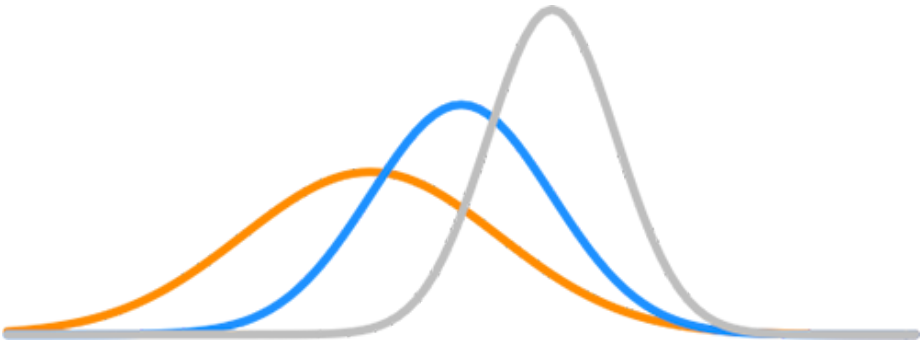
Basics of Statistical Learning

David Dalpiaz

2019-08-23

Contents

Introduction



Welcome to R for Statistical Learning! While this is the current title, a more appropriate title would be "Machine Learning from the Perspective of a Statistician using R" but that doesn't seem as catchy.

About This Book

This book currently serves as a supplement to An Introduction to Statistical Learning for STAT 432 - Basics of Statistical Learning at the University of Illinois at Urbana-Champaign.

The initial focus of this text was to expand on ISL's introduction to using R for statistical learning, mostly through adding to and modifying existing code. This text is currently becoming much more self-contained. Why? A very good question considering that the author consider ISL one of the best undergraduate textbooks available, and was one of the driving forces for the creation of STAT 432. However once the course was created, exact control over content became extremely useful. The main focus of this text is to match the needs of students in that course. Some of those needs include:

- Additional R code examples and explanation
- Simulation studies

- Mathematical rigor that matches the background of the readers
- A book structure that matches the overall structure of the course

In other words, this text seeks to replicate the best parts of An Introduction to Statistical Learning, The Elements of Statistical Learning, and Applied Predictive Modeling that are most needed by a particular set of students.

Organization

The text is organized into roughly seven parts.

1. Prerequisites
2. (Supervised Learning) Regression
3. (Supervised Learning) Classification
4. Unsupervised Learning
5. (Statistical Learning) in Practice
6. (Statistical Learning) in The Modern Era
7. Appendix

Part 1 details the assumed prerequisite knowledge required to read the text. It recaps some of the more important bits of information. It is currently rather sparse.

Parts 2, 3, and 4 discuss the **theory** of statistical learning. Several methods are introduced throughout to highlight different theoretical concepts.

Parts 5 and 6 highlight the use of statistical learning in **practice**. Part 5 focuses on practical usage of the techniques seen in Parts 2, 3, and 4. Part 6 introduces techniques that are most commonly used in practice today.

Who?

This book is targeted at advanced undergraduate or first year MS students in Statistics who have no prior statistical learning experience. While both will be discussed in great detail, previous experience with both statistical modeling and `R` are assumed.

Caveat Emptor

This "book" is under active development. Much of the text was hastily written during the Spring 2017 run of the course. While together with ISL the coverage is essentially complete, significant updates are occurring during Fall 2017.

When possible, it would be best to always access the text online to be sure you are using the most up-to-date version. Also, the html version provides additional features such as changing text size, font, and colors. If you are in need of a local copy, a **pdf version** is continuously maintained. While development is taking place, formatting in the pdf version may not be as well planned as the html version since the html version does not need to worry about pagination.

Since this book is under active development you may encounter errors ranging from typos, to broken code, to poorly explained topics. If you do, please let us know! Simply send an email and we will make the changes as soon as possible. (`dalpiaz2 AT illinois DOT edu`) Or, if you know `rmarkdown` and are familiar with GitHub, make a pull request and fix an issue yourself! This process is partially automated by the edit button in the top-left corner of the html version. If your suggestion or fix becomes part of the book, you will be added to the list at the end of this chapter. We'll also link to your GitHub account, or personal website upon request.

While development is taking place, you may see "TODO" scattered throughout the text. These are mostly notes for internal use, but give the reader some idea of what development is still to come.

Please see the [README] file on GitHub for notes on the development process.

Conventions

0.0.1 Mathematics

This text uses MathJax to render mathematical notation for the web. Occasionally, but rarely, a JavaScript error will prevent MathJax from rendering correctly. In this case, you will see the "code" instead of the expected mathematical equations. From experience, this is almost always fixed by simply refreshing the page. You'll also notice that if you right-click any equation you can obtain the MathML Code (for copying into Microsoft Word) or the TeX command used to generate the equation.

$$a^2 + b^2 = c^2$$

Often the symbol $\triangleq$ will be used to mean "is defined to be."

We use the value p to mean the number of **p**redictors. We will use n for sample size.

0.0.2 Code

`R` code will be typeset using a `monospace` font which is syntax highlighted.

```
a = 3
b = 4
sqrt(a ^ 2 + b ^ 2)
```

`R` output lines, which would appear in the console will begin with `##`. They will generally not be syntax highlighted.

```
## [1] 5
```

For the most part, we will follow the `tidyverse` style guide, however with one massive and obvious exception. Instead of the usual assignment operator, `<-`, we will instead use the more visually appealing and easier to type `=`. Not many do this, but there are dozens of us.

Acknowledgements

The following is a (likely incomplete) list of helpful contributors.

- James Balamuta, Summer 2016 - ???
- Korawat Tanwisuth, Spring 2017
- Yiming Gao, Spring 2017
- Binxiang Ni, Summer 2017
- Ruiqi (Zoe) Li, Summer 2017
- Haitao Du, Summer 2017
- Rachel Banoff, Fall 2017
- Chenxing Wu, Fall 2017
- Wenting Xu, Fall 2017
- Yuanning Wei, Fall 2017
- Ross Drucker, Fall 2017
- Craig Bonsignore, Fall 2018
- Ashish Kumar, Fall 2018

Your name could be here! If you submit a correction and would like to be listed below, please provide your name as you would like it to appear, as well as a link to a GitHub, LinkedIn, or personal website. Pull requests encouraged!

Looking for ways to contribute?

- You'll notice that a lot of the plotting code is not displayed in the text, but is available in the source. Currently that code was written to accomplish a task, but without much thought about the best way to accomplish the task. Try refactoring some of this code.
- Fix typos. Since the book is actively being developed, typos are getting added all the time.
- Suggest edits. Good feedback can be just as helpful as actually contributing code changes.

Figure 1: This work is licensed under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License.

License