Machine Learning:

An Introductory Handbook Using R

Dr. Karen Mazidi

Copyright © 2018 Karen Mazidi

WWW.KARENMAZIDI.COM

This work is released under Creative Commons License CC BY-NC-ND 4.0

You are free to use this work for your personal educational purposes. You may not redistribute any portion of this work without prior permission of the author.

FIRST EDITION, JULY 2018

Created with a LaTeX template by Mathias Legrand `http://www.LaTeXTemplates.com`

Contents

I Part One: Introduction to Machine Learning and R

1 The Craft of Machine Learning 17

1.1 Machine Learning 18

1.1.1 Data 18

1.1.2 Patterns 19

1.1.3 Predictions from Data 19

1.1.4 Accuracy 19

1.1.5 Actions 19

1.2 Machine Learning Scenarios 20

1.2.1 Informative v. Active 20

1.2.2 Supervised v. unsupervised learning 20

1.2.3 Regression v. classification 20

1.3 Machine learning v. traditional algorithms 20

1.4 Terminology 21

1.5 Notation 22

2 Learning R 23

2.1 Installing R, RStudio 23

2.1.1 Getting Started in R 24

2.2 R Data 25

2.2.1 Vector 26

2.2.2 Lists 27

2.2.3 Matrix 28
2.2.4 Arrays 28
2.2.5 Data Frames 28
2.3 Data Exploration 29
2.4 Visual Data Exploration 30
2.5 Factors 32
2.5.1 Adding a Factor Column to a Data Frame 33
2.6 R Notebooks 34
2.7 Control structures 36
2.7.1 if-else 37
2.7.2 Defining and Calling Functions 37
2.7.3 Loops 39
2.8 R Style 40
2.9 Summary 41
3 Data Visualization in R 43
3.1 Data Visualization of One Variable 44
3.1.1 Quantitative Vectors 44
3.1.2 Plotting Qualitative Vectors 47
3.2 Data Visualization of Two Vectors 48
3.2.1 Both X and Y are Qualitative 48
3.2.2 X is Qualitative, Y is Quantitative 50
3.2.3 X is Quantitative, Y is Qualitative 51
3.2.4 X and Y are both Quantitative 52
3.3 Summary 52
II Part Two: Linear Models
4 Linear Regression 57
4.1 Overview 57
4.2 Linear Regression in R 57
4.3 Metrics 59
4.3.1 Metrics for Data Analysis 60
4.3.2 Linear Regression Metrics for Model Fit 60
4.3.3 Metrics for Test Set Evaluation 62
4.4 The Algorithm 64
4.5 Mathematical Foundations 65
4.5.1 Gradient descent 68

4.6 Multiple Linear Regression 68

4.6.1 Interpreting summary() output for a linear model 70

4.6.2 Residuals 71

4.6.3 Dummy variables 73

4.6.4 The anova0 function 73

4.7 Polynomial Linear Regression 74

4.8 Model Fitting and Assumptions 76

4.8.1 Overfitting v. underfitting 76

4.8.2 Bias and variance 76

4.8.3 Linear Model Assumptions 77

4.9 Advanced Topic: Regularization 77

4.10 Summary 79

4.10.1 Terminology 79

4.10.2 Quick Reference 80

4.10.3 Lab 80

4.10.4 Exploring Concepts 81

4.10.5 Going Further 81

5 Logistic Regression 83

5.1 Overview 83

5.2 Logistic Regression in R 84

5.2.1 Plotting factor data 84

5.2.2 Train and Test on the Plasma Data 85

5.2.3 Interpreting summary() in Logistical Regression 86

5.2.4 Evaluation on the Test Data 87

5.2.5 Learning (or Not) From Data 87

5.3 Metrics 88

5.3.1 Accuracy, sensitivity and specificity 88

5.3.2 Kappa 89

5.3.3 ROC Curves and AUC 90

5.3.4 Probability, odds, and log odds 90

5.4 The Algorithm 92

5.5 Mathematical Foundations 93

5.6 Logistic Regression with Multiple Predictors 95

5.7 Advanced Topic: Optimization Methods 97

5.7.1 Gradient Descent 98

5.7.2 Stochastic Gradient Descent 98

5.7.3 Newton’s Method 98

5.7.4 Optimization from Scratch 99

5.8 Multiclass Classification 100

5.9 Generalized Linear Models 103

5.10 **Summary** 103
5.10.1 New Terminology in this Chapter 104
5.10.2 Quick Reference 104
5.10.3 Lab 104
5.10.4 Exploring Concepts 105
5.10.5 Going Further 105

6 Naive Bayes 107
6.1 **Overview** 107
6.2 **Naive Bayes in R** 107
6.3 **Probability Foundations** 110
6.4 **Probability Distributions** 111
6.4.1 Bernouli, Binomial, and Beta Distributions 111
6.4.2 Multinomial and Dirichlet Distributions 113
6.4.3 Gaussian 115
6.5 **The Algorithm** 116
6.6 **Applying the Bayes Theorem** 117
6.7 **Handling Text Data** 118
6.8 **Naive Bayes v. Logistic Regression** 118
6.8.1 Compare to Logistic Regression 118
6.8.2 Naive Bayes Model 120
6.8.3 Generative v. Discriminative Classifiers 121
6.9 **Naive Bayes from Scratch** 121
6.9.1 Probability Tables 121
6.9.2 Conditional Probability for Discrete Data 122
6.9.3 Likelihood for Continuous Data 123
6.9.4 Putting it All Together 123
6.9.5 Results 124
6.10 **Summary** 124
6.10.1 New Terminology in this Chapter 125
6.10.2 Quick Reference 125
6.10.3 Labs 125
6.10.4 Exploring Concepts 126
6.10.5 Going Further 126

7 Support Vector Machines 127
7.1 **Overview** 128
7.2 **SVM in R** 128
7.3 **The Algorithm** 129
7.4 **Mathematical Foundations** 130
7.4.1 Slack Variables and the C Hyperparameter 131

7.5 Kernel Methods and the Gamma Hyperparameter 132

7.6 SVM Regression 133

7.6.1 Radial Kernel SVM 135

7.7 Summary 136

7.7.1 New Terminology in this Chapter 136

7.7.2 Quick Reference 136

7.7.3 Labs 137

7.7.4 Going Further 137

8 The Craft 2: Data Wrangling 139

8.1 Data Organization 139

8.1.1 Reproducible Research 139

8.2 Data Standardization 140

8.2.1 Dealing with NAs 140

8.2.2 Outliers 141

8.3 Time Data 141

8.3.1 Lubridate 141

8.4 Text Data 142

8.4.1 Text Mining Packages tm 142

8.4.2 RTextTools 143

III Part Three: Searching for Similarity

9 Instance-Based Learning with kNN 149

9.1 Overview 149

9.2 kNN in R 150

9.3 The Algorithm 151

9.3.1 Choosing K 151

9.3.2 Curse of Dimensionality 152

9.4 Mathematical Foundations 152

9.4.1 Scaling Data 152

9.5 kNN Regression 153

9.6 Find the Best K 154

9.7 k-fold Cross Validation 155

9.7.1 Creating k Folds 156

9.7.2 Run knnreg() on each Fold 157

9.7.3 Cross Validation with Various K 157

9.7.4 Applying CV to Other Algorithms 158

9.8 Summary 159

9.8.1 Quick Reference 160

9.8.2 Lab 160

9.8.3 Exploring Concepts . . . 161
9.8.4 Going Further . . . 161

10 Clustering . . . 163

10.1 Overview 164

10.2 Clustering in R with k-Means 164

10.3 Metrics 166

10.4 Algorithmic Foundations of k-means 166

10.5 Finding k 167

10.5.1 NbClust . . . 169

10.6 Hierarchical Clustering 169

10.6.1 The Algorithm . . . 170
10.6.2 Cutting the Dendogram . . . 171
10.6.3 Advantages and Disadvantages of Hierarchical Clustering . . . 172

10.7 Summary 172

10.7.1 Quick Reference . . . 172
10.7.2 Lab . . . 173
10.7.3 Exploring Concepts . . . 173
10.7.4 Going Further . . . 173

11 Decision Trees and Random Forests . . . 175

11.1 Overview 175

11.2 Decision Trees in R 176

11.3 The Algorithm 177

11.3.1 Regression . . . 177
11.3.2 Classification . . . 177
11.3.3 Qualitative data . . . 178

11.4 Mathematical Foundations 178

11.4.1 Entropy . . . 178
11.4.2 Information Gain . . . 179
11.4.3 Gini index . . . 180

11.5 Tree Pruning 180

11.6 Random Forests 182

11.7 Cross validation, Bagging, Random Forests in R 182

11.8 Summary 183

11.8.1 Quick Reference . . . 183
11.8.2 Lab . . . 184
11.8.3 Exploring Concepts . . . 185
11.8.4 Going Further . . . 185

12 The Craft 3: Feature Engineering 187

12.1 Feature Selection 187

12.1.1 Feature Selection with caret 187

12.1.2 Ranking Feature Importance 188

12.1.3 Recursive Feature Selection 189

12.2 Feature Engineering 190

12.2.1 Principal Components Analysis 190

12.2.2 Linear Discriminant Analysis 191

IV Part Four: Neural Networks

13 Neural Networks 197

13.1 Neural Network Regression Example 198

13.1.1 Results 199

13.1.2 Hidden Nodes and Layers 200

13.2 The Algorithm 200

13.3 Mathematical Foundations 201

13.4 Neural Network Classification Example 202

13.4.1 Preprocessing for Classification 202

13.5 Neural Network Architecture 204

13.5.1 Training 204

13.5.2 The Backpropagation Algorithm 205

13.5.3 Model Output 205

13.5.4 Visualizing Results 206

13.6 Summary 206

13.6.1 New Terminology 206

13.6.2 Quick Reference 206

13.6.3 Going Further 207

14 Deep Learning 209

14.1 Keras 209

14.1.1 Installing Keras and Tensorflow 209

14.2 Data Representation in TensorFlow 210

14.3 Keras in R Example 210

14.3.1 Data Preparation 211

14.3.2 Build the Model 211

14.3.3 Train and Evaluate 212

14.4 Deep Learning Basics 212

14.4.1 Layers and Units 213

14.4.2 Activation Functions 214

14.4.3 Optimization Schemes and Loss Functions 214

14.5 Bring Your Own Data 214
14.6 Deep Architecture 217
14.6.1 Convolutional Neural Networks 217
14.6.2 Recurrent Neural Networks 217
14.7 Summary 218
14.7.1 Going Further 218

15 The Craft 4: Never Stop Learning 219
15.1 Machine Learning Trends 220
15.2 Inside the Black Box 220

V Part Five: Modeling the World

16 Bayes Nets 225
16.1 Bayes Nets in R 226
16.1.1 Querying the Network 227
16.2 Bayesian Net Semantics 227
16.2.1 Review of Graph Terminology 227
16.2.2 Graph Structure 227
16.2.3 Reasoning with Graphs 228
16.3 The Algorithm 228
16.4 Example: Coronary Data 229
16.5 Summary 229
16.5.1 Going Further 230

17 Markov Models 231
17.1 Overview 231
17.2 Markov Model 231
17.3 Hidden Markov Model 233
17.4 HMM in R 233
17.5 Metrics 235
17.6 The Algorithm 235
17.7 Another HMM in R 235
17.8 Summary 236
17.8.1 Going Further 237

18 Reinforcement Learning 239
18.1 Overview 239
18.2 The Markov Decision Process 240
18.3 MDPtoolbox 241

18.4 Reinforcement Learning 242
18.5 The ReinforcementLearning Package 242
18.6 Summary 243
18.6.1 Going Further 243

19 The Craft 5: Algorithms 245
19.1 The Five Tribes 245
19.2 Choosing an Algorithm 246

VI Part Six: Supplementary Material

20 Modern R 249
20.1 Package dplyr 250
20.1.1 Tibble 250
20.1.2 Function glimpse() 250
20.1.3 Select 251
20.1.4 Mutate 251
20.1.5 Filter 252
20.1.6 Arrange 252
20.1.7 Summarize 252
20.1.8 Pipes 252
20.1.9 Going Further with dplyr 253
20.2 Package ggplot2 253
20.2.1 Improving Appearance 254
20.2.2 Facet Grid 254
20.2.3 Histogram 255
20.2.4 Boxplot and Rug 255
20.2.5 Density Plot 256
20.2.6 Bubble Chart 256
20.2.7 Grid 257
20.2.8 Going Further with Graphs 257
20.3 Shiny 257

21 Big Data with R 261
21.1 Memory, Data and R 261
21.1.1 Limit Data Size 262
21.2 Subset Data Base with dplyr 262
21.3 The ff Package 263
21.3.1 Read in the data 263

21.4 Amazon Cloud Services 265
21.5 Google Cloud Services 265
21.6 Big Data: A Sham? 266

22 Sampling Big Data 267
22.1 What is Sampling? 267
22.2 Sampling Example 268
22.3 Current Practices 269

Index 275

Preface to the Book

I accidentally started writing this book during Spring Break 2018. I was driving down a long Texas road out in the country, thinking about a problem with the undergraduate machine learning class I had been teaching. Chiefly, there was no book suitable for my course. There are several excellent machine learning books for graduate students, but most are too heavy on the math and too light on practical issues for undergraduate students. There are a few books for undergraduate students, but they didn't cover all the topics I was required to teach for my course at The University of Texas at Dallas. The solution I came up with was to format my own notes in latex. I scribbled an outline on a piece of paper as I drove down the road. When I came home I started looking for a suitable latex template and found this one which I really liked. It was designed for a book and I thought, why not just make this a book? And so the book began.

Many years ago I authored several textbooks on microprocessors and microcontrollers with my husband. These books were quite successful and provided a nice second income for our family. I was able to write the books at home while taking care of our two rambunctious sons. However, the days when one can write textbooks as a job are gone. As soon as a book is published, it is pirated, and the pdf is available for free to anyone in the world. That is not entirely a bad thing in my opinion. Yes it is stealing intellectual property from the owner but on the other hand it makes information available to people who want to learn but perhaps can't pay the exhorbitant price that publishing companies charge for textbooks. And so my goals are not monetary. The pdf of this book will remain free, although there is a print version available on Amazon for those who prefer to read from paper. The Amazon print book is printed in grayscale so refer to the pdf when color is important in graphs.

My goals in writing this book are the following:

- To create an undergraduate book that covers the most popular machine learning algorithms in a hands-on approach that reinforces understanding of the algorithms.
- To provide an accessible introduction to the field to professionals who want to get started with machine learning.
- To increase the number of practitioners in the field. I believe that machine learning is to the present and future what computer programming was in the 1970s when I started: a revolutionary new way of solving problems.
- To help people enhance their skill set. If I can introduce people to these skills it will help them in their career goals.

I hope you enjoy the book. This first edition covers what I teach in a one-semester undergraduate Introduction to Machine Learning course. However, I consider it to be a work in progress and I will include additional material in the future, as well as some short video vignettes. Links will be posted on my website: `www.karenmazidi.com` and my blog: `http://karenmazidi.blogspot.com/`

The book has a companion github site for code samples: `https://github.com/kjmazidi/Machine_Learning/`