My Final College Paper

A Thesis
Presented to
The Division of Mathematics and Natural Sciences
Reed College

In Partial Fulfillment
of the Requirements for the Degree
Bachelor of Arts

Emerson H. Webb

May 2018

Approved for the Division
(Mathematics)

_____

Advisor F. Name

Acknowledgements

I want to thank a few people.

Preface

This is an example of a thesis setup to use the reed thesis document class (for LaTeX) and the R bookdown package, in general.

Table of Contents

List of Tables

List of Figures

Abstract

The preface pretty much says it all.

Second paragraph of abstract starts here.

Dedication

You can have a dedication here if you wish.

Chapter 1

Delete line 6 if you only have one advisor

Placeholder

Chapter 2

Introduction to Trees, Random Forests, and the Bootstrap

Placeholder

2.1 CART

2.2 Bootstrap

2.3 Random Forests

2.3.1 How Bagged Forests Work

2.3.2 The Random Forest Algorithm

2.3.3 Variable Importance Measures

2.3.4 Issues with Random Forests

2.4 Focus of this Thesis

2.5 Outline of Remaining Chapters

Chapter 3

Variable Importance and Inference for Random Forests

3.1 Introduction

In this chapter we focus on various approaches towards inference for random forests that have been developed. Generally, we could consider to be two approaches towards inference with random forests.

The first approach involves using variable importance measures of random forests to evaluate the relative importance of different variables in the construction of the forest. Ishwaran, Louppe, Owens, and Strobl et al. have been involved in work in this area. Louppe and Ishwaran focused on theoretical properties of variable importance measures while Owens, and Strobl and colleagues work is centered on developing less biased variable importance measures for hypothesis testing.

The second approach involves utilizing properties of the functional form of the random forest estimator to construct prediction intervals. Once predicton intervals have been constructed, depending on the context, confidence intervals and hypothesis tests can be produced. Mentch and Hooker's work involves interpreting the random forest ensemble as a particular type of U-statistics, and applying theoretical results about U-statistics to construct prediction intervals. Wager's work involves applying the infintisemal jackknife to the random forest estimator to produce prediction intervals. With both approaches there are asymptotic results that we will discuss in some detail. In addition, we will discuss some of the consistency results that have been proven for random forests. It is worth noting here that in mathematically analyzing the random forest algorithm, there are trade-offs between fidelity to the CART algorithm and amenability to mathematical tool. To our best knowledge, at this time there are no theoretical results showing that random forests constructed using CART are consistent. Generally, simplifications need to be made to the random forest algorithm to allow for asymptotic analysis. These simplifications usually involve using a different partitioning scheme than CART and also in working a sampling without replacement framework. We will note when such assumptions are being made when necessary. Otherwise, we assume a setting in which the random forest is constructed using CART

and bootstrap sampling with replacement.

3.2 Some Theory for Variable Importance Measures

In this section we discuss Louppe and Ishwaran's work exploring theoretical properties of variable importance measures.

3.2.1 Ishwaran's Variable Importance Measure

Ishwaran's approach to variable importance measures of random forests, as developed in "Variable Importance in Binary Regression Trees and Forests," differs from the original variable importance measures for random forests, so we require some additional vocabulary.

Suppose T is a binary recursive tree and suppose that T has M many terminal nodes. For each point $\mathbf{x}$ in the feature space, T maps $\mathbf{x}$ to one of the M-many terminal nodes. In particular, if we let $\mathcal{X}$ denote the feature space, then T is a function $\mathcal{X} \to \{1, \ldots, M\}$ defined by the equation

$$T(\mathbf{x}) = \sum_{m=1}^{M} mB_m(\mathbf{x}),$$

where $B_m(\mathbf{x})$ is a $0-1$ basis function which partition the feature space $\mathcal{X}$.

Let $Z = \{(\mathbf{x}_i, Y_i) | i = 1, \ldots, n\}$ denote the training data, where $\mathbf{x}_i$ is a covariate in the feature space and Y_i is the response. We call T a binary regression tree if it is a binary recursive tree grown from Z using using binary recursive splits of the form $x_v \leq c$ and $x_v > c$ where split values c are chosen based on the observed $\mathbf{x}_i$ in the training data Z. The value a_m in the terminal node is the average response of the training observations falling in the mth node. That is,

$$a_m = \frac{\sum_{i=1}^{n} \mathbb{I}\{T(\mathbf{x}_i) = m\} Y_i}{\sum_{i=1}^{n} \mathbb{I}\{T(\mathbf{x}_i) = m\}}.$$

For a binary regression tree, the basis functions $B_m(\mathbf{x})$ are product splines of the following form:

$$B_m(\mathbf{x}) = \prod_{l=1}^{L_m} [x_{l(m)} - c_{l,m}]_{s_{l,m}},$$

where L_m denotes the number of splits used to construct $B_m(\mathbf{x})$. For each split l, there is a splitting variable $\mathbf{x}_{l(m)}$ which denotes the $l(m)$th coordinate of $\mathbf{x}$ and a splitting value $c_{l,m}$. The $s_{l,m}$ are binary ± 1 values, where for a given scalar x, $[x]_{+1} = \mathbb{I}(x > 0)$ and $[x]_{-1} = \mathbb{I}(x \leq 0)$. Note that the basis functions satisfy an orthogonality property, which gives $B_m(\mathbf{x})B_{m'}(\mathbf{x}) = 0$ if $m \neq m'$. Note also that given a tree T, the predictor associated with the tree can be written as a linear combination of basis functions:

$$\hat{\mu}(\mathbf{x}) = \sum_{m=1}^{M} a_m B_m(\mathbf{x}).$$

We are now prepared to define Ishwaran's variable importance measure.

Informally, the MDA variable importance of a variable x_j is the difference between MSE of the tree T when x_j is randomly permuted and MSE of the tree T when x_j is not permuted. As such a scheme of variable importance is difficult to analyze, Ishwaran proposes a surrogate measure. For the variable x_j, we drop $\mathbf{x}$ down the tree and follow the binary splits until either a terminal node is reached or a node with a split depending on x_j is reached. If a node with a split depending on x_j is reached, we then subsequently assign $\mathbf{x}$ randomly to either the left or right daughter node, whenever there is a split, until we reach a terminal node. The difference in MSE between noising up x_j and not noising up x_j to be the variable importance of x_j in the tree T. Denote the tree that results from noising up x_j by T_j.

Such a scheme relies on the following heuristic: if we chose an adequete splitting rule to construct our tree, then we expect that variables that are split earlier in the tree are more important, since prediction will suffer the most from noising up a variable higher up in the tree than a variable close to a terminal node. This is a behavior observed in CART trees and random forests based on CART: splits closer to the root node are more influential than splits close to terminal nodes, so the MDA or MDI variable importance of variables split on close to the root node are expected to be higher than otherwise.

3.2.2 Maximal Subtrees and Theoretical Results

Defining a structure on binary regression trees called subtrees, Ishwaran is able to write the predictor for the noised up tree as a deterministic component relying on terminal nodes for no parent nodes involve a split on x_j, and a random component involving terminal nodes for which there are parent nodes involving a split on x_j. The definition of the subtree is quite intuitive. We call $\tilde{T}_j$ a j-subtree of the tree T, if the root node of $\tilde{T}_j$ has daughters that depend on an x_j split. A j-subtree $\tilde{T}_j$ is a maximal j-subtree of the tree T, if there are no larger j-subtrees continaing $\tilde{T}_j$. For a given tree T and for each variable x_j, there is a set of K_j many distinct maximal j-subtrees, which are denoted by $\tilde{T}_{1,j}, \ldots, \tilde{T}_{K_j,j}$. Note each distinct $\tilde{T}_{k,j}$ maximal j-subtree contains a set of distinct terminal nodes $M_{k,j}$. Each $M_{k,j}$ is distinct for $k = 1, \ldots, K_j$, since we are working with maximal j-subtrees. Define

$$M_j = \bigcup_{k=1}^{K_j} M_{k,j}$$

to be the set of terminal nodes for which there is a parent node involving a split on x_j. Ishwaran proves the following lemma about the functional form of the predictor for the tree T_j.

Lemma 3.1. *Let $\hat{\mu}_j(\mathbf{x})$ denote the predictor for T_j. Then*

$$\hat{\mu}_j(\mathbf{x}) = \sum_{m \notin M_j} a_m B_m(\mathbf{x}) + \sum_{k=1}^{K_j} \tilde{a}_{k,j} \mathbb{I}\{T(\mathbf{x}) \in M_{k,j}\},$$

where $\tilde{a}_{k,j}$ is the random terminal value assigned by $\tilde{T}_{k,j}$ under the random left right path through $\tilde{T}_{k,j}$. We write $\tilde{P}_{k,j}$ to denote the distribution of $\tilde{a}_{k,j}$.

For a proof, the reader is referred to Ishwaran's paper. It is a bit surprising that the functional form of $\hat{mu}_j(\mathbf{x})$ can be separated into the two components. Given this lemma and the definition of j-subtrees, we can more formally define the variable importance of x_j in the tree T.

Let g be a loss function. Often we use the squared error to evaluate loss, which corresponds to MSE, but there is no strict requirement in the definition. Denote the test data by $(Y, \mathbf{x})$. Then the prediction error of the predictor $\hat{\mu}$ is given by $\mathbb{E}(g(Y, \hat{\mu}(\mathbf{x})))$. As a reminder, we assume that there is an underlying regression function

$$Y = \mu(\mathbf{x}) + \varepsilon,$$

where ε is independent error with zero mean and variance $\sigma^2 > 0$. Similarly, we can define the prediction error of the predictor $\hat{\mu}_j$ to be $\mathbb{E}(g(Y, \hat{\mu}_j(\mathbf{x})))$. Set $g(Y, \hat{\mu}(\mathbf{x})) = (Y - \hat{\mu}(\mathbf{x}))^2$ to be the L_2 loss, which corresponds to MSE. Define the variable importance of the variable x_j to be

$$\Delta_j = \mathbb{E}((Y - \hat{\mu}_j(\mathbf{x}))^2) - \mathbb{E}((Y - \hat{\mu})^2).$$

Application of the lemma and some manipulation allows us to write

$$\Delta_j = \mathbb{E}(R_j(\mathbf{x})^2) - 2\mathbb{E}\left(R_j(\mathbf{x})[\mu(\mathbf{x}) - \hat{\mu}(\mathbf{x})]\right),$$

where

$$R_j(\mathbf{x}) = \sum_{k=1}^{K_j} \sum_{m \in M_{k,j}} (\tilde{a}_{k,j} - a_m) B_m(\mathbf{x}).$$

To aid in his analysis, Ishwaran makes the assumption that the true regression function μ is of similar form to T. That is, assume

$$\mu(\mathbf{x}) = \sum_{m=1}^{M} a_{m,0} B_m(\mathbf{x}),$$

where $a_{m,0}$ are the true, but unknown, terminal values. Under this and some other large sample assumptions, Ishwaran finds that asymptotically, each maximal j-subtree will tend to contribute equally to the variable importance Δ_j. In effect, Ishwaran finds that nodes closer to the root of a maximal j-subtree will have a larger effect on Δ_j than nodes closer to the terminal nodes.

3.2.3 Extension to Forest Ensembles

Ishwaran's framework extends naturally to forest ensembles and his theoretical result regarding forest ensembles provides some information of the behavior of variable importance measures for random forests. First for some notation, recall that in the forest ensemble setting, we draw B many bootstrap resamples of the training data to

obtain the bootstrap replicates $Z^b = \{(\mathbf{x}_i^b, Y_i^b) | i = 1, \ldots, n\}$ of the training data for $b = 1, \ldots, B$. We then construct a binary regression tree $T(\mathbf{x}; b)$ on each bootstrap replicate of the data and have the forest $\hat{\mu}_F$ as the average of predictions over the trees $T(\mathbf{x}; b)$:

$$\hat{\mu}_F(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \hat{\mu}(\mathbf{x}; b),$$

where $\hat{\mu}(\mathbf{x}; b)$ denotes the predictor for the tree $T(\mathbf{x}; b)$. Given that each $\hat{\mu}(\mathbf{x}; b)$ is a linear combination of basis functions, we can write

$$\hat{\mu_F}(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{m=1}^{M^b} a_m^b B_m(\mathbf{x}; b).$$

Again assume that the $\mu(\mathbf{x})$ has a similar structure to $\mu(\mathbf{x})$. That is, $\mu(\mathbf{x}) = \frac{1}{B} \sum_{b=1}^{B} \sum_{m=1}^{M^b} a_{m,0}^b B_m(\mathbf{x}; b)$.

Chapter 4

The Bootstrap

Conclusion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Appendix A

The First Appendix

This first appendix includes all of the R chunks of code that were hidden throughout the document (using the `include = FALSE` chunk tag) to help with readibility and/or setup.

In the main Rmd file

In Chapter ??:

Appendix B

The Second Appendix, for Fun

References

Placeholder