


R语言机器学习 参考手册 (影印版)

Machine Learning with R Cookbook

Yu-Wei Chiu (David Chiu) 著

[PACKT]
PUBLISHING

使用简单且易于使用的R代码，
通过110多个参考方案分析数据并构建可预测模型

 东南大学出版社
SOUTHEAST UNIVERSITY PRESS

R 语言机器学习参考手册(影印版)

Yu - Wei Chiu (David Chiu) 著

南京 东南大学出版社

图书在版编目(CIP)数据

R 语言机器学习参考手册:英文/丘祐玮(Yu - wei, Chiu)著. —影印本. —南京:东南大学出版社,2016.1

书名原文:Machine Learning with R Cookbook

ISBN 978 - 7 - 5641 - 6063 - 0

I. ①R… II. ①丘… III. ①程序语言—程序设计—技术手册—英文 IV. ①TP312 - 62

中国版本图书馆 CIP 数据核字(2015)第 243405 号

© 2015 by PACKT Publishing Ltd

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2016. Authorized reprint of the original English edition, 2015 PACKT Publishing Ltd, the owner of all rights to publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2015。

英文影印版由东南大学出版社出版 2016。此影印版的出版和销售得到出版权和销售权的所有者——PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

R 语言机器学习参考手册(影印版)

出版发行:东南大学出版社

地 址:南京四牌楼 2 号 邮编:210096

出 版 人:江建中

网 址:<http://www.seupress.com>

电子邮件:press@seupress.com

印 刷:常州市武进第三印刷有限公司

开 本:787 毫米×980 毫米 16 开本

印 张:27.5

字 数:539 千字

版 次:2016 年 1 月第 1 版

印 次:2016 年 1 月第 1 次印刷

书 号:ISBN 978 - 7 - 5641 - 6063 - 0

定 价:78.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

About the Author

Yu-Wei, Chiu (David Chiu) is the founder of LargitData (www.LargitData.com). He has previously worked for Trend Micro as a software engineer, with the responsibility of building big data platforms for business intelligence and customer relationship management systems. In addition to being a start-up entrepreneur and data scientist, he specializes in using Spark and Hadoop to process big data and apply data mining techniques for data analysis. Yu-Wei is also a professional lecturer and has delivered lectures on Python, R, Hadoop, and tech talks at a variety of conferences.

In 2013, Yu-Wei reviewed *Bioinformatics with R Cookbook*, Packt Publishing. For more information, please visit his personal website at www.ywchiu.com.

I have immense gratitude for my family and friends for supporting and encouraging me to complete this book. I would like to sincerely thank my mother, Ming-Yang Huang (Miranda Huang); my mentor, Man-Kwan Shan; the proofreader of this book, Brendan Fisher; Taiwan R User Group; Data Science Program (DSP); and other friends who have offered their support.

Credits

Author

Yu-Wei, Chiu (David Chiu)

Project Coordinator

Nikhil Nair

Reviewers

Tarek Amr

Abir Datta (data scientist)

Saibal Dutta

Ratanlal Mahanta
(senior quantitative analyst)

Ricky Shi

Jithin S.L

Proofreaders

Simran Bhogal

Joanna McMahon

Jonathan Todd

Indexer

Mariammal Chettiyar

Commissioning Editor

Akram Hussain

Graphics

Sheetal Aute

Abhinash Sahu

Acquisition Editor

James Jones

Production Coordinator

Melwyn D'sa

Content Development Editor

Arvind Koul

Cover Work

Melwyn D'sa

Technical Editors

Tanvi Bhatt

Shashank Desai

Copy Editor

Sonia Cheema

About the Reviewers

Tarek Amr currently works as a data scientist at bidx in the Netherlands. He has an MSc degree from the University of East Anglia in knowledge discovery and data mining. He also volunteers at the Open Knowledge Foundation and School of Data, where he works on projects related to open data and gives training in the field of data journalism and data visualization. He has reviewed another book, *Python Data Visualization Cookbook*, Packt Publishing, and is currently working on writing a new book on data visualization using D3.js.

You can find out more about him at <http://tarekamr.appspot.com/>.

Abir Datta (data scientist) has been working as a data scientist in Cognizant Technology Solutions Ltd. in the fields of insurance, financial services, and digital analytics verticals. He has mainly been working in the fields of analytics, predictive modeling, and business intelligence/analysis in designing and developing end-to-end big data integrated analytical solutions for different verticals to cater to a client's analytical business problems. He has also developed algorithms to identify the latent characteristics of customers so as to take channelized strategic decisions for much more effective business success.

Abir is also involved in risk modeling and has been a part of the team that developed a model risk governance platform for his current organization, which has been widely recognized across the banking and financial service industry.

Saibal Dutta is presently researching in the field of data mining and machine learning at the Indian Institute of Technology, Kharagpur, India. He also holds a master's degree in electronics and communication from the National Institute of Technology, Rourkela, India. He has also worked at HCL Technologies Limited, Noida, as a software consultant. In his 4 years of consulting experience, he has been associated with global players such as IKEA (in Sweden), Pearson (in the U.S.), and so on. His passion for entrepreneurship has led him to start his own start-up in the field of data analytics, which is in the bootstrapping stage. His areas of expertise include data mining, machine learning, image processing, and business consultation.

Ratanlal Mahanta (senior quantitative analyst) holds an MSc in computational finance and is currently working at the GPSK Investment Group as a senior quantitative analyst. He has 4 years of experience in quantitative trading and strategy developments for sell-side and risk consultation firms. He is an expert in high frequency and algorithmic trading.

He has expertise in the following areas:

- ▶ Quantitative trading: FX, equities, futures, options, and engineering on derivatives
- ▶ Algorithms: Partial differential equations, Stochastic Differential Equations, Finite Difference Method, Monte-Carlo, and Machine Learning
- ▶ Code: R Programming, C++, MATLAB, HPC, and Scientific Computing
- ▶ Data analysis: Big-Data-Analytic [EOD to TBT], Bloomberg, Quandl, and Quantopian
- ▶ Strategies: Vol-Arbitrage, Vanilla and Exotic Options Modeling, trend following, Mean reversion, Co-integration, Monte-Carlo Simulations, ValueatRisk, Stress Testing, Buy side trading strategies with high Sharpe ratio, Credit Risk Modeling, and Credit Rating

He has already reviewed two books for Packt Publishing: *Mastering Scientific Computing with R* and *Mastering Quantitative Finance with R*.

Currently, he is reviewing a book for Packt Publishing: *Mastering Python for Data Science*.

Ricky Shi is currently a quantitative trader and researcher, focusing on large-scale machine learning and robust prediction techniques. He obtained a PhD in the field of machine learning and data mining with big data. Concurrently, he conducts research in applied math. With the objective to apply academic research to real-world practice, he has worked with several research institutes and companies, including Yahoo! labs, AT&T Labs, Eagle Seven, Morgan Stanley Equity Trading Lab (ETL), and Engineers Gate Manager LP, supervised by Professor Philip S. Yu.

His research interest covers the following topics:

- ▶ Correlation among heterogeneous data, such as social advertising from both the users' demographic features and users' social networks
- ▶ Correlation among evolving time series objects, such as finding dynamic correlations, finding the most influential financial products (shaker detection, cascading graph), and using the correlation in hedging and portfolio management
- ▶ Correlation among learning tasks, such as transfer learning

Jithin S.L completed his BTech in information technology from Loyola Institute of Technology and Science. He started his career in the field of analytics and then moved to various verticals of big data technology. He has worked with reputed organizations, such as Thomson Reuters, IBM, and Flytxt, under different roles. He has worked in the banking, energy, healthcare, and telecom domains and has handled global projects on big data technology.

He has submitted many research papers on technology and business at national and international conferences.

His motto in life is that learning is always a neverending process that helps in understanding, modeling, and presenting new concepts to the modern world.

I surrender myself to God almighty who helped me to review this book in an effective way. I dedicate my work on this book to my dad, Mr. N. Subbian Asari, my lovable mom, Mrs. M. Lekshmi, and my sweet sister, Ms. S.L Jishma, for coordinating and encouraging me to write this book.

Last but not least, I would like to thank all my friends.

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	vii
Chapter 1: Practical Machine Learning with R	13
Introduction	13
Downloading and installing R	15
Downloading and installing RStudio	23
Installing and loading packages	27
Reading and writing data	29
Using R to manipulate data	32
Applying basic statistics	36
Visualizing data	40
Getting a dataset for machine learning	44
Chapter 2: Data Exploration with RMS Titanic	49
Introduction	49
Reading a Titanic dataset from a CSV file	51
Converting types on character variables	54
Detecting missing values	56
Imputing missing values	59
Exploring and visualizing data	62
Predicting passenger survival with a decision tree	70
Validating the power of prediction with a confusion matrix	75
Assessing performance with the ROC curve	77
Chapter 3: R and Statistics	79
Introduction	79
Understanding data sampling in R	80
Operating a probability distribution in R	81
Working with univariate descriptive statistics in R	86
Performing correlations and multivariate analysis	90
Operating linear regression and multivariate analysis	92

Conducting an exact binomial test	95
Performing student's t-test	97
Performing the Kolmogorov-Smirnov test	101
Understanding the Wilcoxon Rank Sum and Signed Rank test	104
Working with Pearson's Chi-squared test	105
Conducting a one-way ANOVA	109
Performing a two-way ANOVA	112
Chapter 4: Understanding Regression Analysis	117
Introduction	117
Fitting a linear regression model with lm	118
Summarizing linear model fits	120
Using linear regression to predict unknown values	123
Generating a diagnostic plot of a fitted model	124
Fitting a polynomial regression model with lm	127
Fitting a robust linear regression model with rlm	129
Studying a case of linear regression on SLID data	131
Applying the Gaussian model for generalized linear regression	138
Applying the Poisson model for generalized linear regression	141
Applying the Binomial model for generalized linear regression	142
Fitting a generalized additive model to data	144
Visualizing a generalized additive model	146
Diagnosing a generalized additive model	149
Chapter 5: Classification (I) – Tree, Lazy, and Probabilistic	153
Introduction	153
Preparing the training and testing datasets	154
Building a classification model with recursive partitioning trees	156
Visualizing a recursive partitioning tree	159
Measuring the prediction performance of a recursive partitioning tree	161
Pruning a recursive partitioning tree	163
Building a classification model with a conditional inference tree	166
Visualizing a conditional inference tree	167
Measuring the prediction performance of a conditional inference tree	170
Classifying data with the k-nearest neighbor classifier	172
Classifying data with logistic regression	175
Classifying data with the Naïve Bayes classifier	182
Chapter 6: Classification (II) – Neural Network and SVM	187
Introduction	187
Classifying data with a support vector machine	188
Choosing the cost of a support vector machine	191
Visualizing an SVM fit	195

Predicting labels based on a model trained by a support vector machine	197
Tuning a support vector machine	201
Training a neural network with neuralnet	205
Visualizing a neural network trained by neuralnet	209
Predicting labels based on a model trained by neuralnet	211
Training a neural network with nnet	214
Predicting labels based on a model trained by nnet	216
Chapter 7: Model Evaluation	219
Introduction	219
Estimating model performance with k-fold cross-validation	220
Performing cross-validation with the e1071 package	222
Performing cross-validation with the caret package	223
Ranking the variable importance with the caret package	225
Ranking the variable importance with the rminer package	227
Finding highly correlated features with the caret package	229
Selecting features using the caret package	230
Measuring the performance of the regression model	236
Measuring prediction performance with a confusion matrix	239
Measuring prediction performance using ROCR	241
Comparing an ROC curve using the caret package	243
Measuring performance differences between models with the caret package	246
Chapter 8: Ensemble Learning	251
Introduction	251
Classifying data with the bagging method	252
Performing cross-validation with the bagging method	256
Classifying data with the boosting method	257
Performing cross-validation with the boosting method	261
Classifying data with gradient boosting	262
Calculating the margins of a classifier	268
Calculating the error evolution of the ensemble method	272
Classifying data with random forest	274
Estimating the prediction errors of different classifiers	280
Chapter 9: Clustering	283
Introduction	283
Clustering data with hierarchical clustering	284
Cutting trees into clusters	290
Clustering data with the k-means method	294
Drawing a bivariate cluster plot	297
Comparing clustering methods	299

Extracting silhouette information from clustering	302
Obtaining the optimum number of clusters for k-means	303
Clustering data with the density-based method	306
Clustering data with the model-based method	309
Visualizing a dissimilarity matrix	314
Validating clusters externally	317
Chapter 10: Association Analysis and Sequence Mining	321
<hr/>	
Introduction	321
Transforming data into transactions	322
Displaying transactions and associations	324
Mining associations with the Apriori rule	328
Pruning redundant rules	333
Visualizing association rules	335
Mining frequent itemsets with Eclat	339
Creating transactions with temporal information	342
Mining frequent sequential patterns with cSPADE	345
Chapter 11: Dimension Reduction	349
<hr/>	
Introduction	349
Performing feature selection with FSelector	351
Performing dimension reduction with PCA	354
Determining the number of principal components using the scree test	359
Determining the number of principal components using the Kaiser method	361
Visualizing multivariate data using biplot	363
Performing dimension reduction with MDS	367
Reducing dimensions with SVD	371
Compressing images with SVD	375
Performing nonlinear dimension reduction with ISOMAP	378
Performing nonlinear dimension reduction with Local Linear Embedding	383
Chapter 12: Big Data Analysis (R and Hadoop)	387
<hr/>	
Introduction	387
Preparing the RHadoop environment	389
Installing rmr2	392
Installing rhdfs	393
Operating HDFS with rhdfs	395
Implementing a word count problem with RHadoop	397
Comparing the performance between an R MapReduce program and a standard R program	399
Testing and debugging the rmr2 program	401
Installing plyrmr	403

Manipulating data with plymr	404
Conducting machine learning with RHadoop	407
Configuring RHadoop clusters on Amazon EMR	411
Appendix A: Resources for R and Machine Learning	419
Appendix B: Dataset – Survival of Passengers on the Titanic	421
Index	423

Table of Contents

Preface	vii
Chapter 1: Practical Machine Learning with R	13
Introduction	13
Downloading and installing R	15
Downloading and installing RStudio	23
Installing and loading packages	27
Reading and writing data	29
Using R to manipulate data	32
Applying basic statistics	36
Visualizing data	40
Getting a dataset for machine learning	44
Chapter 2: Data Exploration with RMS Titanic	49
Introduction	49
Reading a Titanic dataset from a CSV file	51
Converting types on character variables	54
Detecting missing values	56
Imputing missing values	59
Exploring and visualizing data	62
Predicting passenger survival with a decision tree	70
Validating the power of prediction with a confusion matrix	75
Assessing performance with the ROC curve	77
Chapter 3: R and Statistics	79
Introduction	79
Understanding data sampling in R	80
Operating a probability distribution in R	81
Working with univariate descriptive statistics in R	86
Performing correlations and multivariate analysis	90
Operating linear regression and multivariate analysis	92

Conducting an exact binomial test	95
Performing student's t-test	97
Performing the Kolmogorov-Smirnov test	101
Understanding the Wilcoxon Rank Sum and Signed Rank test	104
Working with Pearson's Chi-squared test	105
Conducting a one-way ANOVA	109
Performing a two-way ANOVA	112
Chapter 4: Understanding Regression Analysis	117
Introduction	117
Fitting a linear regression model with lm	118
Summarizing linear model fits	120
Using linear regression to predict unknown values	123
Generating a diagnostic plot of a fitted model	124
Fitting a polynomial regression model with lm	127
Fitting a robust linear regression model with rlm	129
Studying a case of linear regression on SLID data	131
Applying the Gaussian model for generalized linear regression	138
Applying the Poisson model for generalized linear regression	141
Applying the Binomial model for generalized linear regression	142
Fitting a generalized additive model to data	144
Visualizing a generalized additive model	146
Diagnosing a generalized additive model	149
Chapter 5: Classification (I) – Tree, Lazy, and Probabilistic	153
Introduction	153
Preparing the training and testing datasets	154
Building a classification model with recursive partitioning trees	156
Visualizing a recursive partitioning tree	159
Measuring the prediction performance of a recursive partitioning tree	161
Pruning a recursive partitioning tree	163
Building a classification model with a conditional inference tree	166
Visualizing a conditional inference tree	167
Measuring the prediction performance of a conditional inference tree	170
Classifying data with the k-nearest neighbor classifier	172
Classifying data with logistic regression	175
Classifying data with the Naïve Bayes classifier	182
Chapter 6: Classification (II) – Neural Network and SVM	187
Introduction	187
Classifying data with a support vector machine	188
Choosing the cost of a support vector machine	191
Visualizing an SVM fit	195

Predicting labels based on a model trained by a support vector machine	197
Tuning a support vector machine	201
Training a neural network with neuralnet	205
Visualizing a neural network trained by neuralnet	209
Predicting labels based on a model trained by neuralnet	211
Training a neural network with nnet	214
Predicting labels based on a model trained by nnet	216
Chapter 7: Model Evaluation	219
Introduction	219
Estimating model performance with k-fold cross-validation	220
Performing cross-validation with the e1071 package	222
Performing cross-validation with the caret package	223
Ranking the variable importance with the caret package	225
Ranking the variable importance with the rminer package	227
Finding highly correlated features with the caret package	229
Selecting features using the caret package	230
Measuring the performance of the regression model	236
Measuring prediction performance with a confusion matrix	239
Measuring prediction performance using ROCR	241
Comparing an ROC curve using the caret package	243
Measuring performance differences between models with the caret package	246
Chapter 8: Ensemble Learning	251
Introduction	251
Classifying data with the bagging method	252
Performing cross-validation with the bagging method	256
Classifying data with the boosting method	257
Performing cross-validation with the boosting method	261
Classifying data with gradient boosting	262
Calculating the margins of a classifier	268
Calculating the error evolution of the ensemble method	272
Classifying data with random forest	274
Estimating the prediction errors of different classifiers	280
Chapter 9: Clustering	283
Introduction	283
Clustering data with hierarchical clustering	284
Cutting trees into clusters	290
Clustering data with the k-means method	294
Drawing a bivariate cluster plot	297
Comparing clustering methods	299