tag{9}$$

Leaves'approximation under Lebesgue measure

Corollaire

Suppose that we set

$$w = \frac{1}{2\zeta}\,\frac{k}{n}, \quad \varepsilon = \frac{1}{\sqrt{k}}, \quad \text{and} \quad s = \left\lfloor \frac{\log\left(n/k\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)} \right\rfloor + 1, \tag{10}$$

where $0 < \alpha < 0.5$ and $\zeta \geq 1$ are fixed constants. Then,

$$\log\left(\left|\mathcal{R}_{s,\,w,\,\varepsilon}\right|\right) \leq \frac{\log\left(n/k\right)\left(\log\left(dk\right) + 3\log\log(n)\right)}{\log\left(\left(1-\alpha\right)^{-1}\right)} + \mathcal{O}\left(\log\left(\max\left\{n,\ d\right\}\right)\right). \tag{11}$$

Leaves'approximation under empirical measure

Theorem

Suppose that Assumption 1 holds, and that we have a sequence of problems indexed by n with values of d and k satisfying Assumption 2. Let be as defined in (8) with s, and choose ε and w such that

$$\varepsilon = \frac{1}{\sqrt{k}}, \quad \text{and} \quad w = \frac{1}{2\zeta}\,\frac{k}{n}, \tag{12}$$

where $\zeta \geq 1$ is the constant from Assumption 1. Then, there exists an $n_0 \in \mathbb{N}$ such that, for every $n \geq n_0$, the following statement holds with probability at least $1 - n^{-1/2}$: for every possible leaf $L \in \mathcal{L}_{\alpha,\,k}$, we can select a rectangle $R \in \mathcal{R}_{s,\,w,\,\varepsilon}$ such that $R \subseteq L$, $\lambda(L) \leq e^{\varepsilon}\lambda(R)$, and

$$\#L - \#R \leq 3\,\zeta^2 \varepsilon \#L + 2\sqrt{3\log\left(|\mathcal{R}_{s,\,w,\,\varepsilon}|\right)\#L} + \mathcal{O}\left(\log\left(|\mathcal{R}_{s,\,w,\,\varepsilon}|\right)\right). \tag{13}$$

Lower Bounds

Theorem

For any $r > 0$*, set* $d := d(n) = \lfloor n^r \rfloor$*, and let* $\alpha \leq 0.2$*. Then, there exists a distribution over* (X, Y) *and a sequence* $k(n)$ *satisfying the conditions of main Theorem for which*

$$\lim_{n \to \infty} \mathbb{P}\left[\sup_{x \in [0,1]^d, \Lambda \in \mathcal{V}_{\alpha,k}} |T_\Lambda(x) - T_\Lambda^*(x)| \geq \frac{M}{5} \sqrt{\frac{\log(n)\log(d)}{k}} \right] = 1. \tag{14}$$

Outline

Guess-and-check forest procedure

1. Select a currently unsplit node ν containing at least $2k$ training examples.
2. Pick a candidate splitting variable $j \in \{1, ..., d\}$ uniformly at random.
3. Pick the minimum squared error splitting point $\hat{\theta}$. More specifically,

$$\hat{\theta} = \ \textbf{argmax}\ I(\theta) := \frac{4\, N^{-}(\theta)\, N^{+}(\theta)}{(()\, N^{-}(\theta) + N^{+}(\theta))^2}\, \Delta^2\,(\theta)$$

$$\textbf{such that}\ \theta = (X_i)_j \text{ for some } X_i \in \nu$$

$$\alpha \left|\{i : X_i \in \nu\}\right|,\ k \le N^{-}(\theta),\ N^{+}(\theta)$$

$$\textbf{where}\ \Delta(\theta) = \sum_{\{i:X_i \in \nu(x),\,(X_i)_j > \theta\}} Y_i \,/\, N^{+} \ - \sum_{\{i:X_i \in \nu(x),\,(X_i)_j \le \theta\}} Y_i \,/\, N^{-},$$

$$N^{-}(\theta) = \left|\{i : X_i \in \nu,\ (X_i)_j \le \theta\}\right|,$$

$$N^{+}(\theta) = \left|\{i : X_i \in \nu,\ (X_i)_j > \theta\}\right|.$$

Guess-and-check forest procedure

4 If **either** there has already been a successful split along variable j for some other node **or**

$$\ell\left(\hat{\theta}\right) \geq \left(2 \times 9M \sqrt{\frac{\log(n)\log(d)}{k \log\left((1-\alpha)^{-1}\right)}}\right)^2, \tag{15}$$

the split succeeds and we cut the node ν at $\hat{\theta}$ along the j-th variable ; if not, we do not split the node ν this time.

Guarantee for noise feature

This construction relies on a guarantee from Theorem 1 that no noise feature j will ever appear significant enough to get unlocked at any stage of the forest-generation process.

Assumptions for point-wise consistency theorem

Assumption 3 (sparse signal)

There is a signal set $\mathcal{Q} \in \{1, ..., d\}$ of size $|\mathcal{Q}| \leq q$ such that the set of random variables $\{(X_i)_j : j \notin \mathcal{Q}\}$ is jointly independent of Y_i and the set $\{(X_i)_j : j \in \mathcal{Q}\}$.

Assumption 4 (Monotone signal)

There is a minimum effect size $\beta > 0$ and a set of sign variables $\sigma_j \in \{\pm 1\}$ such that, for all $j \in \mathcal{Q}$ and all $x \in [0, 1]^d$,

$$\sigma_j \Bigg(\mathbb{E}\left[Y_i \,\middle|\, (X_i)_{-j} = x_{-j}, (X_i)_j > \frac{1}{2} \right] - \mathbb{E}\left[Y_i \,\middle|\, (X_i)_{-j} = x_{-j}, (X_i)_j \leq \frac{1}{2} \right] \Bigg) \geq \beta,$$

where $x_{(-j)} \in [0, 1]^{d-1}$ denotes the vector containing all but the j-th coordinate of x.

Theorem 3

Assumption 5

The function $\mathbb{E}\left[Y \mid X = x\right]$ is Lipschitz-continuous in x.

Theorem

Under the conditions of Theorem 1 with $\liminf d/n > 0$, *suppose that* $\hat{y}(x)$ *are estimates for* $\mathbb{E}\left[Y \mid X = x\right]$ *obtained using a guess-and-check forest. Suppose, moreover, that Assumptions 3, 4 and 5 hold. Then,*

$$\lim_{n\, d,\, k \to \infty} \sup_{x \in [0,1]^d} \left|\left|\hat{y}(x) - \mathbb{E}\left[Y \mid X = x\right]\right|\right| = 0.$$

Proof idea

Never split on noise variable

We show that

$$|\Delta(\theta) - \Delta^*(\theta)| = |\Delta(\theta)| \leq 2 \times 9M \sqrt{\frac{\log(n) \log(d)}{k \log\left((1-\alpha)^{-1}\right)}},$$

with probability at least $1 - \mathcal{O}(1/\sqrt{n})$, uniformly over all possible nodes ν with at least $2k$ observations and all variables $j \notin Q$.
We use here assumption 3 to prove that $\Delta^*(\theta) = 0$ and we use theorem 1 to find the correct bound.

Make enough splits along signal variables

Let π_j be the probability that the first time any guess-and-check tree tries to split along j, the split succeeds. We show that $\pi_j = 1 - \mathcal{O}(1/\sqrt{n})$.

Proof idea

Conclusion conditionnally on the event $\mathcal{A}$

We start by defining the following event $\mathcal{A}$: "all trees in the forest never split on any variable $j \notin \mathcal{Q}$, and always split on any variable $j \in \mathcal{Q}$ when j is drawn in phase 2 of the guess-and-check procedure.". We show that conditionnally on $\mathcal{A}$,

$$\sup_{x \in [0,1]^d} |H^*(x) - \mathbb{E}[Y \,|X = x]\,| = o_p(1).$$

From theorem 2, we already know that

$$\sup_{x \in [0,1]^d} |H(x) - H^*(x)| = \mathcal{O}_p\left(\sqrt{\frac{\log(n)\log(d)}{k}}\right)$$

Because $P(\mathcal{A}) \to 1$ we finally obtain the result.

Implementation of the procedure

Variables X taken in $[0,1]^d$ uniformely. We set $Y = X_0 + X_1$.

n, d, k	Bound	Successful split	Energy on X_0, X_1	Energy on noisy variables
$1000, 10, 100$	297	0	~ 0.27	$\sim 10^{-4}$
$5\,10^6, 10, 5\,10^5$	0.13	7	~ 0.25	$\sim 10^{-7}$

- Noisy and real variables are well discriminated
- If M is high, the bound is in $\frac{\log n \log d}{k}$: requires to increase n so as to increase k. Lot of memory quickly becomes necessary.
- Setting $\alpha = 0.5$ greatly increases compute time (only one possible split to test)

The End