

# Machine Learning with R

*Third Edition*

**Expert techniques for predictive modeling**

**Brett Lantz**

**Packt>**

**BIRMINGHAM - MUMBAI**

# Table of Contents

<b><u>Preface</u></b>	<b>jx</b>
<b><u>Chapter 1 - Introducing Machine Learning</u></b>	<b>1</b>
<b>The origins of machine learning</b>	<b>2</b>
<b>Uses and abuses of machine learning</b>	<b>4</b>
Machine learning successes	5
The limits of machine learning	6
Machine learning ethics	7
<b>How machines learn</b>	<b>10</b>
Data storage	12
Abstraction	12
Generalization	14
Evaluation	16
<b>Machine learning in practice</b>	<b>18</b>
Types of input data	19
Types of machine learning algorithms	20
Matching input data to algorithms	23
<b>Machine learning with R</b>	<b>24</b>
Installing R packages	25
Loading and unloading R packages	26
Installing RStudio	27
<b>Summary</b>	<b>28</b>
<b><u>Chapter 2 - Managing and Understanding Data</u></b>	<b>29</b>
<b>R data structures</b>	<b>30</b>
Vectors	30
Factors	32
Lists	34

Data frames	36
Matrices and arrays	39
<b>Managing data with R</b>	<b>41</b>
Saving, loading, and removing R data structures	41
Importing and saving data from CSV files	42
<b>Exploring and understanding data</b>	<b>44</b>
Exploring the structure of data	44
Exploring numeric variables	46
Measuring the central tendency – mean and median	46
Measuring spread – quartiles and the five-number summary	48
Visualizing numeric variables – boxplots	51
Visualizing numeric variables – histograms	52
Understanding numeric data – uniform and normal distributions	54
Measuring spread – variance and standard deviation	55
Exploring categorical variables	57
Measuring the central tendency – the mode	58
Exploring relationships between variables	60
Visualizing relationships – scatterplots	60
Examining relationships – two-way cross-tabulations	62
<b>Summary</b>	<b>64</b>
<b>Chapter 3 - Lazy Learning – Classification Using Nearest Neighbors</b>	<b>65</b>
<b>    Understanding nearest neighbor classification</b>	<b>66</b>
The k-NN algorithm	66
Measuring similarity with distance	70
Choosing an appropriate k	71
Preparing data for use with k-NN	73
Why is the k-NN algorithm lazy?	75
<b>    Example – diagnosing breast cancer with the k-NN algorithm</b>	<b>76</b>
Step 1 – collecting data	77
Step 2 – exploring and preparing the data	78
Transformation – normalizing numeric data	80
Data preparation – creating training and test datasets	81
Step 3 – training a model on the data	82
Step 4 – evaluating model performance	84
Step 5 – improving model performance	85
Transformation – z-score standardization	85
Testing alternative values of k	87
<b>    Summary</b>	<b>88</b>
<b>Chapter 4 - Probabilistic Learning – Classification Using Naive Bayes</b>	<b>89</b>
<b>    Understanding Naive Bayes</b>	<b>90</b>

Basic concepts of Bayesian methods	90
Understanding probability	91
Understanding joint probability	92
Computing conditional probability with Bayes' theorem	94
The Naive Bayes algorithm	96
Classification with Naive Bayes	98
The Laplace estimator	100
Using numeric features with Naive Bayes	102
<b>Example – filtering mobile phone spam with the Naive Bayes algorithm</b>	<b>103</b>
Step 1 – collecting data	103
Step 2 – exploring and preparing the data	105
Data preparation – cleaning and standardizing text data	106
Data preparation – splitting text documents into words	112
Data preparation – creating training and test datasets	114
Visualizing text data – word clouds	115
Data preparation – creating indicator features for frequent words	118
Step 3 – training a model on the data	120
Step 4 – evaluating model performance	121
Step 5 – improving model performance	122
<b>Summary</b>	<b>123</b>
<b>Chapter 5 - Divide and Conquer – Classification Using Decision Trees and Rules</b>	<b>125</b>
<b>Understanding decision trees</b>	<b>126</b>
Divide and conquer	127
The C5.0 decision tree algorithm	131
Choosing the best split	132
Pruning the decision tree	134
<b>Example – identifying risky bank loans using C5.0 decision trees</b>	<b>135</b>
Step 1 – collecting data	136
Step 2 – exploring and preparing the data	136
Data preparation – creating random training and test datasets	138
Step 3 – training a model on the data	139
Step 4 – evaluating model performance	143
Step 5 – improving model performance	144
Boosting the accuracy of decision trees	144
Making some mistakes cost more than others	146
<b>Understanding classification rules</b>	<b>148</b>
Separate and conquer	149
The 1R algorithm	151
The RIPPER algorithm	153
Rules from decision trees	155
What makes trees and rules greedy?	156

<b>Example – identifying poisonous mushrooms with rule learners</b>	<b>158</b>
Step 1 – collecting data	158
Step 2 – exploring and preparing the data	159
Step 3 – training a model on the data	160
Step 4 – evaluating model performance	162
Step 5 – improving model performance	163
<b>Summary</b>	<b>165</b>
<b>Chapter 6 - Forecasting Numeric Data – Regression Methods</b>	<b>167</b>
<b>Understanding regression</b>	<b>168</b>
Simple linear regression	170
Ordinary least squares estimation	173
Correlations	175
Multiple linear regression	177
<b>Example – predicting medical expenses using linear regression</b>	<b>182</b>
Step 1 – collecting data	182
Step 2 – exploring and preparing the data	183
Exploring relationships among features – the correlation matrix	185
Visualizing relationships among features – the scatterplot matrix	186
Step 3 – training a model on the data	189
Step 4 – evaluating model performance	192
Step 5 – improving model performance	193
Model specification – adding nonlinear relationships	194
Transformation – converting a numeric variable to a binary indicator	194
Model specification – adding interaction effects	195
Putting it all together – an improved regression model	196
Making predictions with a regression model	197
<b>Understanding regression trees and model trees</b>	<b>200</b>
Adding regression to trees	201
<b>Example – estimating the quality of wines with regression trees and model trees</b>	<b>203</b>
Step 1 – collecting data	204
Step 2 – exploring and preparing the data	205
Step 3 – training a model on the data	207
Visualizing decision trees	208
Step 4 – evaluating model performance	210
Measuring performance with the mean absolute error	211
Step 5 – improving model performance	212
<b>Summary</b>	<b>216</b>
<b>Chapter 7 - Black Box Methods – Neural Networks and Support</b>	
<b><u>Vector Machines</u></b>	<b><u>217</u></b>
<b>Understanding neural networks</b>	<b>218</b>

From biological to artificial neurons	219
Activation functions	221
Network topology	223
The number of layers	224
The direction of information travel	225
The number of nodes in each layer	226
Training neural networks with backpropagation	227
<b>Example – modeling the strength of concrete with ANNs</b>	<b>229</b>
Step 1 – collecting data	230
Step 2 – exploring and preparing the data	230
Step 3 – training a model on the data	232
Step 4 – evaluating model performance	235
Step 5 – improving model performance	236
<b>Understanding support vector machines</b>	<b>241</b>
Classification with hyperplanes	242
The case of linearly separable data	244
The case of nonlinearly separable data	246
Using kernels for nonlinear spaces	247
<b>Example – performing OCR with SVMs</b>	<b>249</b>
Step 1 – collecting data	250
Step 2 – exploring and preparing the data	251
Step 3 – training a model on the data	252
Step 4 – evaluating model performance	254
Step 5 – improving model performance	256
Changing the SVM kernel function	256
Identifying the best SVM cost parameter	257
<b>Summary</b>	<b>259</b>
<b>Chapter 8 - Finding Patterns – Market Basket Analysis Using Association Rules</b>	<b><u>261</u></b>
<b>Understanding association rules</b>	<b>262</b>
The Apriori algorithm for association rule learning	263
Measuring rule interest – support and confidence	265
Building a set of rules with the Apriori principle	266
<b>Example – identifying frequently purchased groceries with association rules</b>	<b>268</b>
Step 1 – collecting data	268
Step 2 – exploring and preparing the data	269
Data preparation – creating a sparse matrix for transaction data	270
Visualizing item support – item frequency plots	273
Visualizing the transaction data – plotting the sparse matrix	274
Step 3 – training a model on the data	276
Step 4 – evaluating model performance	279

Step 5 – improving model performance	282
Sorting the set of association rules	282
Taking subsets of association rules	283
Saving association rules to a file or data frame	284
<b>Summary</b>	<b>285</b>
<b><u>Chapter 9 - Finding Groups of Data – Clustering with k-means</u></b>	<b><u>287</u></b>
<b>Understanding clustering</b>	<b>288</b>
Clustering as a machine learning task	288
The k-means clustering algorithm	290
Using distance to assign and update clusters	292
Choosing the appropriate number of clusters	296
<b>Finding teen market segments using k-means clustering</b>	<b>298</b>
Step 1 – collecting data	298
Step 2 – exploring and preparing the data	299
Data preparation – dummy coding missing values	300
Data preparation – imputing the missing values	302
Step 3 – training a model on the data	303
Step 4 – evaluating model performance	306
Step 5 – improving model performance	309
<b>Summary</b>	<b>311</b>
<b><u>Chapter 10 - Evaluating Model Performance</u></b>	<b><u>313</u></b>
<b>Measuring performance for classification</b>	<b>314</b>
Understanding a classifier's predictions	314
A closer look at confusion matrices	318
Using confusion matrices to measure performance	320
Beyond accuracy – other measures of performance	322
The kappa statistic	324
Sensitivity and specificity	327
Precision and recall	329
The F-measure	331
Visualizing performance tradeoffs with ROC curves	331
<b>Estimating future performance</b>	<b>337</b>
The holdout method	338
Cross-validation	341
Bootstrap sampling	344
<b>Summary</b>	<b>345</b>
<b><u>Chapter 11 - Improving Model Performance</u></b>	<b><u>347</u></b>
<b>Tuning stock models for better performance</b>	<b>348</b>
Using caret for automated parameter tuning	349
Creating a simple tuned model	351
Customizing the tuning process	355

<b>Improving model performance with meta-learning</b>	<b>359</b>
Understanding ensembles	359
Bagging	362
Boosting	364
Random forests	367
Training random forests	368
Evaluating random forest performance in a simulated competition	370
<b>Summary</b>	<b>374</b>
<b>Chapter 12 - Specialized Machine Learning Topics</b>	<b>375</b>
<b>Managing and preparing real-world data</b>	<b>376</b>
Making data "tidy" with the tidyverse packages	376
Generalizing tabular data structures with tibble	377
Speeding and simplifying data preparation with dplyr	378
Reading and writing to external data files	379
Importing tidy tables with readr	380
Importing Microsoft Excel, SAS, SPSS, and Stata files with rio	380
Querying data in SQL databases	381
The tidy approach to managing database connections	381
Using a database backend with dplyr	384
A traditional approach to SQL connectivity with RODBC	385
<b>Working with online data and services</b>	<b>386</b>
Downloading the complete text of web pages	387
Parsing the data within web pages	389
Parsing XML documents	391
Parsing JSON from web APIs	392
<b>Working with domain-specific data</b>	<b>396</b>
Analyzing bioinformatics data	396
Analyzing and visualizing network data	397
<b>Improving the performance of R</b>	<b>401</b>
Managing very large datasets	401
Making data frames faster with data.table	402
Creating disk-based data frames with ff	403
Using massive matrices with bigmemory	404
Learning faster with parallel computing	404
Measuring execution time	406
Working in parallel with multicore and snow	406
Taking advantage of parallel with foreach and doParallel	409
Training and evaluating models in parallel with caret	411
Parallel cloud computing with MapReduce and Hadoop	412
Parallel cloud computing with Apache Spark	413
Deploying optimized learning algorithms	414
Building bigger regression models with biglm	415
Growing random forests faster with ranger	415

Table of Contents

Growing massive random forests with bigf	416
A faster machine learning computing engine with H2O	416
GPU computing	418
Flexible numeric computing and machine learning with TensorFlow	419
An interface for deep learning with Keras	420
<b>Summary</b>	<b>421</b>
<b><u>Other Books You May Enjoy</u></b>	<b><u>423</u></b>
Leave a review - let other readers know what you think	426
<b>Index</b>	<b>427</b>