

# Machine Learning with R

## *Second Edition*

Discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R

**Brett Lantz**

**[PACKT]**  
PUBLISHING

open source   
community experience distilled

BIRMINGHAM - MUMBAI

# Table of Contents

<b>Preface</b>	<b>ix</b>
<hr/>	
<b>Chapter 1: Introducing Machine Learning</b>	<b>1</b>
<hr/>	
<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	5
Machine learning ethics	7
<b>How machines learn</b>	<b>9</b>
Data storage	10
Abstraction	11
Generalization	13
Evaluation	14
<b>Machine learning in practice</b>	<b>16</b>
Types of input data	17
Types of machine learning algorithms	19
Matching input data to algorithms	21
<b>Machine learning with R</b>	<b>22</b>
Installing R packages	23
Loading and unloading R packages	24
<b>Summary</b>	<b>25</b>
<hr/>	
<b>Chapter 2: Managing and Understanding Data</b>	<b>27</b>
<hr/>	
<b>R data structures</b>	<b>28</b>
Vectors	28
Factors	30
Lists	32
Data frames	35
Matrixes and arrays	37

<b>Managing data with R</b>	<b>39</b>
Saving, loading, and removing R data structures	39
Importing and saving data from CSV files	41
<b>Exploring and understanding data</b>	<b>42</b>
Exploring the structure of data	43
Exploring numeric variables	44
Measuring the central tendency – mean and median	45
Measuring spread – quartiles and the five-number summary	47
Visualizing numeric variables – boxplots	49
Visualizing numeric variables – histograms	51
Understanding numeric data – uniform and normal distributions	53
Measuring spread – variance and standard deviation	54
Exploring categorical variables	56
Measuring the central tendency – the mode	58
Exploring relationships between variables	59
Visualizing relationships – scatterplots	59
Examining relationships – two-way cross-tabulations	61
<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	69
Choosing an appropriate k	70
Preparing data for use with k-NN	72
Why is the k-NN algorithm lazy?	74
<b>Example – diagnosing breast cancer with the k-NN algorithm</b>	<b>75</b>
Step 1 – collecting data	76
Step 2 – exploring and preparing the data	77
Transformation – normalizing numeric data	79
Data preparation – creating training and test datasets	80
Step 3 – training a model on the data	81
Step 4 – evaluating model performance	83
Step 5 – improving model performance	84
Transformation – z-score standardization	85
Testing alternative values of k	86
<b>Summary</b>	<b>87</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
<b>The Naive Bayes algorithm</b>	<b>97</b>
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	104
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	115
Visualizing text data – word clouds	116
Data preparation – creating indicator features for frequent words	119
Step 3 – training a model on the data	121
Step 4 – evaluating model performance	122
Step 5 – improving model performance	123
<b>Summary</b>	<b>124</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	133
Pruning the decision tree	135
<b>Example – identifying risky bank loans using C5.0 decision trees</b>	<b>136</b>
Step 1 – collecting data	136
Step 2 – exploring and preparing the data	137
Data preparation – creating random training and test datasets	138
Step 3 – training a model on the data	140
Step 4 – evaluating model performance	144
Step 5 – improving model performance	145
Boosting the accuracy of decision trees	145
Making mistakes more costlier than others	147
<b>Understanding classification rules</b>	<b>149</b>
Separate and conquer	150
The 1R algorithm	153
The RIPPER algorithm	155
Rules from decision trees	157
What makes trees and rules greedy?	158
<b>Example – identifying poisonous mushrooms with rule learners</b>	<b>160</b>
Step 1 – collecting data	160
Step 2 – exploring and preparing the data	161

Step 3 – training a model on the data	162
Step 4 – evaluating model performance	165
Step 5 – improving model performance	166
<b>Summary</b>	<b>169</b>
<b>Chapter 6: Forecasting Numeric Data – Regression Methods</b>	<b>171</b>
<b>Understanding regression</b>	<b>172</b>
Simple linear regression	174
Ordinary least squares estimation	177
Correlations	179
Multiple linear regression	181
<b>Example – predicting medical expenses using linear regression</b>	<b>186</b>
Step 1 – collecting data	186
Step 2 – exploring and preparing the data	187
Exploring relationships among features – the correlation matrix	189
Visualizing relationships among features – the scatterplot matrix	190
Step 3 – training a model on the data	193
Step 4 – evaluating model performance	196
Step 5 – improving model performance	197
Model specification – adding non-linear relationships	198
Transformation – converting a numeric variable to a binary indicator	198
Model specification – adding interaction effects	199
Putting it all together – an improved regression model	200
<b>Understanding regression trees and model trees</b>	<b>201</b>
Adding regression to trees	202
<b>Example – estimating the quality of wines with regression trees and model trees</b>	<b>205</b>
Step 1 – collecting data	205
Step 2 – exploring and preparing the data	206
Step 3 – training a model on the data	208
Visualizing decision trees	210
Step 4 – evaluating model performance	212
Measuring performance with the mean absolute error	213
Step 5 – improving model performance	214
<b>Summary</b>	<b>218</b>
<b>Chapter 7: Black Box Methods – Neural Networks and Support Vector Machines</b>	<b>219</b>
<b>Understanding neural networks</b>	<b>220</b>
From biological to artificial neurons	221
Activation functions	223

Network topology	225
The number of layers	226
The direction of information travel	227
The number of nodes in each layer	228
Training neural networks with backpropagation	229
<b>Example – Modeling the strength of concrete with ANNs</b>	<b>231</b>
Step 1 – collecting data	232
Step 2 – exploring and preparing the data	232
Step 3 – training a model on the data	234
Step 4 – evaluating model performance	237
Step 5 – improving model performance	238
<b>Understanding Support Vector Machines</b>	<b>239</b>
Classification with hyperplanes	240
The case of linearly separable data	242
The case of nonlinearly separable data	244
Using kernels for non-linear spaces	245
<b>Example – performing OCR with SVMs</b>	<b>248</b>
Step 1 – collecting data	249
Step 2 – exploring and preparing the data	250
Step 3 – training a model on the data	252
Step 4 – evaluating model performance	254
Step 5 – improving model performance	256
<b>Chapter 8: Finding Patterns – Market Basket Analysis Using Association Rules</b>	<b>259</b>
<b>Understanding association rules</b>	<b>260</b>
The Apriori algorithm for association rule learning	261
Measuring rule interest – support and confidence	263
Building a set of rules with the Apriori principle	265
<b>Example – identifying frequently purchased groceries with association rules</b>	<b>266</b>
Step 1 – collecting data	266
Step 2 – exploring and preparing the data	267
Data preparation – creating a sparse matrix for transaction data	268
Visualizing item support – item frequency plots	272
Visualizing the transaction data – plotting the sparse matrix	273
Step 3 – training a model on the data	274
Step 4 – evaluating model performance	277
Step 5 – improving model performance	280
Sorting the set of association rules	280
Taking subsets of association rules	281
Saving association rules to a file or data frame	283
<b>Summary</b>	<b>284</b>

---

<b>Chapter 9: Finding Groups of Data – Clustering with k-means</b>	<b>285</b>
<b>Understanding clustering</b>	<b>286</b>
Clustering as a machine learning task	286
The k-means clustering algorithm	289
Using distance to assign and update clusters	290
Choosing the appropriate number of clusters	294
<b>Example – finding teen market segments using k-means clustering</b>	<b>296</b>
Step 1 – collecting data	297
Step 2 – exploring and preparing the data	297
Data preparation – dummy coding missing values	299
Data preparation – imputing the missing values	300
Step 3 – training a model on the data	302
Step 4 – evaluating model performance	304
Step 5 – improving model performance	308
<b>Summary</b>	<b>310</b>
<b>Chapter 10: Evaluating Model Performance</b>	<b>311</b>
<b>Measuring performance for classification</b>	<b>312</b>
Working with classification prediction data in R	313
A closer look at confusion matrices	317
Using confusion matrices to measure performance	319
Beyond accuracy – other measures of performance	321
The kappa statistic	323
Sensitivity and specificity	326
Precision and recall	328
The F-measure	330
Visualizing performance trade-offs	331
ROC curves	332
<b>Estimating future performance</b>	<b>336</b>
The holdout method	336
Cross-validation	340
Bootstrap sampling	343
<b>Summary</b>	<b>344</b>
<b>Chapter 11: Improving Model Performance</b>	<b>347</b>
<b>Tuning stock models for better performance</b>	<b>348</b>
Using caret for automated parameter tuning	349
Creating a simple tuned model	352
Customizing the tuning process	355
<b>Improving model performance with meta-learning</b>	<b>359</b>
Understanding ensembles	359
Bagging	362
Boosting	366

---

