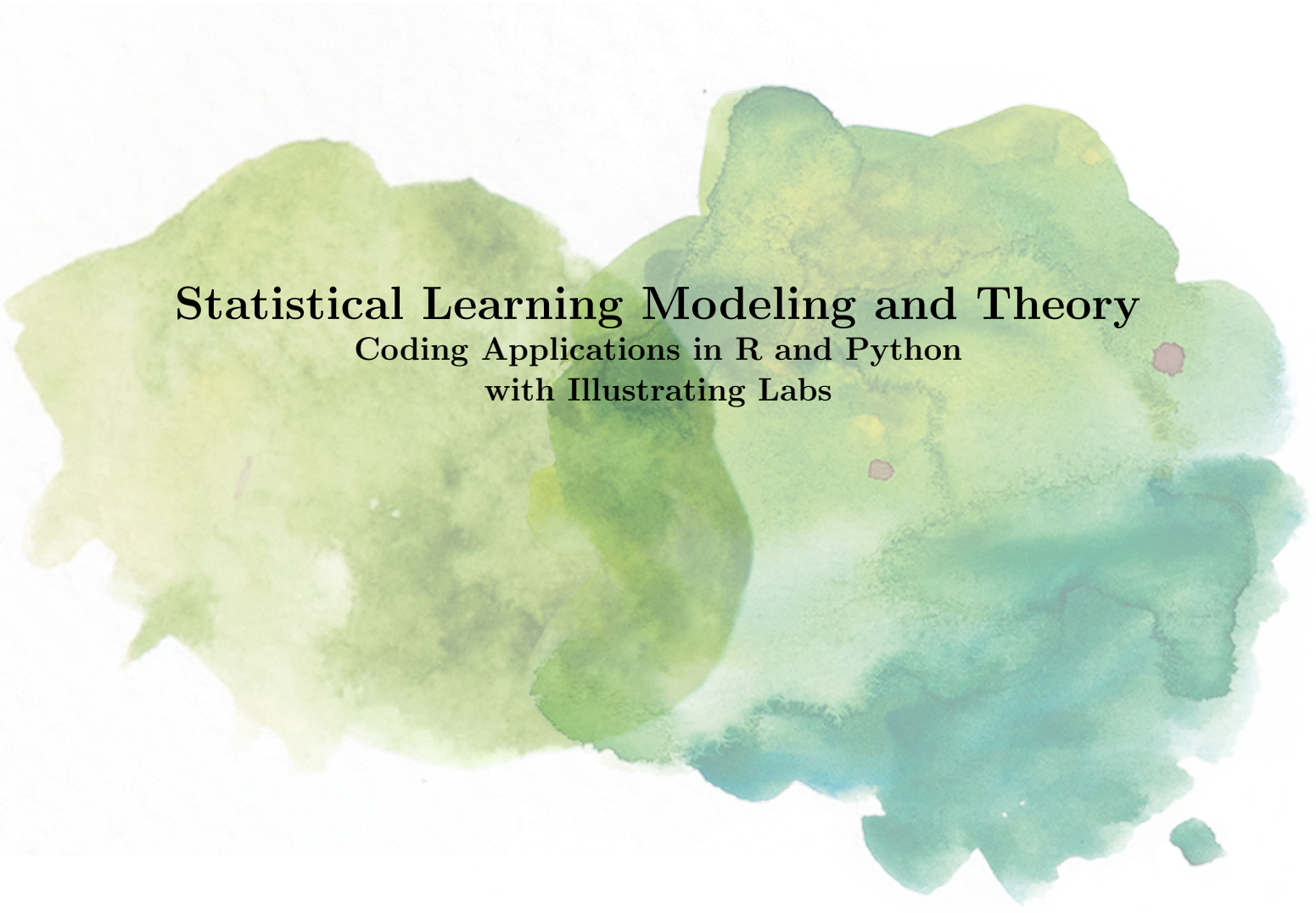

Statistical Learning Modeling and Theory

Coding Applications in R and Python with Illustrating Labs

"Big data isn't about bits,
it's about talent."
– Douglas Merill

Contents

Chapter 1

Introduction

General Outline for the Intro:

1. The growth of information.
2. The rise of data and AI; "Data is the new oil" - Clive Humby, and "AI is the new electricity" - Andrew Ng
3. How statistical learning fits into the picture; what it does; and why it's important.
4. What the book covers, and what the book can offer.
5. Who the book is "for".
6. How to use the book.

Chapter 2

Statistical Learning

2.1 The Basics of Statistical Learning

Generally speaking, statistical learning is a construct that attempts to predict an outcome, or determine a relationship. The act of predicting an outcome falls under a type of learning called *supervised learning*. Supervised learning requires that you have an *output*, Y. Determining a relationship falls under a type of learning called *unsupervised learning*, which does not require an output. Considering *supervised learning*, the output variable Y, can be either *quantitative*, or *qualitative*. *Quantitative* outputs are numeric and continuous, or close in nature. *Qualitative* outputs are categorical, or discrete in nature.
We introduce the notation for input variable(s), X:

$$X = (X_1, X_2, ..., X_P). \tag{2.1}$$

It is worth taking the time to stop, and make a mental note of the various names that are synonomous to the term: *variable*. Within the scope of statistical learning, the term *variable* is interchangeable with *predictors* (hence the P in X_P), *independent variables*, or *features*. While, the output variable is often referred to as the *response*, or *dependent variable*.

2.1.1 Types of Estimation

Supervised learning can be split into two different kinds of estimation: *prediction* and *inference*. When the time comes, you may determine that one estimation better frames your question than the other. Conversely, you may determine that

your question is best answered with the utilization of both types of estimation. Nevertheless, both forms of estimation offer unique insight, and require unique paramaters. We discuss each in depth.

Prediction

The first type of estimation is referred to prediction, which is implemented when we have information about X, and we want to predict Y. Let's say that we have a quantative (numerical) response variable, Y, with predictors $X = (X_1, X_2, ...X_P)$. We can set up a very general equation that predicts Y using

$$\hat{Y} = \hat{f}(X) + \epsilon, \tag{2.2}$$

based on the assumption that there is in fact some relationship between Y, and $X = (X_1, X_2, ...X_P)$, which is assumed to have the form:

$$Y = f(X) + \epsilon. \tag{2.3}$$

Here, $\hat{Y}$ is our resulting prediction for Y, $\hat{f}$ is our estimate for f, which represents the systematic information that X carries about Y, and ϵ is the random error term. The goal in a prediction envorinment is to have values for $\hat{Y}$ that are accurate with respect to Y, or the actual outcome. In this context, we are not concerned with the question of *"what is the true form of f?"*, but rather, how accurate are our precictions of Y. Lastly, ϵ refers to the unavoidable error term that will surely be present in any prediction. This random error is independent of X, has a mean of zero, and can account for the randomness in the real-world that is naturally occurring. Since ϵ has a mean of xero, Y can be predicted by using

$$\hat{Y} = \hat{f}(X) \tag{2.4}$$

Inference

Discussion about inference