

Statistical Learning with Math and Python

Joe Suzuki

Statistical Learning with Math and Python

100 Exercises for Building Logic

 Springer

Joe Suzuki
Graduate School of Eng Sci
Osaka University
Toyonaka, Osaka, Japan

ISBN 978-981-15-7876-2 ISBN 978-981-15-7877-9 (eBook)
<https://doi.org/10.1007/978-981-15-7877-9>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2021

This work is subject to copyright. All rights are solely and exclusively licensed by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

I am currently with the Statistics Laboratory at Osaka University, Japan. I often meet with data scientists who are engaged in machine learning and statistical analyses for research collaborations and introducing my students to them. I recently found out that almost all of them think that (mathematical) logic rather than knowledge and experience is the most crucial ability for grasping the essence in their jobs. Our necessary knowledge is changing every day and can be obtained when needed. However, logic allows us to examine whether each item on the Internet is correct and follow any changes; without it, we might miss even chances.

In 2016, I started teaching statistical machine learning to the undergraduate students of the Mathematics Department. In the beginning, I was mainly teaching them what (statistical) machine learning (ML) is and how to use it. I explained the procedures of ML, such as logistic regression, support vector machines, k-means clustering, etc., by showing figures and providing intuitive explanations. At the same time, the students tried to understand ML by guessing the details. I also showed the students how to execute the ready-made functions in several R packages without showing the procedural details; at the same time, they understood how to use the R packages as black boxes.

However, as time went by, I felt that this manner of teaching should be changed. In other non-ML classes, I focus on making the students consider extending the ideas. I realized that they needed to understand the essence of the subject by mathematically considering problems and building programs. I am both a mathematician and an R/Python programmer and notice the importance of instilling logic inside each student. The basic idea is that the students see that both theory and practice meet and that using logic is necessary.

I was motivated to write this book because I could not find any other book that was inspired by the idea of “instilling logic” in the field of ML.

The closest comparison is “Introduction to Statistical Learning: with Application in R” (ISLR) by Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani (Springer), which is the most popular book in this field. I like this book and have used it for the aforementioned class. In particular, the presentation in the book is splendid (abundant figures and intuitive explanations). I followed this style when

writing this book. However, ISLR is intended for a beginner audience. Compared with ISLR, this book (SLMP) focuses more on mathematics and programming, although the contents are similar: linear regression, classification, information criteria, regularizations, decision trees, support vector machine, and unsupervised learning.

Another similar book is “The Elements of Statistical Learning” (ESL) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (Springer), which provides the most reliable knowledge on statistical learning. I often use it when preparing for my classes. However, the volume of information in ESL is large, and it takes at least 500–1000 h to read it through, although I do recommend reading the book. My book, SLMP, on the other hand, takes at most 100 h, depending on the reader’s baseline ability, and it does not assume the reader has any knowledge of ML. After reading SLMP, it takes at most 300–500 h to read through ESL because the reader will have enough logic to easily understand ESL. ESL contains many equations and procedures but no programming codes.

In this sense, SLMP focuses on both mathematics and programming more than ISLR and ESL. I sincerely wish that the reader of SLMP will develop both logic and statistical learning knowledge.

What Makes SLMP Unique?

I have summarized the features of this book as follows.

1. Developing logic

To grasp the essence of the subject, we mathematically formulate and solve each ML problem and build those programs. The SLMP instills “logic” in the minds of the readers. The reader will acquire both the knowledge and ideas of ML, so that even if new technology emerges, they will be able to follow the changes smoothly. After solving the 100 problems, most of the students would say “I learned a lot.”

2. Not just a story

If programming codes are available, you can immediately take action. It is unfortunate when an ML book does not offer the source codes. Even if a package is available, if we cannot see the inner workings of the programs, all we can do is input data into those programs. In SLMP, the program codes are available for most of the procedures. In cases where the reader does not understand the math, the codes will help them understand what it means.

3. Not just a how-to book: an academic book written by a university professor

This book explains how to use the package and provides examples of executions for those who are not familiar with them. Still, because only the inputs and outputs are visible, we can only see the procedure as a black box. In this sense, the reader will have limited satisfaction because they will not be able to obtain

the essence of the subject. SLMP intends to show the reader the heart of ML and is more of a full-fledged academic book.

4. Solve 100 exercises: problems are improved with feedback from university students

The exercises in this book have been used in university lectures and have been refined based on feedback from students. The best 100 problems were selected. Each chapter (except the exercises) explains the solutions, and you can solve all of the exercises by reading the book.

5. Self-contained

All of us have been discouraged by phrases such as “for the details, please refer to the literature XX.” Unless you are an enthusiastic reader or researcher, nobody will seek out those references. In this book, we have presented the material in such a way that consulting external references is not required. Additionally, the proofs are simple derivations, and the complicated proofs are given in the appendices at the end of each chapter. SLMP completes all discussions, including the appendices.

6. Readers’ pages: questions, discussion, and program files

The reader can ask any question on the book’s Facebook page (<https://bayesnet.org/books>). Additionally, all programs and data can be downloaded from <http://bitbucket.org/prof-joe> (thus, you do not have to copy the programs from the book).

7. Linear algebra

One of the bottlenecks in learning ML and statistics is linear algebra. Except for books for researchers, a few books assume the reader has knowledge of linear algebra, and most books cannot go into the details of this subject. Therefore, SLMP contains a summary of linear algebra. This summary is only 17 pages and is not just an example, but it provides all the proofs. If you already know linear algebra, then you can skip it. However, if you are not confident in the subject, you can read in only one day.

How to Use This Book

Each chapter consists of problems, their explanation (body), and an appendix (proof, program). You can start reading the body and solve the problem. Alternatively, you might want to solve the 100 exercises first and consult the body if necessary. Please read through the entire book until the end.

When used in a lecture, I recommend that the teacher organizes the class into 12, 90 min lectures (or a 1000 min course) as follows: 3 lectures for Chap. 1, 2 lectures for Chap. 6, and 1 lecture for each of the other chapters. You may ask the students to complete the 100 exercises. If you read the text carefully, you will be able to answer any of their questions. I think that the entire book can be fully read in about 12 lectures total.

Acknowledgments

The author wishes to thank Yuske Inaoka, Tianle Yang, Ryosuke Shinmura, Junichi Maruyama, and Kazuya Morishita for checking the manuscript. This English book is largely based on the Japanese book published by Kyoritsu Shuppan Co., Ltd, in 2020. The author would like to thank Kyoritsu Shuppan Co., Ltd, for their generosity. The author also appreciates Ms. Mio Sugino, Springer, for preparing the publication and providing advice on the manuscript.

Osaka, Japan
May 2021

Joe Suzuki

Contents

1	Linear Algebra	1
1.1	Inverse Matrix	1
1.2	Determinant.....	3
1.3	Linear Independence	6
1.4	Vector Spaces and Their Dimensions	8
1.5	Eigenvalues and Eigenvectors	11
1.6	Orthonormal Bases and Orthogonal Matrix	12
1.7	Diagonalization of Symmetric Matrices	13
	Appendix: Proofs of Propositions	15
2	Linear Regression	19
2.1	Least Squares Method.....	19
2.2	Multiple Regression	22
2.3	Distribution of $\hat{\beta}$	25
2.4	Distribution of the RSS Values	26
2.5	Hypothesis Testing for $\hat{\beta}_j \neq 0$	29
2.6	Coefficient of Determination and the Detection of Collinearity ...	35
2.7	Confidence and Prediction Intervals.....	39
	Appendix: Proofs of Propositions	42
	Exercises 1–18	43
3	Classification	53
3.1	Logistic Regression	53
3.2	Newton–Raphson Method	56
3.3	Linear and Quadratic Discrimination.....	61
3.4	k -Nearest Neighbor Method	65
3.5	ROC Curves	66
	Exercises 19–31	69
4	Resampling	77
4.1	Cross-Validation	77
4.2	CV Formula for Linear Regression.....	81

4.3	Bootstrapping	85
	Appendix: Proofs of Propositions	89
	Exercises 32–39	90
5	Information Criteria	95
5.1	Information Criteria	95
5.2	Efficient Estimation and the Fisher Information Matrix	100
5.3	Kullback–Leibler Divergence	103
5.4	Derivation of Akaike’s Information Criterion	105
	Appendix: Proofs of Propositions	107
	Exercises 40–48	110
6	Regularization	115
6.1	Ridge	115
6.2	Subderivative	117
6.3	Lasso	121
6.4	Comparing Ridge and Lasso	124
6.5	Setting the λ Value	126
	Exercises 49–56	127
7	Nonlinear Regression	133
7.1	Polynomial Regression	133
7.2	Spline Regression	136
7.3	Natural Spline Regression	140
7.4	Smoothing Spline	144
7.5	Local Regression	148
7.6	Generalized Additive Models	153
	Appendix: Proofs of Propositions	155
	Exercises 57–68	161
8	Decision Trees	171
8.1	Decision Trees for Regression	171
8.2	Decision Tree for Classification	180
8.3	Bagging	185
8.4	Random Forest	187
8.5	Boosting	189
	Exercises 69–74	193
9	Support Vector Machine	199
9.1	Optimum Boarder	199
9.2	Theory of Optimization	202
9.3	The Solution of Support Vector Machines	205
9.4	Extension of Support Vector Machines Using a Kernel	209
	Appendix: Proofs of Propositions	216
	Exercises 75–87	217

10 Unsupervised Learning	227
10.1 <i>K</i> -means Clustering	227
10.2 Hierarchical Clustering	232
10.3 Principle Component Analysis	239
Appendix: Program	247
Exercises 88–100	248
Index	255