Introduction

Tree-based regression and classification

Random forest

Data analysis showcase

Best practices for health research

References

Practical Session
Prediction Modeling using Random Forest in R

Vanessa McNealis

McGill University

November 10th, 2021

Introduction

Tree-based regression and classification

Random forest

Data analysis showcase

Best practices for health research

References

Introduction

Software prerequisites

For this tutorial, you will need to have installed R and RStudio.

```
install.packages(c("randomForest", "tidyverse", "caret",
"ranger", "pmsampsize", "rms"))
library(randomForest)
library(tidyverse)
library(caret)
library(ranger)
library(pmsampsize)
library(rms)
```

Inference vs Prediction

$$\hat{Y} = x_1\hat{\beta}_1 + x_2\hat{\beta}_2 + \ldots + x_p\hat{\beta}_p$$

- **Inference**: Estimating the effect of a variable on the outcome while adjusting for confounding.
- **Prediction**: Predicting the outcome based on a set of covariate values.

Supervised learning

- Training data set $\mathcal{T}$: $\{(y_i, \boldsymbol{x}_i)\}$
- Using these data, we build a prediction model or *learner*, denoted $\phi(\boldsymbol{x}, \mathcal{T})$ (e.g., decision tree, linear regression model, neural network)
- Given an input $\boldsymbol{x}^*$, a prediction is given by

$$\hat{Y}^* = \phi(\boldsymbol{x}^*, \mathcal{T}).$$

Terminology

Basics

- Supervised learning
- Outcome
- Predictors
- Unsupervised learning

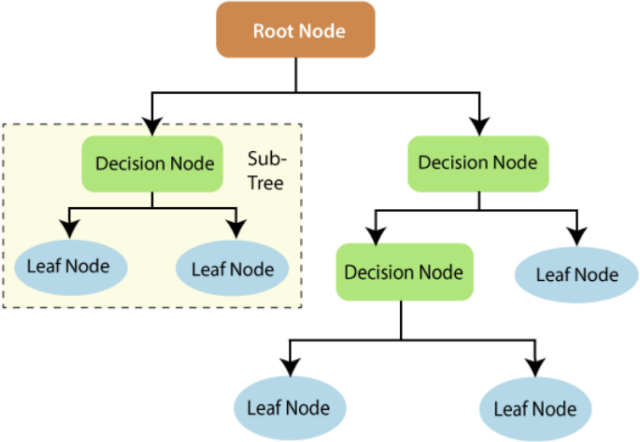
Tree-based learning

- Classification/regression
- Decision node
- Leaf node

Development and validation

- Predictive accuracy
- Hyperparameter tuning
- Cross-validation
- Bootstrap resampling
- Out-of-bag
- Generalization error
- Training/test sets

Decision tree



Source: https://www.tutorialandexample.com/decision-trees/

Bias, variance and model complexity

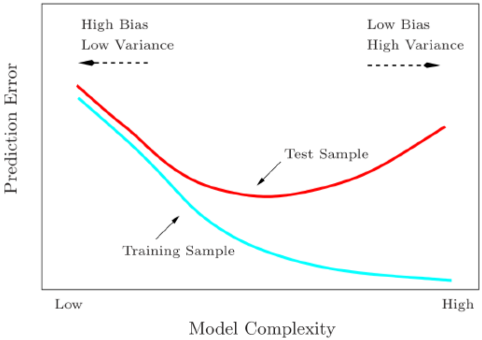


FIGURE 2.11. *Test and training error as a function of model complexity.*

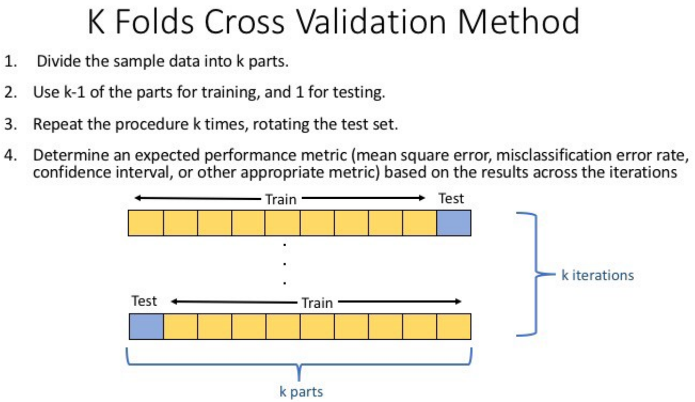
Source: Hastie et al. (2009)

Goal

- Minimize the prediction error, i.e., $\mathbb{E}\left[(\hat{Y} - Y)^2\right]$
- Control for over-fitting

Estimation of the generalization/test error

- Cross-validation
- Bootstrap resampling



Source: https://medium.com/@mtterribile/understanding-cross-validations-purpose-53490faf6a86

Tree-based regression and classification

The CART algorithm

Greedy algorithm:

1. At the root node: Scan through all inputs to find the best combination of splitting variable j and split point s.
   - Criterion for regression: Sum of squared errors $\sum(y_i - f(x_i))^2$
   - Criterion for classification: Measure of node impurity (e.g., Gini index)
2. Partition the data into the two resulting regions and repeat the splitting process on each of the two regions, and so on.

Question: How large should we grow the tree? The deeper the tree is grown, the lower the bias is !

Node impurity measures

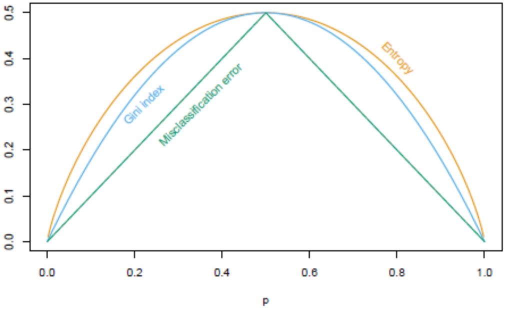


FIGURE 9.3. *Node impurity measures for two-class classification, as a function of the proportion p in class 2. Cross-entropy has been scaled to pass through* (0.5, 0.5).

Source: Hastie et al. (2009)

Pros and cons of trees

Pros:

- Interpretability
- Requires little effort in data preparation

Cons:

- Trees tend to be unstable, i.e., they tend to over-fit the data.
- *Bagging* averages many trees to reduce this variance.

Introduction

Tree-based regression and classification

Random forest

Data analysis showcase

Best practices for health research

References

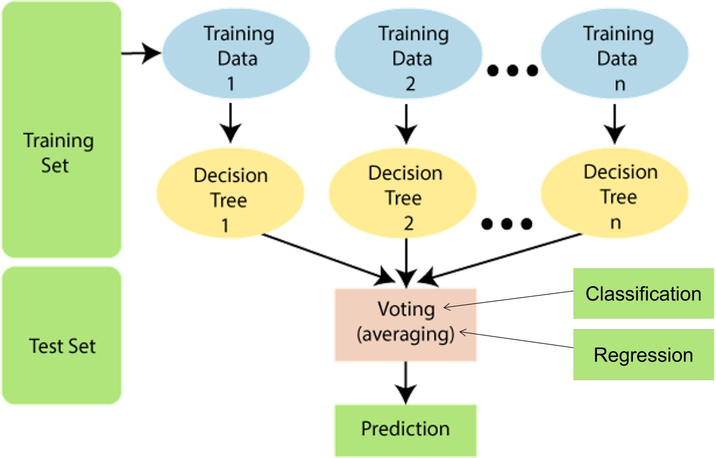
Random forest

Bagging and random forest (Breiman, 2001)

Bagging or *bootstrap aggregation* is a variance reduction method which works well for high-variance, low-bias procedures such as trees.

- Random: random selection of samples and features
- Forest: Ensemble of tree learners

Random forest



Source: https://www.javatpoint.com/machine-learning-random-forest-algorithm

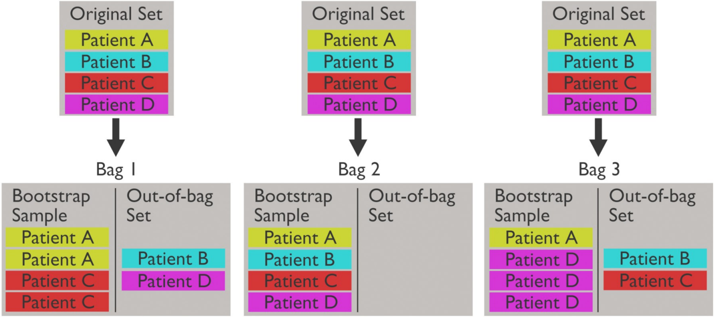
Hyperparameter tuning

A **hyperparameter** is a parameter of the model that is set prior to the start of the learning process.

- How many tree learners should we train? (`ntree` in `randomForest`, default is 500)
- How many features should be considered for splitting a node? (`mtry` in `randomForest`, default is $\sqrt{p}$ for classification)
- How deep should we grow each tree? (controlled by `maxnode` and `nodesize`)

Out-of-bag (OOB) error

Under the OOB method, the model is tested as it is being trained.



Source: https://en.wikipedia.org/wiki/Out-of-bag$_e$*rror*

Comparison of OOB and test errors

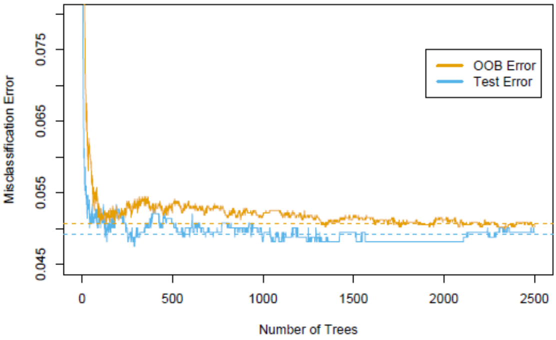


FIGURE 15.4. OOB *error computed on the* spam *training data, compared to the test error computed on the test set.*

Source: Hastie et al. (2009)

Introduction

Tree-based regression and classification

Random forest

Data analysis showcase

Best practices for health research

References

Data analysis showcase

Cardiotocography data set (Dua and Graff, 2017)

See the `markdown` file.

- 2126 fetal cardiotocograms (CTGs) were automatically processed and assessed by three expert obstetricians.
- **Outcome**: Fetal state (N, S, P)
  - Normal
  - Suspect
  - Pathologic
- **Potential predictors**: 21 features, including measurements of fetal heart rate (FHR) and uterine contraction (UC)

Introduction

Tree-based regression and classification

Random forest

Data analysis showcase

Best practices for health research

References

Best practices for health research

Sample size considerations

RESEARCH ARTICLE Open Access

Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints

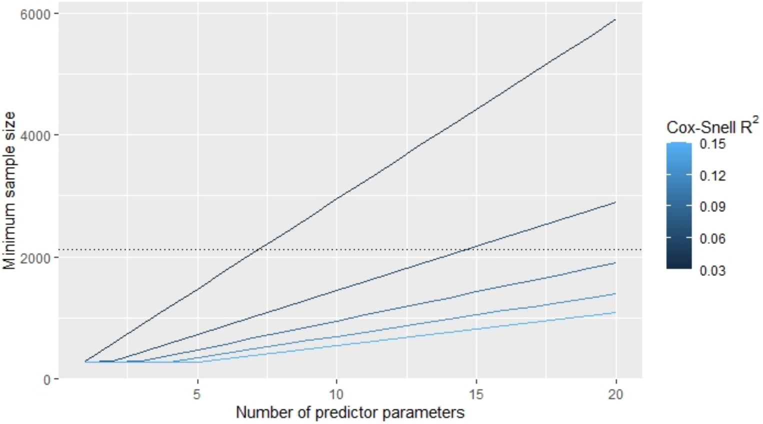
Tjeerd van der Ploeg[1,3*], Peter C Austin[2] and Ewout W Steyerberg[3]

"Modern modelling techniques such as [support vector machine], [neural network] and [random forest] may need over 10 times as many events per variable to achieve a stable AUC and a small optimism than classical modelling techniques such as [logistic regression]."

Sample size considerations

- Sample size required for classification is driven by the **number of events per predictor parameters/variables**.
- Each tree learner in the random forest incorporates complex interactions between features $\rightarrow$ Number of predictor parameters?
- Given a data set, a good practice is to first determine the budget of predictor parameters for conventional methods (regression).
- Package `pmsampsize` (Riley et al., 2020)

Sample size considerations



Selection of predictors

"[Automated solutions] allow us not to think about the problem " - FE Harrell

- To guard against over-fitting, avoid data-driven decisions.
  - Excluding a variable on the basis that it is non-significant.
  - "Exploratory" analyses.
- Prespecify predictor variables and use variables regardless of what the data tell you.
- Correlations between predictors may be examined for selection, but avoid looking at associations with the outcome.

Model validation

- **Apparent performance**: Predictive ability of a model on the same data from which the model was developed.
- Studies developing prediction models for diagnosis or prognosis should include some form of interval validation to quantify **optimism**.
- Randomly splitting a single data set into model training and model test/validation → Weak and **inefficient** approach to internal validation (Collins et al., 2015)
- **Solution**: Perform internal validation through bootstrap resampling

Model validation through the rms package

Table: Output of `rms::validate()` for a logistic regression model fitted on the `CTG` dataset

	index.orig	training	test	**optimism**	index.corrected
Dxy	0.92	0.92	0.91	0.00	0.91
R2	0.72	0.72	0.71	0.01	0.71
Intercept	0.00	0.00	-0.01	0.01	-0.01
Slope	1.00	1.00	0.97	0.03	0.97
Emax	0.00	0.00	0.01	0.01	0.01
D	0.63	0.64	0.63	0.01	0.62
U	-0.00	-0.00	0.00	-0.00	0.00
Q	0.63	0.64	0.63	0.01	0.62
B	0.07	0.07	0.07	-0.00	0.07
g	4.61	4.72	4.56	0.15	4.46
gp	0.32	0.32	0.32	0.00	0.31

Take-away message

- Given its nice variance reduction properties, random forest is a popular algorithm that can be used for classification or regression.
- More studies are warranted to understand the minimal sample size required for Random Forest or other out-of-the-box prediction methods. Should they be restricted to very large data sets?
- During the development phase of a model, data-driven decisions should be minimized to reduce the chance of over-fitting and ensure generalizability of the model.

References

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

Gary S Collins, Johannes B Reitsma, Douglas G Altman, and Karel GM Moons. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *Journal of British Surgery*, 102(3):148–158, 2015.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

Frank E Harrell et al. *Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis*. Springer, New York, 2001.

Frank E Harrell Jr. *rms: Regression Modeling Strategies*, 2021. URL `https://CRAN.R-project.org/package=rms`. R package version 6.2-0.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer, 2009.

Richard D Riley, Joie Ensor, Kym IE Snell, Frank E Harrell, Glen P Martin, Johannes B Reitsma, Karel GM Moons, Gary Collins, and Maarten Van Smeden. Calculating the sample size required for developing a clinical prediction model. *British Medical Journal*, 368, 2020.

Mark J Van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical Applications in Genetics and Molecular Biology*, 6(1), 2007.

Tjeerd van der Ploeg, Peter C Austin, and Ewout W Steyerberg. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC medical research methodology*, 14(1):1–13, 2014.