



C o m m u n i t y E x p e r i e n c e D i s t i l l e d

Learning Bayesian Models with R

Become an expert in Bayesian machine learning methods using R and apply them to solve real-world Big Data problems

Dr. Hari M. Koduvely

[PACKT] open source*
PUBLISHING community experience distilled

Learning Bayesian Models with R

Become an expert in Bayesian machine learning
methods using R and apply them to solve real-world
Big Data problems

Dr. Hari M. Koduvely



BIRMINGHAM - MUMBAI

Learning Bayesian Models with R

Copyright © 2015 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the author, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: October 2015

Production reference: 1231015

Published by Packt Publishing Ltd.
Livery Place
35 Livery Street
Birmingham B3 2PB, UK.

ISBN 978-1-78398-760-3

www.packtpub.com

Credits

Author

Dr. Hari M. Koduvely

Project Coordinator

Bijal Patel

Reviewers

Philip B. Graff

Nishanth Upadhyaya

Proofreader

Safis Editing

Commissioning Editor

Kartikey Pandey

Indexer

Hemangini Bari

Acquisition Editor

Nikhil Karkal

Graphics

Abhinash Sahu

Content Development Editor

Athira Laji

Production Coordinator

Nitesh Thakur

Technical Editor

Taabish Khan

Cover Work

Nitesh Thakur

Copy Editor

Trishya Hajare

About the Author

Dr. Hari M. Koduvely is an experienced data scientist working at the Samsung R&D Institute in Bangalore, India. He has a PhD in statistical physics from the Tata Institute of Fundamental Research, Mumbai, India, and post-doctoral experience from the Weizmann Institute, Israel, and Georgia Tech, USA. Prior to joining Samsung, the author has worked for Amazon and Infosys Technologies, developing machine learning-based applications for their products and platforms. He also has several publications on Bayesian inference and its applications in areas such as recommendation systems and predictive health monitoring. His current interest is in developing large-scale machine learning methods, particularly for natural language understanding.

I would like to express my gratitude to all those who have helped me throughout my career, without whom this book would not have been possible. This includes my teachers, mentors, friends, colleagues, and all the institutions in which I worked, especially my current employer, Samsung R&D Institute, Bangalore. A special mention to my spouse, Prathyusha, and son, Pranav, for their immense moral support during the writing of the book.

About the Reviewers

Philip B. Graff is a data scientist with the Johns Hopkins University Applied Physics Laboratory. He works with graph analytics for a large-scale automated pattern discovery.

Philip obtained his PhD in physics from the University of Cambridge on a Gates Cambridge Scholarship, and a BS in physics and mathematics from the University of Maryland, Baltimore County. His PhD thesis implemented Bayesian methods for gravitational wave detection and the training of neural networks for machine learning.

Philip's post-doctoral research at NASA Goddard Space Flight Center and the University of Maryland, College Park, applied Bayesian inference to the detection and measurement of gravitational waves by ground and space-based detectors, LIGO and LISA, respectively. He also implemented machine learning methods for improved gamma-ray burst data analysis. He has published books in the fields of astrophysical data analysis and machine learning.

I would like to thank Ala for her support while I reviewed this book.

Nishanth Upadhyaya has close to 10 years of experience in the area of analytics, Monte Carlo methods, signal processing, machine learning, and building end-to-end data products. He is active on StackOverflow and GitHub. He has a couple of patents in the area of item response theory and stochastic optimization. He has also won third place in the first ever Aadhaar hackathon organized by Khosla labs.

www.PacktPub.com

Support files, eBooks, discount offers, and more

For support files and downloads related to your book, please visit www.PacktPub.com.

Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details.

At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<https://www2.packtpub.com/books/subscription/packtlib>

Do you need instant solutions to your IT questions? PacktLib is Packt's online digital book library. Here, you can search, access, and read Packt's entire library of books.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Free access for Packt account holders

If you have an account with Packt at www.PacktPub.com, you can use this to access PacktLib today and view 9 entirely free books. Simply use your login credentials for immediate access.

Table of Contents

Preface	v
Chapter 1: Introducing the Probability Theory	1
Probability distributions	2
Conditional probability	6
Bayesian theorem	7
Marginal distribution	8
Expectations and covariance	9
Binomial distribution	9
Beta distribution	10
Gamma distribution	11
Dirichlet distribution	12
Wishart distribution	12
Exercises	13
References	14
Summary	15
Chapter 2: The R Environment	17
Setting up the R environment and packages	18
Installing R and RStudio	18
Your first R program	19
Managing data in R	19
Data Types in R	19
Data structures in R	20
Importing data into R	22
Slicing and dicing datasets	23
Vectorized operations	24

Writing R programs	25
Control structures	25
Functions	25
Scoping rules	26
Loop functions	27
lapply	28
sapply	28
mapply	29
apply	29
tapply	30
Data visualization	30
High-level plotting functions	31
Low-level plotting commands	32
Interactive graphics functions	33
Sampling	33
Random uniform sampling from an interval	34
Sampling from normal distribution	34
Exercises	35
References	35
Summary	36
Chapter 3: Introducing Bayesian Inference	37
Bayesian view of uncertainty	37
Choosing the right prior distribution	42
Non-informative priors	42
Subjective priors	44
Conjugate priors	46
Hierarchical priors	47
Estimation of posterior distribution	48
Maximum a posteriori estimation	48
Laplace approximation	49
Monte Carlo simulations	51
Variational approximation	57
Prediction of future observations	59
Exercises	59
References	60
Summary	60
Chapter 4: Machine Learning Using Bayesian Inference	61
Why Bayesian inference for machine learning?	63
Model overfitting and bias-variance tradeoff	65
Selecting models of optimum complexity	66
Subset selection	66
Model regularization	67

Bayesian averaging	68
An overview of common machine learning tasks	70
References	72
Summary	72
Chapter 5: Bayesian Regression Models	73
Generalized linear regression	73
The arm package	74
The Energy efficiency dataset	74
Regression of energy efficiency with building parameters	75
Ordinary regression	77
Bayesian regression	77
Simulation of the posterior distribution	79
Exercises	81
References	81
Summary	81
Chapter 6: Bayesian Classification Models	83
Performance metrics for classification	84
The Naïve Bayes classifier	85
Text processing using the tm package	87
Model training and prediction	88
The Bayesian logistic regression model	91
The BayesLogit R package	93
The dataset	93
Preparation of the training and testing datasets	94
Using the Bayesian logistic model	95
Exercises	96
References	96
Summary	97
Chapter 7: Bayesian Models for Unsupervised Learning	99
Bayesian mixture models	100
The bgmm package for Bayesian mixture models	103
Topic modeling using Bayesian inference	105
Latent Dirichlet allocation	106
R packages for LDA	107
The topicmodels package	108
The lda package	109
Exercises	110
References	110
Summary	111

Chapter 8: Bayesian Neural Networks	113
Two-layer neural networks	114
Bayesian treatment of neural networks	116
The brnn R package	118
Deep belief networks and deep learning	119
Restricted Boltzmann machines	120
Deep belief networks	123
The darch R package	124
Other deep learning packages in R	126
Exercises	127
References	127
Summary	128
Chapter 9: Bayesian Modeling at Big Data Scale	129
Distributed computing using Hadoop	130
RHadoop for using Hadoop from R	130
Spark – in-memory distributed computing	132
SparkR	133
Linear regression using SparkR	133
Computing clusters on the cloud	134
Amazon Web Services	134
Creating and running computing instances on AWS	134
Installing R and RStudio	135
Running Spark on EC2	136
Microsoft Azure	137
IBM Bluemix	137
Other R packages for large scale machine learning	137
The parallel R package	138
The foreach R package	138
Exercises	139
References	140
Summary	141
Index	143

Preface

Bayesian inference provides a unified framework to deal with all sorts of uncertainties when learning patterns from data using machine learning models and using it for predicting future observations. However, learning and implementing Bayesian models is not easy for data science practitioners due to the level of mathematical treatment involved. Also, applying Bayesian methods to real-world problems requires high computational resources. With the recent advancements in cloud and high-performance computing and easy access to computational resources, Bayesian modeling has become more feasible to use for practical applications today. Therefore, it would be advantageous for all data scientists and data engineers to understand Bayesian methods and apply them in their projects to achieve better results.

What this book covers

This book gives comprehensive coverage of the Bayesian machine learning models and the R packages that implement them. It begins with an introduction to the fundamentals of probability theory and R programming for those who are new to the subject. Then, the book covers some of the most important machine learning methods, both supervised learning and unsupervised learning, implemented using Bayesian inference and R. Every chapter begins with a theoretical description of the method, explained in a very simple manner. Then, relevant R packages are discussed and some illustrations using datasets from the UCI machine learning repository are given. Each chapter ends with some simple exercises for you to get hands-on experience of the concepts and R packages discussed in the chapter. The state-of-the-art topics covered in the chapters are Bayesian regression using linear and generalized linear models, Bayesian classification using logistic regression, classification of text data using Naïve Bayes models, and Bayesian mixture models and topic modeling using Latent Dirichlet allocation.

The last two chapters are devoted to the latest developments in the field. One chapter discusses deep learning, which uses a class of neural network models that are currently at the frontier of artificial intelligence. The book concludes with the application of Bayesian methods on Big Data using frameworks such as Hadoop and Spark.

Chapter 1, Introducing the Probability Theory, covers the foundational concepts of probability theory, particularly those aspects required for learning Bayesian inference, which are presented to you in a simple and coherent manner.

Chapter 2, The R Environment, introduces you to the R environment. After reading through this chapter, you will learn how to import data into R, make a selection of subsets of data for its analysis, and write simple R programs using functions and control structures. Also, you will get familiar with the graphical capabilities of R and some advanced capabilities such as loop functions.

Chapter 3, Introducing Bayesian Inference, introduces you to the Bayesian statistic framework. This chapter includes a description of the Bayesian theorem, concepts such as prior and posterior probabilities, and different methods to estimate posterior distribution such as MAP estimates, Monte Carlo simulations, and variational estimates.

Chapter 4, Machine Learning Using Bayesian Inference, gives an overview of what machine learning is and what some of its high-level tasks are. This chapter also discusses the importance of Bayesian inference in machine learning, particularly in the context of how it can help to avoid important issues such as model overfit and how to select optimum models.

Chapter 5, Bayesian Regression Models, presents one of the most common supervised machine learning tasks, namely, regression modeling, in the Bayesian framework. It shows by using an example how you can get tighter confidence intervals of prediction using Bayesian regression models.

Chapter 6, Bayesian Classification Models, presents how to use the Bayesian framework for another common machine learning task, classification. The two Bayesian models of classification, Naïve Bayes and Bayesian logistic regression, are discussed along with some important metrics for evaluating the performance of classifiers.

Chapter 7, Bayesian Models for Unsupervised Learning, introduces you to the concepts behind unsupervised and semi-supervised machine learning and their Bayesian treatment. The two most important Bayesian unsupervised models, the Bayesian mixture model and LDA, are discussed.

Chapter 8, Bayesian Neural Networks, presents an important class of machine learning model, namely neural networks, and their Bayesian implementation. Neural network models are inspired by the architecture of the human brain and they continue to be an area of active research and development. The chapter also discusses deep learning, one of the latest advances in neural networks, which is used to solve many problems in computer vision and natural language processing with remarkable accuracy.

Chapter 9, Bayesian Modeling at Big Data Scale, covers various frameworks for performing large-scale Bayesian machine learning such as Hadoop, Spark, and parallelization frameworks that are native to R. The chapter also discusses how to set up instances on cloud services, such as Amazon Web Services and Microsoft Azure, and run R programs on them.

What you need for this book

To learn the examples and try the exercises presented in this book, you need to install the latest version of the R programming environment and the RStudio IDE. Apart from this, you need to install the specific R packages that are mentioned in each chapter of this book separately.

Who this book is for

This book is intended for data scientists who analyze large datasets to generate insights and for data engineers who develop platforms, solutions, or applications based on machine learning. Although many data science practitioners are quite familiar with machine learning techniques and R, they may not know about Bayesian inference and its merits. This book, therefore, would be helpful to even experienced data scientists and data engineers to learn about Bayesian methods and incorporate them in to their projects to get better results. No prior experience is required in R or probability theory to use this book.

Conventions

In this book, you will find a number of text styles that distinguish between different kinds of information. Here are some examples of these styles and an explanation of their meaning.

Code words in text, database table names, folder names, filenames, file extensions, pathnames, dummy URLs, user input, and Twitter handles are shown as follows:
"The first function is `gibbs_met`."


A block of code is set as follows:


```
myMean <-function(x) {  
  s <-sum(x)  
  l <-length(x)  
  mean <-s/l  
  mean  
}  
>x <-c(10,20,30,40,50)  
>myMean(x)  
[1] 30
```

Any command-line input or output is written as follows:

```
setwd("directory path")
```

New terms and **important words** are shown in bold. Words that you see on the screen, for example, in menus or dialog boxes, appear in the text like this: "You can also set this from the menu bar of RStudio by clicking on **Session | Set Working Directory**."

[ Warnings or important notes appear in a box like this.]

[ Tips and tricks appear like this.]

Reader feedback

Feedback from our readers is always welcome. Let us know what you think about this book—what you liked or disliked. Reader feedback is important for us as it helps us develop titles that you will really get the most out of.

To send us general feedback, simply e-mail feedback@packtpub.com, and mention the book's title in the subject of your message.

If there is a topic that you have expertise in and you are interested in either writing or contributing to a book, see our author guide at www.packtpub.com/authors.

Customer support

Now that you are the proud owner of a Packt book, we have a number of things to help you to get the most from your purchase.

Downloading the example code

You can download the example code files from your account at <http://www.packtpub.com> for all the Packt Publishing books you have purchased. If you purchased this book elsewhere, you can visit <http://www.packtpub.com/support> and register to have the files e-mailed directly to you.

Errata

Although we have taken every care to ensure the accuracy of our content, mistakes do happen. If you find a mistake in one of our books – maybe a mistake in the text or the code – we would be grateful if you could report this to us. By doing so, you can save other readers from frustration and help us improve subsequent versions of this book. If you find any errata, please report them by visiting <http://www.packtpub.com/submit-errata>, selecting your book, clicking on the **Errata Submission Form** link, and entering the details of your errata. Once your errata are verified, your submission will be accepted and the errata will be uploaded to our website or added to any list of existing errata under the Errata section of that title.

To view the previously submitted errata, go to <https://www.packtpub.com/books/content/support> and enter the name of the book in the search field. The required information will appear under the **Errata** section.

Piracy

Piracy of copyrighted material on the Internet is an ongoing problem across all media. At Packt, we take the protection of our copyright and licenses very seriously. If you come across any illegal copies of our works in any form on the Internet, please provide us with the location address or website name immediately so that we can pursue a remedy.

Please contact us at copyright@packtpub.com with a link to the suspected pirated material.

We appreciate your help in protecting our authors and our ability to bring you valuable content.

Questions

If you have a problem with any aspect of this book, you can contact us at questions@packtpub.com, and we will do our best to address the problem.