Random Forest for Regression of a Censored Variable

Yohann Le Faou
Univ. Pierre et Marie Curie Paris VI, 4 place Jussieu, 75005 Paris, France

Supervisor(s): Olivier Lopez (UPMC) and Guillaume Gerber (Forsides)

Ph.D. expected duration: 2015-2018

Adress: 200, rue Championnet, 75018, Paris

Email: yohann.lefaou@forsides.fr

Abstract:

1 Introduction

Given a quantitative random variable T, a function ϕ and a vector of covariates X, a common problem in statistics, called regression, is to estimate $E[\phi(T)|X]$ as a function of X. A well known regression technique brought by L. Breiman in the early 2000s ([1]) is the Random Forest algorithm. We propose to adapt the Random Forest method to the case where T is right-censored by a random variable C. Our method inspires from [4] which describes a CART algorithm for the study of a censored variable. We emphasize practical aspects of our work, as one of our purposes is to build a scoring system for the use of an insurance broker. Of particular interest is the calibration of the observation weights we use in our method, which we carefully discuss. We also compare performances of our method to other state of the art models in real and simulated data studies.

Random Survival Forest have been proposed in [3] and [2] to extend Random Forests to the censored case. This algorithm aims to model the entire survival function of T, given X, and thus can be used to estimate $E[\phi(T)|X]$. Our approach is more direct than the latter since it doesn't rely on the estimation of the whole conditional distribution of T. Indeed, our algorithm relies on the weighting of the observations by the inverse-probability-of-censoring weighting principle. The same idea is studied in [5] but this article restricts to single tree model and doesn't go into details about the practical computation of the weights, two subjects we believe we bring new contributions.

Our work is motivated by an application to insurance that we describe in part 3

2 Mathematical formulation

Let T a right-censored random variable. We call C the censoring variable of T. This means each experiment doesn't lead to an observation of T. In fact, each experiment leads to an observation of $Y = \min(T, C)$ and $\delta = \mathbb{1}_{T \leq C}$

Let $X \in \mathbb{R}^p$ a vector of covariates and ϕ a real valued function. In this context, we are interested in estimating the influence of X on $\phi(T)$.

We have observations $(Y_i, \delta_i, X_i)_{i=1,\dots,n}$ and we look for estimations of $f(x) = E[\phi(T)|X = x]$. As we know, f is the solution to the optimization problem : $f = \underset{g}{\operatorname{argmin}}\, E\left[(\phi(T) - g(X))^2\right]$

We then choose the Random Forest algorithm with the mean squared error splitting criteria to estimate f. This leads us in looking for estimators of quantities of the form : $E[\psi(T, X)]$. Under some hypothesis it is possible to estimate the quantity $E[\psi(T, X)]$ asymptotically without bias, with ψ a real valued function. We

use the inverse-probability-of-censoring weighting principle to do so. It is a general principal that provide unbiased estimate of the law of a couple (Z, X) when observation of X is complete and observation of Z is censored.

In our case, the probability of being non-censored given X and T is $P(\delta = 1|X, T) = P(T \leq C|X, T)$. In the survival censoring scheme, it's impossible to infer the latter since it is well known it's impossible to estimate the dependence between T and C. Therefore, we have to make assumptions about the dependence between T and C. Let **H1** and **H2** denotes the following hypothesis :

$$\begin{aligned} &\textbf{H1} : P(T \leq C|X, T) = P(T \leq C) \\ &\textbf{H2} : P(T \leq C|X, T) = P(T \leq C|X) \end{aligned}$$

Sufficient conditions for these hypothesis to be satisfied are, respectively, $T \perp\!\!\!\perp C$ (**H1**) and $T \perp\!\!\!\perp C$ conditionally on X (**H2**).

Let S_C the survival function of C, $S_C(\cdot|X)$ the survival function of C given X, and denote by $\hat{S}_C$ and $\hat{S}_C(\cdot|X)$ estimators of these functions. Then, depending on the hypothesis we make, let $\hat{W}_i = \frac{\delta_i}{\hat{S}_C(Y_i)}$ or $\frac{\delta_i}{\hat{S}_C(Y_i|X_i)}$. We estimate $E[(\phi(T) - g(X))^2]$ by

$$\frac{1}{n}\sum_{i=1}^{n} \hat{W}_i \cdot (\phi(Y_i) - g(X_i))^2 \tag{1}$$

The Random Forest adaptation we propose is then a weighted Random Forest. Weights are taken into account in the bootstrap procedure. Indeed, during the sampling of a bootstrap set, we do a sample with replacement where each observation has probability $\hat{W}_i$ of being sampled. This way, each observation accounts in the growing of the forest proportionally to its weight.

3 Application scheme

Our work is motivated by an application in insurance where T corresponds to termination time of a contract and ϕ gives the amounts of commissions received by an insurance broker per unity of premium. ϕ then represents the impact of the termination time of a contract on the turnover this contract brings.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Hemant Ishwaran and Udaya B Kogalur. Random survival forests for r. *New Functions for Multivariate Analysis*, page 25, 2007.

[3] Hemant Ishwaran, Udaya B Kogalur, Eugene H Blackstone, and Michael S Lauer. Random survival forests. *The annals of applied statistics*, pages 841–860, 2008.

[4] Olivier Lopez, Xavier Milhaud, and Pierre-Emmanuel Thérond. Tree-based censored regression with applications to insurance. *disponible sur le Hal*, 2015.

[5] Annette M Molinaro, Sandrine Dudoit, and Mark J Van der Laan. Tree-based multivariate regression and density estimation with right-censored data. *Journal of Multivariate Analysis*, 90(1):154–177, 2004.

Short biography – After mathematical studies at ENS Rennes, I began a CIFRE thesis in september 2015 in the fields of statistics and actuarial science. I do my PhD at Forsides Innovation and with UPMC doctoral program.