Selected Topics In Data Science

Bruce Campbell

2020-12-29

Contents

1 Preface **5**

2 Introduction **7**

3 On Model Averaging **9**

4 Sensitivity Analysis and Shapley Values **11**
4.1 Relationship between Sobol indices and Shapley values 11

5 Applications **13**
5.1 Example one 13
5.2 Example two 13

6 Final Words **15**

Chapter 1

Preface

This is the first installment on my promise to elucidate less popular topics in statistics and machine learning. I wrote this as a way to solidify my understanding of some of the topics that are treated here. Hopefully others will find value here.

Chapter 2

Introduction

This is a living book. It's under development. We are using the **bookdown** package (Xie, 2020) in this book, which was built on top of R Markdown and **knitr** (Xie, 2015).

Chapter 3

On Model Averaging

Recall that we can break down model error into the bias an variance $bias(\hat{Y}) = E[\hat{Y} - E[Y]]$

If we are averaging models $i = 1, \cdots, k$ then

$\text{MSE}\left(\hat{Y}_i\right) = \left\{\text{bias}\left(\hat{Y}_i\right)\right\}^2 + \text{var}\left(\hat{Y}_i\right)$

Chapter 4

Sensitivity Analysis and Shapley Values

Global sensitivity analysis measures the importance of input variables to a function. This is an important task in quantifying the uncertainty in which target variables can be predicted from their inputs. Sobol indices are a popular approach to this. It turns out that there's a relationship between Sobol indices and Shapley values. We explore this relationship here and demonstrate their effectiveness on some linear and non-linear models.

4.1 Relationship between Sobol indices and Shapley values

Shapley values are based on $f(x) - E[f(x)]$ while Sobol indices decompose output variance into fractions contributed by the inputs. The Sobol index is a global measure of feature importance while Shapley values focus on local explanations although we could combine local Shapley values to achieve a global importance measure. Sobol indices are based on expectations and can be used for features not included in the model / function of interest. In this way we could query for important features correlated with those that the model does use.

Chapter 5

Applications

Some *significant* applications are demonstrated in this chapter.

5.1 Example one

5.2 Example two

Chapter 6

Final Words

We have finished a nice book.

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2020). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.21.