Bornes de risque pour les forêts purement uniformément aléatoires

Robin Genuer

Bornes de risque pour les forêts purement uniformément aléatoires

Robin Genuer

Laboratoire de mathématiques, Bât.425, Université Paris-Sud 11, 91405 Orsay.

MOTS CLES

Modèles semi et non paramétriques, Apprentissage

RESUME

Introduites par Leo Breiman en 2001, les forêts aléatoires sont une méthode statistique très performante. D'un point de vue théorique, leur analyse est difficile, du fait de la complexité de l'algorithme. Pour expliquer ces performances, des versions de forêts aléatoires simplifiées (et donc plus faciles à analyser) ont été introduites : les forêts purement aléatoires. Dans cet article, nous introduisons une autre version simplifiée, que nous appelons forêts purement uniformément aléatoires. Dans un contexte de régression avec une seule variable explicative, nous montrons que les arbres aléatoires ainsi que les forêts aléatoires atteignent la vitesse de convergence minimax. Et plus important, nous prouvons que les forêts aléatoires améliorent les performances des arbres aléatoires, en réduisant la variance des estimateurs associés d'un facteur trois quarts.

ABSTRACT

Random forests, introduced by Leo Breiman in 2001, are a very effective statistical method. The complex mechanism of the method makes theoretical analysis difficult. To give some insights, people introduced simplified random forests, called purely random forests, which can be theoretically handled more easily. In this paper we introduce random forests of this kind, that we call purely uniform random forests. In the context of regression with an one dimensional input vector, we show that both random trees and random forests reach minimax rate of convergence. More importantly, we prove that compared to random trees, random forests improve accuracy by reducing the estimator variance by a factor of three fourths.

1 Framework

The framework we consider all along the paper is the classical regression in random design framework.

More precisely, we consider a learning set $\mathcal{L}_n = ((X_1, Y_1), \ldots, (X_n, Y_n))$ made of n i.i.d. observations of a vector (X, Y) of an unknown distribution. We consider the following statistical model:

$$Y_i = s(X_i) + \varepsilon_i \quad i = 1, \ldots, n \; .$$

And we make the following assumptions. The strongest one is the fact that we deal only with one dimensional input vectors.

Hypothesis 1

- $X \in [0, 1]$ *of marginal distribution* μ*;*
- $Y \in \mathbb{R}$*;*
- $(\varepsilon_1, \ldots, \varepsilon_n)$ *are i.i.d. observations of* ε*, independent of* $\mathcal{L}_n$*, whith* $\mathbb{E}[\varepsilon] = 0$ *and* $\mathrm{Var}(\varepsilon) = \sigma^2$ *assumed known.*

s is the unknown regression function and the goal is to estimate s.

This paper aims at comparing performances in estimating s between a random tree-predictor and random forest-predictor of a special kind, described in the next section.

2 Purely uniform random tree

Purely random forests (PRF) (see Cutler and Zhao (2001), Biau et al. (2008)) are a simplified variant of random forests-RI (see Breiman (2001)). The biggest difference is that in PRF, the splits of tree nodes are randomly drawn *independently* of the learning set $\mathcal{L}_n$. Whereas in random forests-RI, the splits are optimised with $\mathcal{L}_n$. And people introduced PRF cause with this independence between splits and $\mathcal{L}_n$, they can be theoretically handled more easily.

We introduce Purely Uniform Random Forests (PURF), another simple variant of random forests. The principle of such RF is that for each tree we draw k uniform random variables, which form the partition of the input space $[0, 1]$. Then we build a regressogram on this partition, that we call a tree. Then, a PURF is defined (as all RF variants) as the mean of q such trees.

Note that, unlike PRF or random forests-RI, the tree structure of individual predictors is not obvious. This comes from the fact that in PURT the partition is not obtained in a recursive manner. Nevertheless we keep the vocabulary of trees and forests to distinguish individual predictors from aggregated ones.

More precisely, let $\mathbb{U} = (U_1, \ldots, U_k)$ be k i.i.d. random variables of distribution $\mathcal{U}([0,1])$, where $k \in \mathbb{N}$ will be a parameter of interest.

A Purely Uniform Random Tree (PURT), associated to $\mathbb{U}$, is define for $x \in [0,1]$ as:

$$\hat{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \hat{\beta}_j \mathbb{1}_{U_{(j)} < x \leq U_{(j+1)}}$$

where

$$\hat{\beta}_j = \frac{1}{\sharp\{i : \ U_{(j)} < X_i \leq U_{(j+1)}\}} \sum_{i:\ U_{(j)} < X_i \leq U_{(j+1)}} Y_i$$

and $(U_{(1)}, \ldots, U_{(k)})$ is the ordered statistics of $(U_1, \ldots, U_k)$ and $U_{(0)} = 0$, $U_{(k+1)} = 1$.

And let us define, for $x \in [0,1]$:

$$\tilde{s}_{\mathbb{U}}(x) = \sum_{j=0}^{k} \beta_j \mathbb{1}_{U_{(j)} < x \leq U_{(j+1)}}$$

where

$$\beta_j = \mathbb{E}[Y | U_{(j)} < X \leq U_{(j+1)}] \ .$$

Conditionnaly on $\mathbb{U}$, $\tilde{s}_{\mathbb{U}}$ is the best estimator of s, but of course it depends of the unknown distribution of (X, Y). Note that $\tilde{s}_{\mathbb{U}}$ is random.

With these notations, we can write a bias-variance decompostion as follows:

$$\begin{aligned} \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] &= \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] + \mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \\ &= \text{variance term} + \text{bias term} \end{aligned}$$

2.1 Variance of a tree

Conditionnaly on $\mathbb{U}$, we are in the case of a regressogram on a deterministic partition. Using a proposition from Arlot (2008) about the variance of such regressogram, we manage to establish the following proposition.

Proposition 1 *If* $k \xrightarrow[n \to +\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n \to +\infty]{} 0$, $\forall x \in [0,1] \quad \mu(x) > 0$ *and* s *is* C*-lipschitz, we have:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] = \frac{\sigma^2(k+1)}{n} + \underset{n \to +\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \ .$$

Let us note that the conditions on k and n are the same as in consistency results of Biau et al. (2008).

2.2 Bias of a tree

We show that the bias term behaves as in the case of a deterministic regular partition.

Proposition 2 *If μ the marginal density of X is bounded by $M > 0$ and s is C-lipschitz, we have:*
$\mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \leq \frac{6MC^2}{(k+1)^2}$.

2.3 Risk bound of a tree

Putting together the previous results, we get the following risk bound.

Theorem 1 *If $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\forall x \in [0,1] \quad 0 < \mu(x) \leq M$ and s is C-lipschitz, we have:*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leq \frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) .$$

And the classical balance between the two terms leads to take $(k+1) = n^{1/3}$, and gives the upper bound for the risk:

Corollary 1 *On the same assumptions,*

$$\mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] \leq Kn^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

where K is a positive constant.

This is the well-know minimax rate associated to the class of Lipschitz functions.

3 Purely uniform random forest

A random forest is the aggregation of a collection of random trees. So, in the context of PURF, the principle is to generate several PURT by drawing several random partitions given by uniforms random variables, and to aggregate these trees.

Let $\mathbb{V} = (\mathbb{U}_1, \ldots, \mathbb{U}_q)$ be q i.i.d. random vectors of the same distribution as $\mathbb{U}$.
A PURF, associated to $\mathbb{V}$, is define for $x \in [0,1]$ as follows:

$$\hat{s}(x) = \frac{1}{q}\sum_{l=1}^{q} \hat{s}_{\mathbb{U}_l}(x) .$$

And let us define, for $x \in [0,1]$:

$$\tilde{s}(x) = \frac{1}{q}\sum_{l=1}^{q} \tilde{s}_{\mathbb{U}_l}(x) \ .$$

Again, we have a bias-variance decomposition:

$$\begin{aligned} \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - s(X))^2] &= \mathbb{E}[(\hat{s}_{\mathbb{U}}(X) - \tilde{s}_{\mathbb{U}}(X))^2] + \mathbb{E}[(\tilde{s}_{\mathbb{U}}(X) - s(X))^2] \\ &= \text{variance term} + \text{bias term} \end{aligned}$$

3.1 Variance of a forest

Theorem 2 *If* $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\mu > 0$, s *is* C*-Lipschitz and* $q \xrightarrow[n\to+\infty]{} +\infty$ *we have,*

$$\mathbb{E}[(\hat{s}(X) - \tilde{s}(X))^2] \leq \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \ .$$

3.2 Bias of a forest

A convex inequality gives that the bias of a forest is not larger than the bias of one single tree:

$$\begin{aligned} \mathbb{E}[(\tilde{s}(X) - s(X))^2] &\leq \frac{1}{q}\sum_{l=1}^{q} \mathbb{E}[(\tilde{s}_{\mathbb{U}_l}(X) - s(X))^2] \\ &= \mathbb{E}[(\tilde{s}_{\mathbb{U}_1}(X) - s(X))^2] \ . \end{aligned}$$

And from Propostion 2, we deduce that:

Proposition 3 *If* μ *the marginal density of* X *is bounded by* $M > 0$ *and* s *is* C*-lipschitz:*

$$\mathbb{E}[(\tilde{s}(X) - s(X))^2] \leq \frac{6MC^2}{(k+1)^2} \ .$$

3.3 Risk bound of a forest

The previous upper bounds on the variance and the bias of a PURF allow us to state that:

Theorem 3 *If* $k \xrightarrow[n\to+\infty]{} +\infty$, $\frac{k}{n} \xrightarrow[n\to+\infty]{} 0$, $\forall x \in [0,1] \quad 0 < \mu(x) \leq M$, s *is* C*-lipschitz and* $q \xrightarrow[n\to+\infty]{} +\infty$, *we have:*

$$\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq \frac{3}{4}\frac{\sigma^2(k+1)}{n} + \frac{6MC^2}{(k+1)^2} + \underset{n\to+\infty}{\mathrm{o}}\left(\frac{k}{n}\right) \ .$$

Again, taking $(k+1) = n^{1/3}$ gives the upper bound for the risk:

Corollary 2 *On the same assumptions,*

$$\mathbb{E}[(\hat{s}(X) - s(X))^2] \leq Kn^{-2/3} + \underset{n\to+\infty}{\mathrm{o}}(n^{-2/3})$$

where K is a positive constant.

So a PURF reaches the minimax rate of convergence for C-lipschitz functions.

Secondly, as the variance of a PURF is systematically reduced compared to a PURT and as the bias of a PURF is not larger than the one of a PURT: the risk of a PURF is actually lower than the risk of a PURT.

4 Conclusion

We exhibit, for a very simple version of random forests, the actual gain of using a random forest instead of using one single random tree. First, we show that both trees and forests reach the minimax rate. Then, we manage to highlight a reduction of the variance of a forest, compared to the variance of a tree. This is, for the simple version of random forest we considered here, a proof of the well-known conjecture for random forests: "a random forest, by aggregating several random trees, reduces variance and leaves the bias unchanged" which can be found for example in Hastie et al. (2009).

Bibliographie

[1] Arlot, S. (2008) *V-fold cross-validation improved: V-fold penalization*,
Preprint, arXiv : 0802.0566v2
[2] Breiman, L. (2001) *Random Forests*, Machine Learning, 45, 5-32.
[3] Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J. (1984) *Classification And Regression Trees*, Chapman & Hall, New York.
[4] Biau, G., Devroye, L. and Lugosi, G. (2008) *Consistency of random forests and other averaging classifiers*, Journal of Machine Learning Research, 9, 2039-2057
[5] Cutler, A. and Zhao, G. (2001) *Pert - Perfect random tree ensembles*, Computing Science and Statistics, 33, 490-497
[6] Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning*, Second edition, Springer.