Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

A First Course in Machine Learning

Simon Rogers
Mark Girolami

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

A First Course in Machine Learning

常州大学图书馆
藏书章

Simon Rogers
Mark Girolami

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

MATLAB® is a trademark of The MathWorks, Inc. and is used with permission. The MathWorks does not warrant the accuracy of the text or exercises in this book. This book's use or discussion of MATLAB® software or related products does not constitute endorsement or sponsorship by The MathWorks of a particular pedagogical approach or particular use of the MATLAB® software.

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

© 2012 by Taylor & Francis Group, LLC
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works

Printed in the United States of America on acid-free paper
Version Date: 2011908

International Standard Book Number: 978-1-4398-2414-6 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access www.copyright.com (http://www.copyright.com/) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

Trademark Notice: Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

A First Course in Machine Learning

Chapman & Hall/CRC
Machine Learning & Pattern Recognition Series

SERIES EDITORS

Ralf Herbrich and Thore Graepel
Microsoft Research Ltd.
Cambridge, UK

AIMS AND SCOPE

This series reflects the latest advances and applications in machine learning and pattern recognition through the publication of a broad range of reference works, textbooks, and handbooks. The inclusion of concrete examples, applications, and methods is highly encouraged. The scope of the series includes, but is not limited to, titles in the areas of machine learning, pattern recognition, computational intelligence, robotics, computational/statistical learning theory, natural language processing, computer vision, game AI, game theory, neural networks, computational neuroscience, and other relevant topics, such as machine learning applied to bioinformatics or cognitive science, which might be proposed by potential contributors.

PUBLISHED TITLES

MACHINE LEARNING: An Algorithmic Perspective
Stephen Marsland

HANDBOOK OF NATURAL LANGUAGE PROCESSING,
Second Edition
Nitin Indurkhya and Fred J. Damerau

UTILITY-BASED LEARNING FROM DATA
Craig Friedman and Sven Sandow

A FIRST COURSE IN MACHINE LEARNING
Simon Rogers and Mark Girolami

试读结束，需要全本PDF请购买 www.ertongbook.com

Preface

Machine learning is rapidly becoming one of the most important areas of general practice, research and development activity within computing science. This is reflected in the scale of the academic research area devoted to the subject and the active recruitment of machine learning specialists by major international banks and financial institutions as well as companies such as Microsoft®, Google®, Yahoo® and Amazon®.

This growth can be partly explained by the increase in the quantity and diversity of measurements we are able to make of the world. A particularly fascinating example arises from the wave of new biological measurement technologies that preceded the sequencing of the first genomes. It is now possible to measure the detailed molecular state of an organism in ways that would have been hard to imagine only a short time ago. Such measurements go far beyond our understanding of these organisms and machine learning techniques have been heavily involved in the distillation of useful structures from them.

This book is based on material presented in a machine learning course in the School of Computing Science at the University of Glasgow, UK. The course, presented to final year undergraduates and taught by postgraduates, is made up of 20 hour-long lectures and 10 hour-long laboratory sessions. In such a short teaching period, it is impossible to cover more than a small fraction of the material that now comes under the banner of machine learning. Our intention when teaching this course, therefore, is to present the core mathematical and statistical techniques required to understand some of the most popular machine learning algorithms and then present a few of these algorithms that span the main problem areas within machine learning: classification, clustering and projection. At the end of the course, the students should have the knowledge and confidence to be able to explore machine learning literature to find methods that are more appropriate for them. The same is hopefully true of readers of this book.

Due to the varying mathematical literacy of students taking the course, we assume only very minor mathematical pre-requisites. An undergraduate student from computer science, engineering, physics (or any other numerical subject) should have no problem. This does not preclude those without such experience – additional mathematical explanations appear throughout the text in comment boxes. In addition, important equations have been highlighted – it is worth spending time understanding these equations before proceeding.

Students attending this course often find the practical sessions very useful. Experimenting with the various algorithms and concepts helps transfer them from an abstract set of equations into something that could be used to solve real problems. We have attempted to transfer this to the book through an extensive collection of MATLAB®/Octave[1] scripts, available from the associated web page and referenced throughout the text. These scripts enable the user to recreate plots that appear in the book and investigate changing model specifications and parameter values.

Finally, the machine learning methods that are covered in this book are our choice of those we feel students should understand. In limited space and time, we think that it is more worthwhile to give detailed descriptions and derivations for a small number of algorithms than attempt to cover many algorithms at a lower level of detail – many people will not find their favourite algorithms within this book!

MATLAB® is a registered trademark of The MathWorks, Inc.
For product information, please contact:

The MathWorks, Inc.
3 Apple Hill Drive
Natick MA 01760-2098 USA
Tel: 508-647-7000
Fax: 508-647-7001
E-mail: info@mathworks.com
Web: www.mathworks.com

Simon Rogers and Mark Girolami

[1] A free mathematical software environment, available from `www.gnu.org/software/octave/`

Contents

List of Tables **xi**

List of Figures **xiii**

Preface **xix**

1 Linear Modelling: A Least Squares Approach **1**

1.1 Linear modelling 1

1.1.1 Defining the model 2

1.1.2 Modelling assumptions 3

1.1.3 Defining what a *good* model is 4

1.1.4 The least squares solution – a worked example 6

1.1.5 Worked example 9

1.1.6 Least squares fit to the Olympics data 10

1.1.7 Summary 11

1.2 Making predictions 12

1.2.1 A second Olympics dataset 12

1.2.2 Summary 15

1.3 Vector/matrix notation 15

1.3.1 Example 22

1.3.2 Numerical example 23

1.3.3 Making predictions 24

1.3.4 Summary 24

1.4 Nonlinear response from a linear model 25

1.5 Generalisation and over-fitting 28

1.5.1 Validation data 29

1.5.2 Cross-validation 29

1.5.3 Computational scaling of K-fold cross-validation . . . 32

1.6 Regularised least squares 33

1.7 Exercises 35

Further reading 37

2 Linear Modelling: A Maximum Likelihood Approach **39**

2.1 Errors as noise 39

2.1.1 Thinking generatively 40

2.2 Random variables and probability 41

2.2.1 Random variables 41
2.2.2 Probability and distributions 42
2.2.3 Adding probabilities 44
2.2.4 Conditional probabilities 44
2.2.5 Joint probabilities 45
2.2.6 Marginalisation 47
2.2.7 Aside – Bayes' rule 49
2.2.8 Expectations 50
2.3 Popular discrete distributions 53
2.3.1 Bernoulli distribution 53
2.3.2 Binomial distribution 53
2.3.3 Multinomial distribution 54
2.4 Continuous random variables – density functions 55
2.5 Popular continuous density functions 58
2.5.1 The uniform density function 58
2.5.2 The beta density function 60
2.5.3 The Gaussian density function 61
2.5.4 Multivariate Gaussian 62
2.5.5 Summary 65
2.6 Thinking generatively...continued 65
2.7 Likelihood 67
2.7.1 Dataset likelihood 68
2.7.2 Maximum likelihood 69
2.7.3 Characteristics of the maximum likelihood solution . . 71
2.7.4 Maximum likelihood favours complex models 74
2.8 The bias-variance tradeoff 75
2.8.1 Summary 75
2.9 Effect of noise on parameter estimates 76
2.9.1 Uncertainty in estimates 78
2.9.2 Comparison with empirical values 81
2.9.3 Variability in model parameters – Olympics data . . . 82
2.10 Variability in predictions 83
2.10.1 Predictive variability – an example 85
2.10.2 Expected values of the estimators 86
2.10.3 Summary 90
2.11 Exercises 90
Further reading 93

3 The Bayesian Approach to Machine Learning 95
3.1 A coin game 95
3.1.1 Counting heads 97
3.1.2 The Bayesian way 98
3.2 The exact posterior 103
3.3 The three scenarios 104
3.3.1 No prior knowledge 104

3.3.2 The fair coin scenario 111
3.3.3 A biased coin 114
3.3.4 The three scenarios – a summary 116
3.3.5 Adding more data 116
3.4 Marginal likelihoods 117
3.4.1 Model comparison with the marginal likelihood 118
3.5 Hyper-parameters 119
3.6 Graphical models 120
3.6.1 Summary 121
3.7 A Bayesian treatment of the Olympics 100 m data 122
3.7.1 The model 122
3.7.2 The likelihood 124
3.7.3 The prior 124
3.7.4 The posterior 124
3.7.5 A first-order polynomial 126
3.7.6 Making predictions 129
3.8 Marginal likelihood for polynomial model order selection . . 131
3.9 Chapter summary 133
3.10 Exercises 133
Further reading 137

4 Bayesian Inference **139**
4.1 Non-conjugate models 139
4.2 Binary responses 140
4.2.1 A model for binary responses 140
4.3 A point estimate – the MAP solution 143
4.4 The Laplace approximation 149
4.4.1 Laplace approximation example: Approximating a gamma density 150
4.4.2 Laplace approximation for the binary response model 151
4.5 Sampling techniques 154
4.5.1 Playing darts 154
4.5.2 The Metropolis–Hastings algorithm 156
4.5.3 The art of sampling 164
4.6 Summary 165
4.7 Exercises 165
Further reading 167

5 Classification **169**
5.1 The general problem 169
5.2 Probabilistic classifiers 170
5.2.1 The Bayes classifier 170
5.2.1.1 Likelihood – class-conditional distributions . 171
5.2.1.2 Prior class distribution 171
5.2.1.3 Example – Gaussian class-conditionals 172

5.2.1.4 Making predictions 173
5.2.1.5 The naive Bayes assumption 175
5.2.1.6 Example – classifying text 175
5.2.1.7 Smoothing 177
5.2.2 Logistic regression 179
5.2.2.1 Motivation 180
5.2.2.2 Nonlinear decision functions 181
5.2.2.3 Nonparametric models - the Gaussian process 182
5.3 Nonprobabilistic classifiers 183
5.3.1 K-nearest neighbours 183
5.3.1.1 Choosing K 184
5.3.2 Support vector machines and other kernel methods . . 186
5.3.2.1 The margin 186
5.3.2.2 Maximising the margin 187
5.3.2.3 Making predictions 190
5.3.2.4 Support vectors 191
5.3.2.5 Soft margins 192
5.3.2.6 Kernels 193
5.3.3 Summary 197
5.4 Assessing classification performance 198
5.4.1 Accuracy – 0/1 loss 198
5.4.2 Sensitivity and specificity 198
5.4.3 The area under the ROC curve 199
5.4.4 Confusion matrices 201
5.5 Discriminative and generative classifiers 203
5.6 Summary 203
5.7 Exercises 203
Further reading 205

6 Clustering **207**
6.1 The general problem 207
6.2 K-means clustering 208
6.2.1 Choosing the number of clusters 210
6.2.2 Where K-means fails 212
6.2.3 Kernelised K-means 212
6.2.4 Summary 214
6.3 Mixture models 215
6.3.1 A generative process 216
6.3.2 Mixture model likelihood 217
6.3.3 The EM algorithm 219
6.3.3.1 Updating π_k 220
6.3.3.2 Updating $\boldsymbol{\mu}_k$ 221
6.3.3.3 Updating $\boldsymbol{\Sigma}_k$ 222
6.3.3.4 Updating q_{nk} 223
6.3.3.5 Some intuition 224

试读结束，需要全本PDF请购买 www.ertongbook.com

6.3.4 Example . . . 225
6.3.5 EM finds local optima . . . 226
6.3.6 Choosing the number of components . . . 228
6.3.7 Other forms of mixture components . . . 230
6.3.8 MAP estimates with EM . . . 232
6.3.9 Bayesian mixture models . . . 233
6.4 Summary . . . 234
6.5 Exercises . . . 234
Further reading . . . 237

7 Principal Components Analysis and Latent Variable Models 239
7.1 The general problem . . . 239
7.1.1 Variance as a proxy for interest . . . 239
7.2 Principal components analysis . . . 242
7.2.1 Choosing D . . . 247
7.2.2 Limitations of PCA . . . 247
7.3 Latent variable models . . . 248
7.3.1 Mixture models as latent variable models . . . 248
7.3.2 Summary . . . 249
7.4 Variational Bayes . . . 249
7.4.1 Choosing $Q(\boldsymbol{\theta})$ . . . 251
7.4.2 Optimising the bound . . . 252
7.5 A probabilistic model for PCA . . . 252
7.5.1 $Q_\tau(\tau)$ . . . 254
7.5.2 $Q_{\mathbf{x}_n}(\mathbf{x}_n)$ . . . 256
7.5.3 $Q_{\mathbf{w}_m}(\mathbf{w}_m)$ . . . 257
7.5.4 The required expectations . . . 258
7.5.5 The algorithm . . . 258
7.5.6 An example . . . 260
7.6 Missing values . . . 260
7.6.1 Missing values as latent variables . . . 262
7.6.2 Predicting missing values . . . 264
7.7 Non-real-valued data . . . 264
7.7.1 Probit PPCA . . . 264
7.7.2 Visualising parliamentary data . . . 268
7.7.2.1 Aside – relationship to classification . . . 272
7.8 Summary . . . 273
7.9 Exercises . . . 273
Further reading . . . 275

Glossary 277

Index 283

List of Tables

1.1 Synthetic dataset for linear regression example. 9

1.2 Olympics men's 100 m data. 11

1.3 Olympics women's 100 m data. 13

1.4 Some useful identities when differentiating with respect to a vector. 21

2.1 Events we might want to model with random variables. . . . 42

5.1 Likelihood and priors for $\mathbf{x}_{\mathsf{new}} = [2, 0]^{\mathsf{T}}$ for the Gaussian class-conditional Bayesian classification example. 174

5.2 A binary confusion matrix. 201

5.3 Confusion matrix for the 20 class newsgroup data. 202

List of Figures

1.1 Winning men's 100 m times at the Summer Olympics since 1896. 2

1.2 Effect of varying w_0 and w_1 in the linear model defined by Equation 1.1. 4

1.3 Example loss function of one parameter (w). 5

1.4 Data and function for the worked example of Section 1.1.5. . 10

1.5 The least squares fit ($f(x; w_0, w_1) = 36.416 - 0.013x$) to the men's 100 m Olympics dataset. 12

1.6 Zoomed-in plot of the winning time in the Olympics men's 100 m sprint from 1980 showing predictions for both the 2012 and 2016 Olympics. 13

1.7 Women's Olympics 100 m data with a linear model that minimises the squared loss. 14

1.8 Male and female functions extrapolated into the future. . . . 14

1.9 Example of linear and quadratic models fitted to a dataset generated from a quadratic function. 26

1.10 8th order polynomial fitted to the Olympics 100 m men's sprint data. 27

1.11 Least squares fit of $f(x; \mathbf{w}) = w_0 + w_1 x + w_2 \sin\left(\frac{x-a}{b}\right)$ to the 100 m sprint data ($a = 2660, b = 4.3$). 28

1.12 Training and validation loss for Olympics men's 100 m data. . 29

1.13 Generalisation ability of 1st, 4th and 8th order polynomials on Olympics men's 100 m data. 30

1.14 Cross-validation. 30

1.15 Mean LOOCV loss as polynomials of increasing order are fitted to the Olympics men's 100 m data. 31

1.16 The training, testing and leave-one-out loss curves obtained for a noisy cubic function where a sample size of 50 is available for training and LOOCV estimation. 32

1.17 Effect of varying the regularisation parameter λ for a 5th order polynomial function. 34

2.1 Linear fit to the Olympics men's 100 m data with errors highlighted. 40

2.2 Dataset generated from a linear model. 41

2.3 An example of the probability distribution function for a binomial random variable when $N = 50$ and $q = 0.7$. 54
2.4 An example of the uniform pdf. 59
2.5 Effect of increasing the number of samples on the approximation to the expectation given in Equation 2.25 where $p(y) = \mathcal{U}(0, 1)$. 60
2.6 Examples of beta pdfs with three different pairs of parameters. 61
2.7 Three Gaussian pdfs with different means and variances. . . . 61
2.8 Example surface (left) and contour (right) plots for two different two-dimensional Gaussian pdfs. 63
2.9 Dataset generated from a linear model with Gaussian errors. 66
2.10 Likelihood function for the year 1980. 68
2.11 Model complexity example with Olympics men's 100 m data. 74
2.12 Data generated from the model given in Equation 2.39 and the true function. 76
2.13 Variability in $\widehat{\mathbf{w}}$ for 10,000 datasets generated from the model described in Equation 2.39. 77
2.14 Functions inferred from 10 datasets generated from the model given in Equation 2.39 as well as the true function. 77
2.15 Two example datasets with different noise levels and the corresponding likelihood function. 81
2.16 Ten samples of $\mathbf{w}$ using the distribution given in Equation 2.48. 83
2.17 (a) Example data set. (b), (c) and (d) Predictive error bars for a linear, cubic and 6th order model, respectively. 85
2.18 Examples of functions with parameters drawn from a Gaussian with mean $\widehat{\mathbf{w}}$ and covariance $\mathsf{cov}\{\widehat{\mathbf{w}}\}$ for the example data set shown in Figure 2.17(a). 86
2.19 Evolution of the theoretical and empirical estimates of $\mathbf{E}_{p(\mathbf{t}|\mathbf{X},\mathbf{w},\sigma^2)}\left\{\widehat{\sigma^2}\right\}$ as the number of data points increases. . . 89

3.1 The binomial density function when $N = 10$ and $r = 0.5$. . . 96
3.2 The binomial density function when $N = 10$ and $r = 0.9$. . . 97
3.3 Examples of the likelihood $p(y_N|r)$ as a function of r for two scenarios. 99
3.4 Examples of prior densities, $p(r)$, for r for three different scenarios. 100
3.5 Examples of three possible posterior distributions $p(r|y_N)$. . . 102
3.6 Evolution of $p(r|y_N)$ as the number of observed coin tosses increases. 106
3.7 Evolution of expected value (a) and variance (b) of r as coin toss data is added to the posterior. 108
3.8 The posterior after six and seven tosses. 109
3.9 Posterior distribution after observing 10 tosses and 20 tosses 111

3.10 Evolution of the posterior $p(r|y_N)$ as more coin tosses are observed for the fair coin scenario. 112

3.11 Evolution of $\mathbf{E}_{p(r|y_N)}\{R\}$ (a) and $\mathsf{var}\{R\}$ (b) as the 20 coin tosses are observed for the fair coin scenario. 113

3.12 Evolution of the posterior $p(r|y_N)$ as more coin tosses are observed for the biased coin scenario. 115

3.13 Evolution of $\mathbf{E}_{p(r|y_N)}\{R\}$ (a) and $\mathsf{var}\{R\}$ (b) as the 20 coin tosses are observed for the biased coin scenario. 116

3.14 The posterior densities for the three scenarios after 100 coin tosses and 1000 coin tosses. 117

3.15 Marginal likelihood contours for the coin example. 119

3.16 Graphical model examples. 121

3.17 Graphical model for the Bayesian model of the Olympics men's 100 m data. 123

3.18 Olympics data with rescaled x values. 126

3.19 Gaussian prior used for the Olympics 100 m data (a) and some functions created with samples drawn from the prior (b). . . 127

3.20 Evolution of the posterior density and example functions drawn from the posterior for the Olympics data after increasing numbers of observations have been added. 128

3.21 Posterior density (a) and sampled functions (b) for the Olympics data when all 27 data points have been added. . . . 129

3.22 Posterior density (a) and sampled functions (b) for the Olympics data when all 27 data points have been added with more realistic noise variance, $\sigma^2 = 0.05$. 130

3.23 Predictive distribution for the winning time in the men's 100 m sprint at the 2012 London Olympics. 131

3.24 Dataset sampled from the function $t = 5x^3 - x^2 + x$ (a) and marginal likelihoods for polynomials of increasing order (b). . 132

3.25 Marginal likelihoods for the 3rd order polynomial example with $\mathbf{\Sigma}_0 = \sigma_0^2\mathbf{I}$ as σ_0^2 is decreased. 133

4.1 An example of a dataset with a binary response. 140

4.2 The sigmoid function that squashes a real value to always be between 0 and 1. 142

4.3 Evolution of the components of $\mathbf{w}$ throughout the Newton–Raphson procedure to find the $\mathbf{w}$ corresponding to the maximum of the posterior density. 147

4.4 Inferred function in the binary response example. 148

4.5 Examples of the Laplace approximation to the gamma density function given in Equation 4.14. 152

4.6 The Laplace approximation for the binary problem. 152

4.7 Decision boundaries sampled from the Laplace approximation and the predictive probability contours. 153

4.8 A dartboard. 155

4.9 Two examples of random walks where the distribution over the next location is a Gaussian centred at the current location. . 158
4.10 The Metropolis–Hastings algorithm. 159
4.11 Example of the Metropolis–Hastings algorithm in operation. . 160
4.12 Results of applying the MH sampling algorithm to the binary response model. 162
4.13 Two densities that would be tricky to sample from with MH. 164

5.1 Three class classification dataset. 172
5.2 Three class classification dataset with the density contours for the three class-conditional distributions fitted using Equations 5.4 and 5.5. 173
5.3 Contour plots of the classification probabilities for the Bayesian classifier with Gaussian class-conditional distributions. 174
5.4 Density contours for Gaussian class-conditionals with the naive Bayes assumption. 176
5.5 Contour plots of the classification probabilities for the Bayesian classifier with Gaussian class-conditional distributions and the naive Bayes assumption. 176
5.6 Graphical representation of the predictive probabilities for the Bayesian classifier on the 20 newsgroups data. 179
5.7 Binary data and classification probability contours for the logistic regression model described by Equation 5.10. 181
5.8 Cartoon depicting the operation of KNN ($K = 3$). 183
5.9 Binary classification dataset and decision boundaries for $K = 1$ and $K = 5$. 184
5.10 Second binary classification dataset and decision boundaries for $K = 5$ and $K = 39$. 185
5.11 Using cross-validation to find the best value of K. 186
5.12 The classification margin γ, defined as the perpendicular distance from the decision boundary to the closest points on either side. 187
5.13 Illustrating the steps taken to compute the margin. 188
5.14 Decision boundary and support vectors for a linear SVM. . . 191
5.15 Decision boundary and support vectors for a linear SVM. . . 192
5.16 Decision boundary and support vectors for a linear SVM with a soft margin for two values of the margin parameter C. . . . 194
5.17 A binary dataset for which a linear decision boundary would not be appropriate. 194
5.18 Decision boundary and support vectors for the dataset in Figure 5.17 using a Gaussian kernel with the kernel parameter $\gamma = 1$ and $C = 10$. 196
5.19 Decision boundary and support vectors for the dataset in Figure 5.17 using a Gaussian kernel with different values of the kernel parameter γ and $C = 10$. 197

试读结束，需要全本PDF请购买 www.ertongbook.com