DET
RAN
OTE

DETRA NOTE 2020-4

THE REASON WHY BAGGING TREES OUTPERFORM DECISION TREE...

By Candy Mahirwe, Michel Denuit and Julien Trufin

Detralytics

DISCLAIMER

The content of the Detra Notes for a pedagogical use only. Each business case is so specific that a careful analysis of the situation is needed before implementing a possible solution. Therefore, Detralytics does not accept any liability for any commercial use of the present document. Of course, the entire team remain available if the techniques presented in this Detra Note required your attention.

Detralytics
Rue Belliard/Belliardstraat 2
1040 Brussels
www.detralytics.com
info@detralytics.eu

ABSTRACT

One of the objectives of ensemble techniques is to improve model accuracy by driving down the variance without affecting too much the bias. In this note, we consider bagging trees. Bagging trees is an ensemble technique which consists in combining several regression trees fitted on different bootstrap samples of the training set. We demonstrate that bagging trees performs better than one of its constituent trees in the sense of the expected generalization error. Moreover, we show through an example that bagging trees outperforms not only one of its constituent tree but also the best decision tree built on the entire training set.

Keywords: Bagging trees, regression tree, generalization error, Poisson deviance loss.

Contents

Chapter 1

Introduction

The expected generalization error of a model could be reduced by driving down the variance of the model without increasing too much the bias. Ensemble methods are relevant tools to perform this task. The principle of ensemble methods based on randomization consists in introducing random perturbations into the training procedure in order to get different models from a single training set $\mathcal{D}$ and combining them to obtain the estimate of the ensemble.

Bagging is one of the first ensemble methods proposed in the literature. This algorithm is used for reducing the variance of an estimate. Typically, it works well for high variance and low bias procedures, such as regression trees.

In this note, we introduce the concepts of expected generalization error and randomized learning procedures in Chapter 2. Then, Chapter 3 is devoted to bagging trees and demonstrates the advantage of this ensemble method compared to one of its constituent trees. We also show through an example that bagging trees outperforms the best decision tree built on the entire training set. The final chapter briefly concludes the note.

Chapter 2

Model Performance

2.1 Generalization error

We denote by

$$\mathcal{L} = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\} \tag{2.1}$$

the set of observations available to the insurer, where y_i and $\boldsymbol{x}_i$ are the response and the features available for policyholder i. This dataset is called the learning set. The general problem of supervised learning can be stated as finding a model $\widehat{\mu}$ built on the learning set $\mathcal{L}$ (or only on a part of $\mathcal{L}$, as discussed thereafter) which minimizes the generalization error. The generalization error, also known as expected prediction error, of $\widehat{\mu}$ is defined as follows:

Définition 2.1.1. *The generalization error of the model $\widehat{\mu}$ is*

$$Err(\widehat{\mu}) = E[L(Y, \widehat{\mu}(\boldsymbol{X}))], \tag{2.2}$$

where $L(.,.)$ is a function measuring the discrepancy between its two arguments, called loss function, $\boldsymbol{X}$ is the random vector gathering the observable features and Y is the response variable.

The goal is thus to find a function of the covariates which predicts at best the response, that is, which minimizes the generalization error. The model performance is evaluated according to the generalization error which depends on a predefined loss function. A simple estimate of the generalization error is given by

$$\widehat{Err}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{\mu}(\boldsymbol{x}_i)). \tag{2.3}$$

In the ED (Exponential Dispersion) family setting, the appropriate choice for the loss function is then related to the deviance. It suffices to observe that the regression model $\widehat{\mu}$ maximizing the log-likelihood function $LF(\widehat{\mu})$ also minimizes the corresponding deviance $D(\widehat{\mu})$, so that (2.3) becomes

$$\widehat{Err}(\widehat{\mu}) = \frac{D(\widehat{\mu})}{n}. \tag{2.4}$$

Notice that the expectation in (2.2) is taken over all possible data, that is, with respect to the probability distribution of the random vector $(Y, \boldsymbol{X})$ assumed to be independent of the learning set $\mathcal{L}$.

2.2 Estimates

The performance of a model is evaluated throughout the generalization error $Err(\widehat{\mu})$. In practice, we usually do not know the probability distribution from which the observations are drawn, making the direct evaluation of the generalization error $Err(\widehat{\mu})$ not feasible. Hence, the set of observations available to the insurer often constitutes the only data on which the model needs to be fitted and its generalization error estimated.

2.2.1 Training sample estimate

The learning set

$$\mathcal{L} = \{(y_1, \boldsymbol{x}_1), (y_2, \boldsymbol{x}_2), \ldots, (y_n, \boldsymbol{x}_n)\} \tag{2.5}$$

constitutes the only data available to the insurer. When the whole learning set is used to fit the model $\widehat{\mu}$, the generalization error $Err(\widehat{\mu})$ can only be estimated on the same data as the ones used to build the model, that is,

$$\widehat{Err}^{\text{train}}(\widehat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} L(y_i, \widehat{\mu}(\boldsymbol{x}_i)). \tag{2.6}$$

This estimate is called the training sample estimate and has been introduced in (2.3). In our setting, we thus have

$$\widehat{Err}^{\text{train}}(\widehat{\mu}) = \frac{D(\widehat{\mu})}{n}. \tag{2.7}$$

Typically, the training sample estimate (2.6) will be less that the true generalization error, because the same data is being used to fit the model and assess its error. A model typically adapts to the data used to train it, and hence the training sample estimate will be an overly optimistic estimate of the generalization error.

2.2.2 Validation sample estimate

The training sample estimate (2.6) directly evaluates the accuracy of the model on the dataset used to build the model. While the training sample estimate is useful to fit the model, as we aim to minimize it (the deviance in our context) when we build the model, the resulting estimate for the generalization error is likely to be very optimistic since the model is precisely built to reduce it. This is of course an issue when we aim to assess the predictive performance of the model, namely its accuracy on new data.
As actuaries generally deal with massive amounts of data, a better approach is to divide the learning set $\mathcal{L}$ into two disjoint sets $\mathcal{D}$ and $\overline{\mathcal{D}}$, called training set and validation set, and to use the training set for fitting the model and the validation set for estimating the generalization error of the model. The learning set is thus partitioned into a training set

$$\mathcal{D} = \{(y_i, \boldsymbol{x}_i); i \in \mathcal{I}\}$$

and a validation set

$$\overline{\mathcal{D}} = \{(y_i, \boldsymbol{x}_i); i \in \overline{\mathcal{I}}\},$$

with $\mathcal{I} \subset \{1, \ldots, n\}$ labeling the observations in $\mathcal{D}$ considered for fitting the model and $\overline{\mathcal{I}} = \{1, \ldots, n\} \backslash \mathcal{I}$ labelling the remaining observations of $\mathcal{L}$ used to assess the predictive accuracy of the model. The validation sample estimate of the generalization error of the model $\widehat{\mu}$ that has been built on the training set $\mathcal{D}$ is then given by

$$\widehat{Err}^{\text{val}}(\widehat{\mu}) = \frac{1}{|\overline{\mathcal{I}}|} \sum_{i \in \overline{\mathcal{I}}} L(y_i, \widehat{\mu}(\boldsymbol{x}_i)), \tag{2.8}$$

while the training sample estimate (2.9) now writes

$$\begin{aligned} \widehat{Err}^{\text{train}}(\widehat{\mu}) &= \frac{1}{|\mathcal{I}|} \sum_{i \in \mathcal{I}} L(y_i, \widehat{\mu}(\boldsymbol{x}_i)) \\ &= \frac{D^{\text{train}}(\widehat{\mu})}{|\mathcal{I}|}, \end{aligned} \tag{2.9}$$

where we denote by $D^{\text{train}}(\widehat{\mu})$ the deviance computed from the observations composing the training set. As a rule-of-thumb, the training set usually represents 80% of the learning set and the validation set the remaining 20%. Of course, this allocation depends on the problem under consideration. In any case, the splitting of the learning set must be done in a way that observations in the training set can be considered independent from those in the validation set and drawn from the same population. Usually, this is guaranteed by drawing both sets at random from the learning set.

Training and validation sets should be as homogeneous as possible. Creating those two sets by taking simple random samples, as mentioned above, is usually sufficient to guarantee similar data sets. However, when considering the annual number of claims in MTPL insurance for instance, the distribution of the response can be quite different between the training and validation sets. Typically, the vast majority of the policyholders makes no claim over the year (say 95%). Some policyholders experience one claim (say 4%) while only a few of them have more than one claim (say 1% with two claims). In such a situation, because the proportions of policyholders with one or two claims are small compared to the proportion of policyholders with no claim, the distribution of the response can be very different between the training and validation sets.

To address this potential issue, random sampling can be applied within subgroups, a subgroup being a set of observations with the same response. In our example, we would thus have three subgroups: a first one made of the observations with no claim (95% of the observations), a second one with the policyholders having one claim (4% of the observations) and a third one with the policyholders having two claims (1% of the observations). Applying the randomization within these subgroups is called stratified random sampling.

2.3 Decomposition of the generalized error

The generalization error $Err(\widehat{\mu})$ of a model $\widehat{\mu}$ is thus defined as

$$Err(\widehat{\mu}) = \mathrm{E}\left[L(Y, \widehat{\mu}(\boldsymbol{X}))\right]. \tag{2.10}$$

In the same way, the generalization error of $\widehat{\mu}$ can be defined for a fixed value $\boldsymbol{X} = \boldsymbol{x}$ as

$$Err(\widehat{\mu}(\boldsymbol{x})) = \mathrm{E}\left[L(Y, \widehat{\mu}(\boldsymbol{X}))|\boldsymbol{X} = \boldsymbol{x}\right]. \tag{2.11}$$

Notice that averaging the local errors $Err(\widehat{\mu}(\boldsymbol{x}))$ enables to recover the generalization error $Err(\widehat{\mu})$, that is,

$$Err(\widehat{\mu}) = \mathrm{E}\left[Err(\widehat{\mu}(\boldsymbol{X}))\right]. \tag{2.12}$$

2.3.1 Squared error loss

Consider that the loss function is the squared error loss. In our ED family setting, it amounts to assume that the responses are normally distributed. The generalization error of model $\widehat{\mu}$ at $\boldsymbol{X} = \boldsymbol{x}$ becomes

$$\begin{aligned}
Err(\widehat{\mu}(\boldsymbol{x})) = &= \mathrm{E}\left[(Y - \widehat{\mu}(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathrm{E}\left[(Y - \mu(\boldsymbol{x}) + \mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathrm{E}\left[(Y - \mu(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] + \mathrm{E}\left[(\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&\quad +2\,\mathrm{E}\left[(Y - \mu(\boldsymbol{x}))\,(\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x})) \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= \mathrm{E}\left[(Y - \mu(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] + \mathrm{E}\left[(\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x}))^2 \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right]
\end{aligned}$$

since

$$\begin{aligned}
&\mathrm{E}\left[(Y - \mu(\boldsymbol{x}))\,(\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x})) \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= (\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x}))\ \mathrm{E}\left[(Y - \mu(\boldsymbol{x})) \,\middle|\, \boldsymbol{X} = \boldsymbol{x}\right] \\
&= (\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x}))\ (\mathrm{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] - \mu(\boldsymbol{x})) \\
&= 0
\end{aligned}$$

by definition of $\mu(\boldsymbol{x}) = \mathrm{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right]$. So, it comes

$$Err\left(\widehat{\mu}(\boldsymbol{x})\right) = Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \widehat{\mu}(\boldsymbol{x})\right)^2. \tag{2.13}$$

By (2.12), the generalization error $Err(\widehat{\mu})$ thus writes

$$Err\left(\widehat{\mu}\right) = Err\left(\mu\right) + \mathrm{E}\left[\left(\mu(\boldsymbol{X}) - \widehat{\mu}(\boldsymbol{X})\right)^2\right]. \tag{2.14}$$

The generalization error of $\widehat{\mu}$ can be expressed as the sum of two terms, the first one corresponding to the generalization error of the true model μ and the second one representing the estimation error, that is, the discrepancy of $\widehat{\mu}$ from the true model μ. The further our model from the true one, the larger the generalization error. The generalization error of the true model is called the residual error and is irreducible. Indeed, we have

$$Err\left(\widehat{\mu}\right) \geq Err\left(\mu\right),$$

which means that the smallest generalization error coincides with the one associated to the true model.

2.3.2 Poisson deviance loss

Consider that the loss function is the Poisson deviance. This choice is appropriate when the responses are assumed to be Poisson distributed, as when examining the number of claims for instance. The generalization error of model $\widehat{\mu}$ at $\boldsymbol{X} = \boldsymbol{x}$ is then given by

$$\begin{aligned} Err\left(\widehat{\mu}(\boldsymbol{x})\right) &= 2\mathrm{E}\left[Y \ln\left(\frac{Y}{\widehat{\mu}(\boldsymbol{x})}\right) - (Y - \widehat{\mu}(\boldsymbol{x}))\Big|\boldsymbol{X} = \boldsymbol{x}\right] \\ &= 2\mathrm{E}\left[Y \ln\left(\frac{Y}{\mu(\boldsymbol{x})}\right) - (Y - \mu(\boldsymbol{x}))\Big|\boldsymbol{X} = \boldsymbol{x}\right] \\ &\quad +2\mathrm{E}\left[\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x}) - Y \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\Big|\boldsymbol{X} = \boldsymbol{x}\right] \\ &= 2\mathrm{E}\left[Y \ln\left(\frac{Y}{\mu(\boldsymbol{x})}\right) - (Y - \mu(\boldsymbol{x}))\Big|\boldsymbol{X} = \boldsymbol{x}\right] \\ &\quad +2\left(\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\right) - 2\mathrm{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right] \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right). \end{aligned}$$

Replacing $\mathrm{E}\left[Y|\boldsymbol{X} = \boldsymbol{x}\right]$ by $\mu(\boldsymbol{x})$, we get

$$\begin{aligned} Err\left(\widehat{\mu}(\boldsymbol{x})\right) &= Err\left(\mu(\boldsymbol{x})\right) + 2\left(\widehat{\mu}(\boldsymbol{x}) - \mu(\boldsymbol{x})\right) - 2\mu(\boldsymbol{x}) \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right) \\ &= Err\left(\mu(\boldsymbol{x})\right) + 2\mu(\boldsymbol{x})\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})} - 1 - \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right). \end{aligned} \tag{2.15}$$

The generalization error $Err(\widehat{\mu})$ thus writes

$$Err\left(\widehat{\mu}\right) = Err\left(\mu\right) + 2\mathrm{E}\left[\mu(\boldsymbol{X})\left(\frac{\widehat{\mu}(\boldsymbol{X})}{\mu(\boldsymbol{X})} - 1 - \ln\left(\frac{\widehat{\mu}(\boldsymbol{X})}{\mu(\boldsymbol{X})}\right)\right)\right]. \tag{2.16}$$

As for the squared error loss, the generalization error of $\widehat{\mu}$ can be decomposed as the sum of the generalization error of the true model and an estimation error $\mathrm{E}[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}(\boldsymbol{X})\right)]$, where

$$\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}(\boldsymbol{x})\right) = 2\mu(\boldsymbol{x})\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})} - 1 - \ln\left(\frac{\widehat{\mu}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right).$$

Notice that $\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}(\boldsymbol{x})\right)$ is always positive because $y \to y - 1 - \ln y$ is positive on $\mathbb{R}^+$, so that we have

$$Err\left(\widehat{\mu}\right) \geq Err\left(\mu\right).$$

2.4 Expected generalization error

The model $\widehat{\mu}$ under consideration is estimated on the training set $\mathcal{D}$ so that it depends on $\mathcal{D}$. To make explicit the dependence on the training set, we use from now on both notations $\widehat{\mu}$ and $\widehat{\mu}_{\mathcal{D}}$ for the model under interest. We assume in a first time there is only one model which corresponds to a given training set, that is, we consider learning procedures that are said to be deterministic. Learning procedures that can produce different models for a fixed training set are discussed in Section 2.4.4.

The generalization error $Err(\widehat{\mu}_{\mathcal{D}})$ is evaluated conditional on the training set. That is, the model $\widehat{\mu}_{\mathcal{D}}$ under study is first fitted on the training set $\mathcal{D}$ before computing the expectation over all possible observations independently from the training set $\mathcal{D}$. In that sense, the generalization error $Err(\widehat{\mu}_{\mathcal{D}})$ gives an idea of the general accuracy of the learning procedure for the particular training set $\mathcal{D}$. In order to study the general behavior of our learning procedure, and not only its behavior for a specific training set, it is interesting to evaluate the learning procedure on different training sets of the same size.
The training set $\mathcal{D}$ is itself a random variable sampled from a distribution usually unknown in practice, so that the generalization error $Err(\widehat{\mu}_{\mathcal{D}})$ is in its turn a random variable. In order to study the general performance of the learning procedure, it is then of interest to take the average of the generalization error $Err\left(\widehat{\mu}_{\mathcal{D}}\right)$ over $\mathcal{D}$, that is, to work with the expected generalization error $\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right]$ over the models learned from all possible training sets and produced with the learning procedure under investigation.
The expected generalization error is thus given by

$$\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right] = \mathrm{E}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{X}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right)\right]\right], \tag{2.17}$$

which can also be expressed as

$$\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right] = \mathrm{E}_{\boldsymbol{X}}\left[\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right)\right]\right]. \tag{2.18}$$

We can first determine the expected local error $\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right)\right]$ in order to get the expected generalization error.

2.4.1 Squared error loss

When the loss function is the squared error loss, we know from equation (2.13) that the generalization error at $\boldsymbol{X} = \boldsymbol{x}$ writes

$$Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right) = Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2. \tag{2.19}$$

The true model μ is independent of the training set, so is the generalization error $Err\left(\mu(\boldsymbol{x})\right)$. The expected generalization error of $\widehat{\mu}$ at $\boldsymbol{X} = \boldsymbol{x}$ is then given by

$$\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] = Err\left(\mu(\boldsymbol{x})\right) + \mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right].$$

The first term is the local generalization error of the true model while the second term is the expected estimation error at $\boldsymbol{X} = \boldsymbol{x}$, which can be re-expressed as

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] + \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)^2\right] + \mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&\quad +2\mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] \\
&= \mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)^2\right] + \mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right] \\
&= \left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)^2 + \mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]
\end{aligned}$$

since

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D}}\left[\left(\mu(\boldsymbol{x})-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]-\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] \\
&=\left(\mu(\boldsymbol{x})-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)\mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]-\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] \\
&=\left(\mu(\boldsymbol{x})-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right) \\
&=0. \qquad (2.20)
\end{aligned}$$

Therefore, the expected generalization error at $\boldsymbol{X}=\boldsymbol{x}$ is given by

$$\begin{aligned}
\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] &= Err\left(\mu(\boldsymbol{x})\right)+\left(\mu(\boldsymbol{x})-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right)^2 \\
&\quad +\mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]-\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)^2\right]. \qquad (2.21)
\end{aligned}$$

This is the bias-variance decomposition of the expected generalization error.
The first term in (2.21) is the local generalization error of the true model, that is, the residual error. The residual error is independent of the learning procedure and the training set, which provides in any case a lower bound for the expected generalization error. Notice that in practice, the computation of this lower bound is often unfeasible since the true model is usually unknown. The second term measures the discrepancy between the average estimate $\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]$ and the value of the true model $\mu(\boldsymbol{x})$, and corresponds to the bias term. The third term measures the variability of the estimate $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ over the models trained from all possible training sets, and corresponds to the variance term.
From (2.21), the expected generalization error writes

$$\begin{aligned}
\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right] &= Err\left(\mu\right)+\mathrm{E}_{\boldsymbol{X}}\left\{\left(\mu(\boldsymbol{X})-\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right]\right)^2\right\} \\
&\quad +\mathrm{E}_{\boldsymbol{X}}\left\{\mathrm{E}_{\mathcal{D}}\left[\left(\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right]-\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})\right)^2\right]\right\}. \qquad (2.22)
\end{aligned}$$

2.4.2 Poisson deviance loss

In the case of the Poisson deviance loss, we know from equation (2.15) that the local generalization error writes

$$Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)=Err\left(\mu(\boldsymbol{x})\right)+\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right) \qquad (2.23)$$

where

$$\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)=2\mu(\boldsymbol{x})\left(\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}-1-\ln\left(\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right). \qquad (2.24)$$

Because the true model μ is independent of the training set, the expected generalization error $\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right]$ can be expressed as

$$\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right]=Err\left(\mu(\boldsymbol{x})\right)+\mathrm{E}_{\mathcal{D}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right] \qquad (2.25)$$

with

$$\mathrm{E}_{\mathcal{D}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right]=2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D}}\left[\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right]-1-\mathrm{E}_{\mathcal{D}}\left[\ln\left(\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]\right).$$

Locally, the expected generalization error is equal to the generalization error of the true model plus the expected estimation error which can be attributed to the bias and the estimation

fluctuation. Notice that the expected estimation error $\mathrm{E}_{\mathcal{D}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)\right]$ is positive since we have seen that the estimation error $\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right)$ is always positive. The generalization error of the true model is again a theoretical lower bound for the expected generalization error.
From (2.25) and (2.26), the expected generalization error writes

$$\mathrm{E}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right] = Err\left(\mu\right) + 2\mathrm{E}_{\boldsymbol{X}}\left\{\mu(\boldsymbol{X})\left(\mathrm{E}_{\mathcal{D}}\left[\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})}{\mu(\boldsymbol{X})}\right] - 1 - \mathrm{E}_{\mathcal{D}}\left[\ln\left(\frac{\widehat{\mu}_{\mathcal{D}}(\boldsymbol{X})}{\mu(\boldsymbol{X})}\right)\right]\right)\right\}. \quad (2.26)$$

2.4.3 Bias and variance

In order to minimise the expected generalization error, it might appear desirable to sacrifice a bit on the bias provided we can reduce to a large extend the variability of the prediction over the models trained from all possible training sets. The bias-variance decomposition of the expected generalization error is used for justifying the performances of ensemble learning techniques.

2.4.4 Randomized learning procedures

A learning procedure which always produces the same model $\widehat{\mu}_{\mathcal{D}}$ for a given training set $\mathcal{D}$ is said to be deterministic.
There also exist randomized learning procedures that can produce different models for a fixed training set, such as random forests and boosting. In order to account for the randomness of the learning procedure, we introduce a random vector $\boldsymbol{\Theta}$ which is assumed to fully capture the randomness of the algorithm. The model $\widehat{\mu}$ resulting from the randomized learning procedure depends on the training set $\mathcal{D}$ and also on the random vector $\boldsymbol{\Theta}$, so that we use both notations $\widehat{\mu}$ and $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}$ for the model under consideration.
The generalization error $Err(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}})$ is thus evaluated conditional on the training set $\mathcal{D}$ and the random vector $\boldsymbol{\Theta}$. The expected generalization error, which aims to assess the general accuracy of the learning procedure, is now obtained by taking the average of the generalization error $Err(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}})$ over $\mathcal{D}$ and $\boldsymbol{\Theta}$. Expression (2.17) becomes

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] = \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathrm{E}_{\boldsymbol{X}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{X})\right)\right]\right], \quad (2.27)$$

which can also be expressed as

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] = \mathrm{E}_{\boldsymbol{X}}\left[\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{X})\right)\right]\right]. \quad (2.28)$$

Again, we can first determine the expected local error $\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{X})\right)\right]$ in order to get the expected generalization error.
Taking into account the additional source of randomness in the learning procedure, expressions (2.21) and (2.25) become respectively

$$\begin{aligned}\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] &= Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right]\right)^2 \\ &\quad + \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] - \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)^2\right], \quad (2.29)\end{aligned}$$

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] = Err\left(\mu(\boldsymbol{x})\right) + \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right], \quad (2.30)$$

and

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] = 2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right] - 1 - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\ln\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]\right). \tag{2.31}$$

Chapter 3

Bagging trees

3.1 Introduction

The expected generalization error of a model could be reduced by driving down the variance of the model without increasing too much the bias. Ensemble methods are relevant tools to perform this task. The principle of ensemble methods based on randomization consists in introducing random perturbations into the training procedure in order to get different models from a single training set $\mathcal{D}$ and combining them to obtain the estimate of the ensemble.
One ensemble method is considered in this note, namely bagging trees. One issue with trees is their high variance. There is a high variability of the mean estimate $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ over the trees trained from all possible training sets $\mathcal{D}$. Ensemble methods like bagging trees and random forests aim to reduce the variance without too much altering bias.
The average estimate $\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]$ has the same bias as $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ since

$$\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] = \mathrm{E}_{\mathcal{D}}\left[\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right], \tag{3.1}$$

and zero variance, that is,

$$\mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]\right] = 0. \tag{3.2}$$

This motivates the fact of finding a training procedure that produces a good approximation of the average model in order to stabilize model estimates.
If we assume that we can draw as many training sets as we want, so that we have B training sets $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^B$ available, then an approximation of the average model can be obtained by averaging the regression trees built on these training sets, that is,

$$\widehat{\mathrm{E}}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x}). \tag{3.3}$$

In such a case, the average of the estimate (3.3) with respect to the training sets $\mathcal{D}^1, \ldots, \mathcal{D}^B$ is the average estimate $\mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]$, that is,

$$\begin{aligned}\mathrm{E}_{\mathcal{D}^1,\ldots,\mathcal{D}^B}\left[\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x})\right] &= \frac{1}{B}\sum_{b=1}^{B}\mathrm{E}_{\mathcal{D}^b}\left[\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x})\right] \\ &= \mathrm{E}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right],\end{aligned} \tag{3.4}$$

while the variance of (3.3) with respect to $\mathcal{D}^1, \ldots, \mathcal{D}^B$ is given by

$$\begin{aligned} \mathrm{Var}_{\mathcal{D}^1,\ldots,\mathcal{D}^B}\left[\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x})\right] &= \frac{1}{B^2}\mathrm{Var}_{\mathcal{D}^1,\ldots,\mathcal{D}^B}\left[\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x})\right] \\ &= \frac{1}{B^2}\sum_{b=1}^{B}\mathrm{Var}_{\mathcal{D}^b}\left[\widehat{\mu}_{\mathcal{D}^b}(\boldsymbol{x})\right] \\ &= \frac{\mathrm{Var}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]}{B} \end{aligned} \tag{3.5}$$

since estimates $\widehat{\mu}_{\mathcal{D}^1}(\boldsymbol{x}), \ldots, \widehat{\mu}_{\mathcal{D}^B}(\boldsymbol{x})$ are independent and identically distributed. So, averaging over B estimates fitted on different training sets leaves the bias unchanged compared to each individual estimate while it divides the variance by B. The estimate (3.3) is then less variable than each individual one.
In practice, the probability distribution from which the observations of the training set are drawn is usually not known so that there is only one training set available. In this context, the bootstrap approach, used both in bagging trees and random forests, appears to be particularly useful.

3.2 Bootstrap

Suppose we have independent random variables $Y_1, Y_2, \ldots, Y_n$ with common distribution function F that is unknown and that we are interested in using them to estimate some quantity $\theta(F)$ associated with F. An estimator

$$\widehat{\theta} = g(Y_1, Y_2, \ldots, Y_n)$$

is available for $\theta(F)$. The distributional properties of $\widehat{\theta}$ in terms of the variables $Y_1, Y_2, \ldots, Y_n$ cannot be determined since the distribution function F is not known. The idea of bootstrap is to estimate F.
The empirical counterpart to F is defined as

$$\widehat{F}_n(x) = \frac{\#\{Y_i \text{ such that } Y_i \leq x\}}{n} = \frac{1}{n}\sum_{i=1}^{n} I[Y_i \leq x].$$

Where I is the indicator function.
Thus, the empirical distribution function $\widehat{F}_n$ puts an equal probability $\frac{1}{n}$ on each of the observed data points $Y_1, \ldots, Y_n$. The idea behind the non-parametric bootstrap is to simulate sets of independent random variables

$$Y_1^{(*b)}, Y_2^{(*b)}, \ldots, Y_n^{(*b)}$$

obeying the distribution function $\widehat{F}_n$, $b = 1, 2, \ldots, B$. This can be done by simulating $U_i \sim \mathcal{U}ni(0, 1)$ and setting

$$Y_i^{(*b)} = y_I \text{ with } I = [nU_i] + 1.$$

Then, for each $b = 1, \ldots, B$, we calculate

$$\widehat{\theta}^{(*b)} = g(Y_1^{(*b)}, Y_2^{(*b)}, \ldots, Y_n^{(*b)}),$$

$\lvert\mathcal{I}\rvert$	$1-\left(\frac{\lvert\mathcal{I}\rvert-1}{\lvert\mathcal{I}\rvert}\right)^{\lvert\mathcal{I}\rvert}$
10	0.651322
100	0.633968
1000	0.632305
10 000	0.632139
100 000	0.632122

Table 3.1: Probability in (3.6) with respect to $|\mathcal{I}|$.

so that the corresponding bootstrap distribution of $\widehat{\theta}$ is given by

$$F^*_{\widehat{\theta}}(x) = \frac{1}{B}\sum_{b=1}^{B} I[\widehat{\theta}^{(*b)} \leq x].$$

3.3 Bagging trees

Bagging is one of the first ensemble methods proposed in the literature. Consider a model fitted to our training set $\mathcal{D}$, obtaining the prediction $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ at point $\boldsymbol{x}$. Bootstrap aggregation or bagging averages this prediction over a set of bootstrap samples in order to reduce its variance. The probability distribution of the random vector $(Y, \boldsymbol{X})$ is usually not known. This latter distribution is then approximated by its empirical version which puts an equal probability $\frac{1}{|\mathcal{I}|}$ on each of the observations $\{(y_i, \boldsymbol{x}_i); i \in \mathcal{I}\}$ of the training set $\mathcal{D}$. Hence, instead of simulating B training sets $\mathcal{D}^1, \mathcal{D}^2, \ldots, \mathcal{D}^B$ from the probability distribution of $(Y, \boldsymbol{X})$, which is not possible in practice, the idea of bagging is rather to simulate B bootstrap samples $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ of the training set $\mathcal{D}$ from its empirical counterpart. Specifically, a bootstrap sample of $\mathcal{D}$ is obtained by simulating independently $|\mathcal{I}|$ observations from the empirical distribution of $(Y, \boldsymbol{X})$ defined above. A bootstrap sample is thus a random sample of $\mathcal{D}$ taken with replacement which has the same size as $\mathcal{D}$. Notice that, on average, 63.2% of the observations of the training set are represented at least once in a bootstrap sample. Indeed,

$$1-\left(\frac{|\mathcal{I}|-1}{|\mathcal{I}|}\right)^{|\mathcal{I}|}, \tag{3.6}$$

which is computed in Table 3.1 for different values of $|\mathcal{I}|$, is the probability that a given observation of the training set is represented at least once. One can see that the value of 63.2% is already attained for values of $|\mathcal{I}|$ around 1000.
Let $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots, \mathcal{D}^{*B}$ be B bootstrap samples of the training set $\mathcal{D}$. For each $\mathcal{D}^{*b}$, $b = 1, \ldots, B$, we fit our model, giving estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}) = \widehat{\mu}_{\mathcal{D}^{*b}}(\boldsymbol{x})$. The bagging estimate is then defined by

$$\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}), \tag{3.7}$$

where $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_B)$. Random vectors $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_B$ fully capture the randomness of the training procedure. For bagging, $\boldsymbol{\Theta}_1, \ldots, \boldsymbol{\Theta}_B$ are independent and identically distributed so

that $\boldsymbol{\Theta}_b$ is a vector of $|\mathcal{I}|$ integers randomly and uniformly drawn in $\{1, 2, \ldots, |\mathcal{I}|\}$. Each component of $\boldsymbol{\Theta}_b$ indexes one observation of the training set selected in $\mathcal{D}^{*b}$.
In this note, bagging is applied to regression trees. This provides the following algorithm:

Algorithm: Bagging Trees.

For $b = 1$ to B **do**

1. Generate a bootstrap sample $\mathcal{D}^{*b}$ of $\mathcal{D}$.
2. Fit an unpruned tree on $\mathcal{D}^{*b}$, which gives estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$.

End for

Output: $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x}) = \frac{1}{B}\sum_{b=1}^{B} \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$.

3.3.1 Bias

For bagging, the bias is the same as the bias of the individual sampled models. Indeed,

$$\begin{aligned} Bias(\boldsymbol{x}) &= \mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] \\ &= \mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\ &= \mu(\boldsymbol{x}) - \frac{1}{B}\sum_{b=1}^{B}\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\ &= \mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \end{aligned} \tag{3.8}$$

since estimates $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_1}(\boldsymbol{x}), \ldots, \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_B}(\boldsymbol{x})$ are identically distributed.
However, the bias of $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ is typically greater in absolute terms than the bias of $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ fitted on $\mathcal{D}$ since the reduced sample $\mathcal{D}^{*b}$ imposes restrictions. The improvements in the estimation obtained by bagging will be a consequence of variance reduction.
Notice that trees are ideal candidates for bagging. They can handle complex interaction structures in the data and they have relatively low bias if grown sufficiently deep. Because they are noisy, they will greatly benefit from the averaging.

3.3.2 Variance

The variance of $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ can be written as

$$\begin{aligned}
\text{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] &= \text{Var}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\
&= \frac{1}{B^2}\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\
&= \frac{1}{B^2}\Bigg\{\text{Var}_{\mathcal{D}}\left[\text{E}_{\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right] \\
&\qquad +\text{E}_{\mathcal{D}}\left[\text{Var}_{\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right]\Bigg\} \\
&= \text{Var}_{\mathcal{D}}\left[\text{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right] + \frac{1}{B}\text{E}_{\mathcal{D}}\left[\text{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right]
\end{aligned} \tag{3.9}$$

since conditionally to $\mathcal{D}$, estimates $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_1}(\boldsymbol{x}),\ldots,\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_B}(\boldsymbol{x})$ are independent and identically distributed. The second term is the within-$\mathcal{D}$ variance, a result of the randomization due to the bootstrap sampling. The first term is the sampling variance of the bagging ensemble, a result of the sampling variability of $\mathcal{D}$ itself. As the number of aggregated estimates gets arbitrarily large, i.e. as $B \to \infty$, the variance of $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ reduces to $\text{Var}_{\mathcal{D}}\left[\text{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right]$.
From (3.9) and

$$\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] = \text{Var}_{\mathcal{D}}\left[\text{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right] + \text{E}_{\mathcal{D}}\left[\text{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right], \tag{3.10}$$

we see that

$$\text{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] \leq \text{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]. \tag{3.11}$$

The variance of the bagging estimate $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ is smaller than the variance of an individual estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$. Actually, we learn from (3.9) and (3.10) that the variance reduction is given by

$$\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] - \text{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] = \frac{B-1}{B}\,\text{E}_{\mathcal{D}}\left[\text{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right], \tag{3.12}$$

which increases as B increases and tends to $\text{E}_{\mathcal{D}}\left[\text{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\Big|\mathcal{D}\right]\right]$ when $B \to \infty$.
Let us introduce the correlation coefficient $\rho(\boldsymbol{x})$ between any pair of estimates used in the averaging which are built on the same training set but fitted on two different bootstrap samples. Using the definition of the Pearson's correlation coefficient, we get

$$\begin{aligned}
\rho(\boldsymbol{x}) &= \frac{\text{Cov}_{\mathcal{D},\boldsymbol{\Theta}_b,\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}),\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})\right]}{\sqrt{\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]}\sqrt{\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})\right]}} \\
&= \frac{\text{Cov}_{\mathcal{D},\boldsymbol{\Theta}_b,\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}),\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})\right]}{\text{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]}
\end{aligned} \tag{3.13}$$

as $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ and $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})$ are identically distributed. By the law of total covariance, the numerator in (3.13) can be rewritten as

$$\begin{aligned}\mathrm{Cov}_{\mathcal{D},\boldsymbol{\Theta}_b,\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}),\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})\right] &= \mathrm{E}_{\mathcal{D}}\left[\mathrm{Cov}_{\boldsymbol{\Theta}_b,\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x}),\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})|\mathcal{D}\right]\right] \\ &\quad +\mathrm{Cov}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right],\mathrm{E}_{\boldsymbol{\Theta}_{b'}}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})|\mathcal{D}\right]\right] \\ &= \mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right] \qquad (3.14)\end{aligned}$$

since conditionally to $\mathcal{D}$, estimates $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ and $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_{b'}}(\boldsymbol{x})$ are independent and identically distributed. Hence, combining (3.10) and (3.14), the correlation coefficient in (3.13) becomes

$$\rho(\boldsymbol{x}) = \frac{\mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right]}{\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]} \qquad (3.15)$$

$$= \frac{\mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right]}{\mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\middle|\mathcal{D}\right]\right]+\mathrm{E}_{\mathcal{D}}\left[\mathrm{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\middle|\mathcal{D}\right]\right]}. \qquad (3.16)$$

The correlation coefficient $\rho(\boldsymbol{x})$ measures the correlation between a pair of estimates in the ensemble induced by repeatedly making training sample draws $\mathcal{D}$ from the population and then drawing a pair of bootstrap samples from $\mathcal{D}$.
When $\rho(\boldsymbol{x})$ is close to 1, the estimates are highly correlated, suggesting that the randomization due to the bootstrap sampling has no significant effect on the estimates. On the contrary, when $\rho(\boldsymbol{x})$ is close to 0, the estimates are de-correlated, suggesting that the randomization due to the bootstrap sampling has a strong impact on the estimates.
One sees that $\rho(\boldsymbol{x})$ is the ratio between the variance due to the training set and the total variance. The total variance is the sum of the variance due to the training set and the variance due to randomization induced by the bootstrap samples. A correlation coefficient close to 1 and hence correlated estimates means that the total variance is mostly driven by the training set. On the contrary, a correlation coefficient close to 0 and hence de-correlated estimates means that the total variance is mostly due to the randomization induced by the bootstrap samples.
Alternatively, the variance of $\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ given in (3.9) can be re-expressed in terms of the correlation coefficient. Indeed, from (3.15) and (3.16), we have

$$\mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right] = \rho(\boldsymbol{x})\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \qquad (3.17)$$

and

$$\mathrm{E}_{\mathcal{D}}\left[\mathrm{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right] = (1-\rho(\boldsymbol{x}))\,\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right], \qquad (3.18)$$

such that (3.9) can be rewritten as

$$\begin{aligned}\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] &= \mathrm{Var}_{\mathcal{D}}\left[\mathrm{E}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right]+\frac{1}{B}\mathrm{E}_{\mathcal{D}}\left[\mathrm{Var}_{\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})|\mathcal{D}\right]\right] \\ &= \rho(\boldsymbol{x})\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]+\frac{(1-\rho(\boldsymbol{x}))}{B}\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]. \qquad (3.19)\end{aligned}$$

As B increases, the second term disappears, but the first term remains. Hence, when $\rho(\boldsymbol{x}) < 1$, one sees that the variance of the ensemble is strictly smaller than the variance of an individual

model. Let us mention that assuming $\rho(\boldsymbol{x}) < 1$ amounts to suppose that the randomization due to the bootstrap sampling influences the individual estimates.
Notice that the random perturbation introduced by the bootstrap sampling induces a higher variance for an individual estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$ than for $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$, so that

$$\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \geq \mathrm{Var}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]. \tag{3.20}$$

Therefore, bagging averages models with higher variances. Nevertheless, the bagging estimate $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ has generally a smaller variance than $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$. This comes from the fact that, typically, the correlation coefficient $\rho(\boldsymbol{x})$ in (3.19) compensates for the variance increase $\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] - \mathrm{Var}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]$, so that the combined effect of $\rho(\boldsymbol{x}) < 1$ and $\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \geq \mathrm{Var}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right]$ often leads to a variance reduction

$$\mathrm{Var}_{\mathcal{D}}\left[\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})\right] - \rho(\boldsymbol{x})\mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \tag{3.21}$$

that is positive. Because of their high variance, regression trees very likely benefit from the averaging procedure .

3.3.3 Expected generalization error

For some loss functions, such as the squared error and Poisson deviance losses, we can show that the expected generalization error for the bagging estimate $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ is smaller than the expected generalization error for an individual estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$, that is,

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right]. \tag{3.22}$$

However, while it is typically the case with bagging trees, we cannot highlight some situations where the estimate $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ performs always better than $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ in the sense of the expected generalization error, even for the squared error and Poisson deviance losses.

3.3.3.1 Squared error loss

For the squared error loss, from (2.29), the expected generalization error for $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ is given by

$$\begin{aligned}\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] &= Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right]\right)^2 \\ &\quad + \mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right].\end{aligned} \tag{3.23}$$

From (3.8) and (3.11), one observes that the bias term remains unchanged while the variance decreases compared to the individual estimate $\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})$, so that we get

$$\begin{aligned}&\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \\ &\quad = Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]\right)^2 + \mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}}\left[\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] \\ &\quad \leq Err\left(\mu(\boldsymbol{x})\right) + \left(\mu(\boldsymbol{x}) - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right]\right)^2 + \mathrm{Var}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\ &\quad = \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right].\end{aligned} \tag{3.24}$$

For every value of $\boldsymbol{X}$, the expected generalization error of the ensemble is smaller than the expected generalization error of an individual model.
Taking the average of (3.24) over $\boldsymbol{X}$ leads to

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}\right)\right]. \tag{3.25}$$

3.3.3.2 Poisson deviance loss

For the Poisson deviance loss, from (2.30) and (2.31), the expected generalization error for $\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ is given by

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] = Err\left(\mu(\boldsymbol{x})\right) + \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \tag{3.26}$$

with

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] = 2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\frac{\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right] - 1 - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\ln\left(\frac{\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]\right). \tag{3.27}$$

We have

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\frac{\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right] = \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right], \tag{3.28}$$

so that (3.27) can be expressed as

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \\
&= 2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right] - 1 - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]\right) \\
&\quad -2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\ln\left(\frac{\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right] - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\left(\frac{\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})}{\mu(\boldsymbol{x})}\right)\right]\right) \\
&= \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right] \\
&\quad -2\mu(\boldsymbol{x})\left(\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\ln\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right]\right).
\end{aligned} \tag{3.29}$$

Jensen's inequality implies

$$\begin{aligned}
&\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\ln\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right] - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\
&= \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\ln\left(\frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right] - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\
&\geq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_1,\ldots,\boldsymbol{\Theta}_B}\left[\frac{1}{B}\sum_{b=1}^{B}\ln\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] - \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\ln\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right] \\
&= 0,
\end{aligned} \tag{3.30}$$

so that combining (3.29) and (3.30) leads to

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[\mathcal{E}^{\mathcal{P}}\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right] \tag{3.31}$$

and hence

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}(\boldsymbol{x})\right)\right]. \tag{3.32}$$

For every value of $\boldsymbol{X}$, the expected generalization error of the ensemble is smaller than the expected generalization error of an individual model.

Taking the average of (3.32) over $\boldsymbol{X}$ leads to

$$\mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] \leq \mathrm{E}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}\right)\right]. \tag{3.33}$$

Example

We consider an example in car insurance. Four features $\boldsymbol{X} = (X_1; X_2; X_3; X_4)$ are supposed to be available, that are

- X_1 = Gender : policyholder's gender (female or male) ;
- X_2 = Age : policyholder's age (integer values from 18 to 65) ;
- X_3 = Split : whether the policyholder splits its annual premium or not (yes or no) ;
- X_4 = Sport : whether the policyholder's car is a sports car or not (yes or no).

The variables X_1 ,X_2, X_3 and X_4 are assumed to be independent and distributed as follows :

$$\begin{aligned}
&\mathrm{P}[X_1 = female] = \mathrm{P}[X_1 = male] = 0.5;\\
&\mathrm{P}[X_2 = 18] = \mathrm{P}[X_2 = 19] = \ldots = \mathrm{P}[X_2 = 65] = 1/48;\\
&\mathrm{P}[X_3 = yes] = \mathrm{P}[X_3 = no] = 0.5;\\
&\mathrm{P}[X_4 = yes] = \mathrm{P}[X_4 = no] = 0.5.
\end{aligned} \tag{3.34}$$

The values taken by a feature are thus equiprobable. The response Y is supposed to be the number of claims. Given $\boldsymbol{X} = \boldsymbol{x}$, Y is assumed to be Poisson distributed with expected claim frequency given by

$$\begin{aligned}
\lambda(x) = \quad & 0.1 \times (1 + 0.1 I[x_1 = male])\\
& \times \left(1 + \frac{1}{\sqrt{x_2 - 17}}\right)\\
& \times (1 + 0.15 I[x_4 = yes]),
\end{aligned} \tag{3.35}$$

where $I[.]$ is the indicator function.

Being a male increases the expected claim frequency by 10%. The expected claim frequency smoothly decreases with the age, young drivers being more risky. Splitting its premium does not influence the expected claim frequency while driving a sports car increases the expected claim frequency by 15%. In this example, the true model $\lambda(x)$ is known and we can simulate realizations of the random vector $(Y, \boldsymbol{X})$.

We simulate training sets $\mathcal{D}$ made of 100 000 observations and validation sets $\overline{\mathcal{D}}$ of the same size. For each simulated training set $\mathcal{D}$, we build the corresponding tree $\widehat{\mu}_{\mathcal{D}}$ with tree depth being equal to 5 and we estimate its generalization error on a validation set $\overline{\mathcal{D}}$. Also, we generate bootstrap samples $\mathcal{D}^{*1}, \mathcal{D}^{*2}, \ldots$ of $\mathcal{D}$ and we produce the corresponding trees $\widehat{\mu}_{\mathcal{D}^{*1}}, \widehat{\mu}_{\mathcal{D}^{*2}}, \ldots$ with tree depth being equal to 5. We estimate their generalization errors on a validation set $\overline{\mathcal{D}}$, together with the generalization errors of the corresponding bagging models. Note that in this example, we use the `R` package `rpart` to build the different trees described above.

Figure 3.1 displays estimates of the expected generalization errors for $\widehat{\mu}_{\mathcal{D}}$, $\widehat{\mu}_{\mathcal{D}^{*b}} = \widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}$ and $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}$ for $B = 1, 2, \ldots, 10$ obtained by Monte-Carlo simulations. As expected, we notice that

$$\widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] \leq \widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}\right)\right].$$

For $B \geq 2$, bagging trees outperforms individual sample trees. Also, we note that

$$\widehat{\mathrm{E}}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right] \leq \widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}\right)\right],$$

showing that the restriction imposed by the reduced sample $\mathcal{D}^{*b}$ does not allow to build trees as predictive as trees built on the entire training set $\mathcal{D}$. Finally, from $B = 4$, we note that

$$\widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right] \leq \widehat{\mathrm{E}}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right],$$

meaning that for $B \geq 4$, bagging trees also outperforms single trees built on the entire training set.

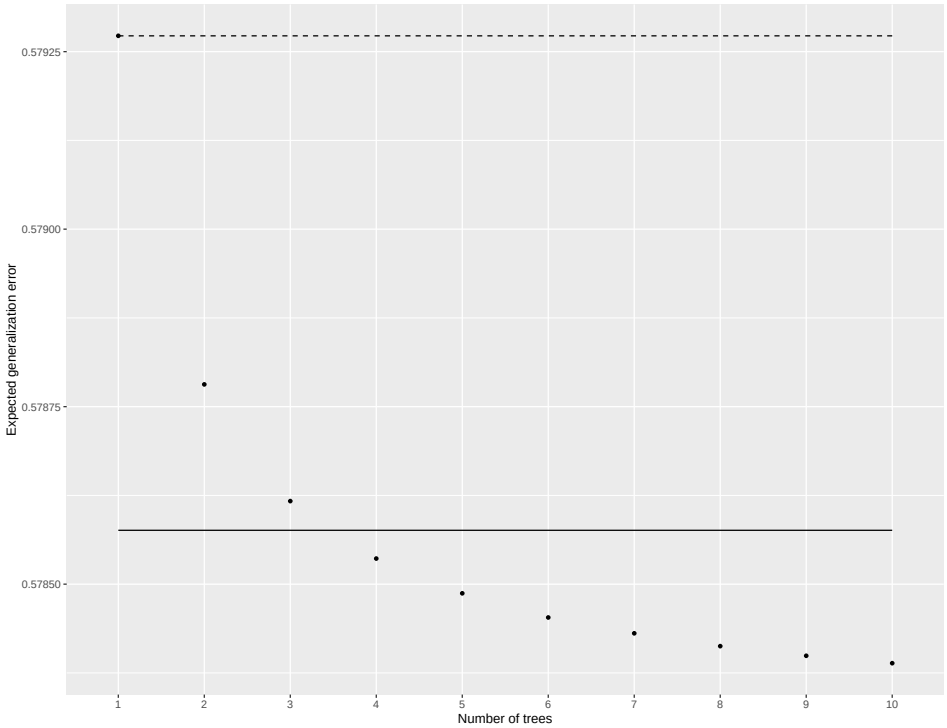

Figure 3.1: $\widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}}\left[Err\left(\widehat{\mu}^{\mathrm{bag}}_{\mathcal{D},\boldsymbol{\Theta}}\right)\right]$ with respect to the number of trees B, together with $\widehat{\mathrm{E}}_{\mathcal{D},\boldsymbol{\Theta}_b}\left[Err\left(\widehat{\mu}_{\mathcal{D},\boldsymbol{\Theta}_b}\right)\right]$ (dotted line) and $\widehat{\mathrm{E}}_{\mathcal{D}}\left[Err\left(\widehat{\mu}_{\mathcal{D}}\right)\right]$ (solid line).

Chapter 4

Conclusion

Two main drawbacks of regression trees are that they produce piece-wise constant estimates and that they are rather unstable under a small change in the observations of the training set. The construction of an ensemble of trees produces more stable and smoothed estimates under averaging.
Bagging is a technique used for reducing the variance of an estimate. Typically, it works well for high variance and low-bias procedures, such as regression trees.

In this note, we demonstrate and illustrate two elements:

- The expected generalization error of bagging trees is smaller than the expected generalization error of an individual estimate that constitutes the ensemble;
- The expected generalization error of the best regression tree fitted on the entire training set is smaller than the expected generalization error of an individual estimate constituting the ensemble.

Furthermore, we also illustrate by means of an example that bagging trees performs better than the best single tree built on the entire training set after a certain number of trees in the ensemble. Note that we cannot demonstrate that the estimate $\widehat{\mu}^{\text{bag}}_{\mathcal{D},\boldsymbol{\Theta}}(\boldsymbol{x})$ performs always better than $\widehat{\mu}_{\mathcal{D}}(\boldsymbol{x})$ in the sense of the expected generalization error.

References

Denuit, M., Hainaut, D., Trufin, J. (2020). Effective Statistical Learning Methods for Actuaries II: Tree-based Methods and Extensions. Springer Actuarial Lecture Notes Series.

Hastie, T., Tibshirani, R., Friedman, J. (2009). The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. Springer Series in Statistics.

Louppe, G. (2014). Understanding random forests: from theory to practice. arXiv:14077502.

Wüthrich, M. V., Buser, C. (2019). Data analytics for non-life insurance pricing. Lecture notes.

Chapter 5

About the serie and the authors...

5.1 The DetraNotes

The Detra Notes are a series of educational papers dedicated to the insurance sector. Those notes are published by members of the Detralytics team and written in a clear and accessible language. The team combines academic expertise and business knowledge. Detralytics was founded to support companies in the advancement of actuarial science and the solving of the profession's future challenges. It is within the scope of this mission that we make our work available through our DetraNotes and FAQctuary's series.

5.2 Authors' biographies

Candy Mahirwe

Candy is part of the Talent Consolidation Program (TCP) at Detralytics. During her various missions, Candy has worked on the ORSA report of a health insurance company in order to assess the adequacy and the completeness of the standard formula; on the creation of business requirements for the migration of the actuarial platform used for life cash-flows projections; and as a life product manager in the actuarial department of an insurance company. Prior to joining Detralytics, Candy worked as an intern at AG Insurance. Candy holds a Master's degree in Actuarial Sciences from ULB University.

Michel Denuit

Michel is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the Université Catholique de Louvain. Michel has established an international career for some two decades and has promoted many technical projects in collaboration with different actuarial market participants. He has written and co-written various books and publications. A full list of his publications is available at : `https://uclouvain.be/en/directories/michel.denuit` .

Julien Trufin

Julien is Scientific Director at Detralytics, as well as a Professor in Actuarial Science at the department of mathematics of the Université Libre de Bruxelles. Julien is a qualified actuary ofthe Instute of Actuaries in Belgium (IA|BE) and has experience as a consultant, as well as a compelling academic background developed in prominent universities such as Université Laval (Canada), UCL and ULB (Belgium).He has written and co-written various books and publications. A full list of his publications is available at : http://homepages.ulb.ac.be/~jtrufin/.

Expertise and innovation at the service of your future

www.detralytics.com - info@detralytics.eu