

Inhalt

Vorwort	XIII
Zusatzmaterial online	XIV
Teil I Einstieg	1
1 Einleitung	3
1.1 Informatik und künstliche Intelligenz	3
1.2 Expertensysteme	4
1.3 Maschinelles Lernen	5
1.3.1 Überwachtes Lernen	7
1.3.2 Unüberwachtes Lernen	7
1.3.3 Verstärkendes Lernen	7
1.4 Methoden und Werkzeuge für das maschinelle Lernen	7
1.5 Zielsetzungen dieses Buchs und Vorgehensweise	9
2 Einführung in R und RStudio	11
2.1 Installation unter Windows	11
2.2 Installation unter Ubuntu Linux	12
2.3 Die RStudio-Oberfläche	15
2.3.1 Das Konsolenfenster	15
2.3.2 Das Source-Fenster	16
2.3.3 Das Files-/Packages-Fenster	18
2.3.3.1 Die Files-Registerkarte	18
2.3.3.2 Die Plot-Registerkarte	20
2.3.3.3 Die Packages-Registerkarte	20
2.3.3.4 Das Environment-Fenster	22
2.3.4 Daten einlesen und speichern	23

2.4	Zuweisungen, Variablen und elementare Datentypen	26
2.5	Zusammengesetzte Datentypen	27
2.5.1	Vektoren	27
2.5.1.1	Union, intersect, setdiff	29
2.5.1.2	Umwandeln des Datentyps, Faktorisierung.....	30
2.5.2	Matrizen	31
2.5.3	Frames	32
2.5.3.1	Hinzufügen von Spalten	32
2.5.3.2	Hinzufügen von Zeilen	33
2.5.3.3	Auswahl von Spalten	33
2.5.3.4	Auswahl von Zeilen und Spalten.....	33
2.5.3.5	Löschen von Spalten oder Zeilen	34
2.5.4	tibbles.....	34
2.5.5	Listen	35
2.5.6	Zeichenketten	36
2.5.7	Datum und Zeit.....	36
2.6	Programmablaufsteuerung.....	37
2.6.1	Blöcke	37
2.6.2	Bedingte Ausführung	37
2.6.3	Schleifen	38
2.6.4	Apply()	38
2.6.5	Gefahren und Benachrichtigungen.....	39
2.6.6	Funktionen.....	39
2.6.7	Der Pipe-Operator	41
2.7	Debugging	41
2.8	Programmierstil.....	43
2.8.1	Variablen- und Funktionsbenennung.....	43
2.8.2	Abstände	44
2.8.3	Blöcke	44
2.8.4	Abschnitte	45
2.9	Formatierhilfen	45
2.9.1	Styler	45
2.9.2	Lintr	45
2.10	Zum Nachschlagen	46
2.11	Einstiegsaufgabe	46

Teil II	Datenvorbereitung	47
3	Daten visualisieren und vorbereiten	49
3.1	Plots mit ggplot2	50
3.1.1	Plot-Architekturen	50
3.1.2	Erster Plot mit ggplot2	50
3.1.3	Aufbau eines Plots	51
3.1.4	Ebenen	54
3.1.5	Geome	55
3.1.6	Stats	57
3.1.7	Themes	60
3.2	Datenaufbereitung	61
3.3	Strukturierung der Daten	62
3.3.1	Werte in Spaltenüberschriften	63
3.3.2	Mehrere Werte in einer Spalte	63
3.3.3	Imputation und Grafikdarstellung	65
3.3.4	Überwachung von Datentypen/Bereichsgrenzen	66
3.4	Datenaufbereitung an Zeitreihen – ein Beispiel	66
3.4.1	Erster Überblick	66
3.4.2	Messfehler	67
3.4.3	Überprüfen der zeitlichen Abfolge	68
3.4.4	Zeitreihe generieren	69
3.4.4.1	Bestimmung der mittleren Tagestemperatur	69
3.4.4.2	Bestimmung der mittleren Jahrestemperatur	70
3.4.4.3	Imputation – Ersetzen von NA-Werten	70
3.5	Zwei kleine Checklisten	72
3.6	Aufgaben	73
4	Datenplausibilität	75
4.1	Hypothesen-Betrachtung	76
4.2	Allgemeine Kenngrößen als Hilfsmittel	77
4.2.1	Mittelwert und Median	77
4.2.2	Varianz	79
4.2.3	Momente höherer Ordnung	80
4.3	Grafische Hilfsmittel	81
4.3.1	Histogramme	81
4.3.2	Box-Plots	81
4.3.3	Summenhäufigkeiten und QQ-Plots	82
4.4	R-Packages zum Erkennen von Ausreißern	83

4.4.1	Das Package „Outliers“	83
4.4.1.1	scores()	84
4.4.1.2	Dixon()	85
4.4.1.3	Cochran.test()	86
4.4.1.4	Grubbs-Test.....	87
4.4.1.5	Weitere Funktionen.....	88
4.5	Datenplausibilität bei mehreren Variablen und weitere Funktionen.....	88
4.6	Aufgaben	89
Teil III Statistische Lernmodelle.....		91
5	Regression.....	93
5.1	Regression mit einer unabhängigen Variablen	94
5.1.1	Methode der kleinsten Fehlerquadrate	94
5.1.2	Homoskedastizität	96
5.1.3	Modellvalidierung.....	98
5.1.3.1	Residuen	101
5.1.4	Vorbeugende Wartung.....	105
5.1.4.1	Das Modellbeispiel	105
5.1.4.2	Etwas mehr Residuenanalyse	106
5.1.4.3	Vorhersagen	110
5.1.5	Erweiterung der Regression auf nichtlineare Funktionen	113
5.1.6	Kreuzvalidierung	114
5.2	Regression mit mehreren unabhängigen Variablen.....	116
5.2.1	Der Boston-Datensatz	116
5.2.2	Durchführung der Regression	122
5.2.3	Modellvalidierung.....	124
5.2.4	Regressionen robuster machen	128
5.2.4.1	Regularisierung	128
5.2.4.2	M-Schätzer.....	130
5.2.4.3	Weitere Alternativen.....	131
5.3	Aufgaben	133
6	Klassifikation	135
6.1	Logistische Regression	136
6.2	Der Perceptron-Algorithmus	140
6.3	Support Vector Machines	143
6.3.1	Der Iris-Datensatz.....	145

6.4	Entscheidungsbaumverfahren.....	146
6.4.1	Entscheidungsbäume	146
6.4.2	Der Iris-Datensatz.....	148
6.4.3	Bagging	148
6.4.4	Random Forests	149
6.4.5	Boosted Regression Trees	149
6.5	Naive Bayes-Klassifikatoren	150
6.5.1	Multinomiale naive Bayes-Klassifikatoren	150
6.5.2	Das spam-Beispiel	152
6.5.3	Likelihood	155
6.5.4	Gauß'sche naive Bayes-Klassifizierer	156
6.6	KNN Nächste-Nachbarn-Klassifikation	158
6.7	Modellbewertung: Devianzen und universellere Kenngrößen	161
6.7.1	Receiver Operating Characteristic, ROC und AUC.....	161
6.7.2	Die R^2 -Metrik.....	163
6.7.3	Eine generalisierte Metrik.....	164
6.7.3.1	Pseudo- R^2	164
6.7.3.2	Devianzen.....	167
6.8	Aufgaben	167
7	Objekte clustern, Merkmale reduzieren und Zeitreihen zerlegen....	169
7.1	K-means-Clustering.....	170
7.2	Korrelationen und Merkmalsreduktion durch Hauptkomponentenanalyse	173
7.2.1	Der beste Standpunkt	173
7.2.2	Kovarianzen.....	174
7.2.3	Kovarianzmatrizen	176
7.2.4	Anwendungen	180
7.2.4.1	Eine kleine Weinprobe	180
7.2.4.2	Der Boston-Datensatz.....	185
7.2.4.3	Erweiterungen der Hauptkomponentenanalyse	188
7.2.4.4	Ausreißererkennung.....	189
7.3	Zeitreihen	189
7.3.1	Komponentenmodelle	189
7.3.2	Glättungsverfahren bei Zeitreihen	190
7.3.2.1	Gleitender Durchschnitt	191
7.3.2.2	Exponentielle Glättung.....	192
7.3.2.3	Holt-Winters-Glättung	194
7.3.2.4	Weitere Glättungsmethoden	195

7.3.3	AR-Modelle.....	196
7.3.3.1	Autokorrelationsfunktionen.....	197
7.3.3.2	Partielle Autokorrelationsfunktion	198
7.3.4	MA-Modelle.....	198
7.3.5	ARMA- und ARIMA-Modelle	200
7.4	Aufgaben	202

Teil IV Lernen mit neuronalen Netzen..... 203

8	Neuronale Netze	205
8.1	Das Perceptron.....	205
8.2	Layers.....	207
8.3	Aktivierungsfunktionen.....	208
8.4	Warum das Stapeln linearer Funktionen nicht sinnvoll ist	211
8.5	Der Regelkreis eines Lernvorgangs	213
8.5.1	Verlustfunktionen und Metriken	213
8.5.2	Epochen und Batches	214
8.5.3	Gradienten, lokale Gradienten und Autodiff	215
8.5.3.1	Manuelle Bestimmung des Gradienten eines Perceptrons	216
8.5.3.2	Berechnung des Gradienten mit CAS oder Differenzenquotienten	217
8.5.3.3	Berechnung des Gradienten mit Forward-Mode Autodiff	217
8.5.3.4	Berechnung des Gradienten mit Backpropagation Autodiff ...	218
8.5.4	Der Optimizer	224
8.5.4.1	Stochastisches Gradientenverfahren (SGD).....	225
8.5.4.2	Momenten-Update	226
8.5.4.3	Nesterov-Moment	226
8.5.4.4	Adagrad	226
8.5.4.5	Rmsprop	227
8.5.4.6	Adadelat.....	227
8.5.4.7	Adam	228
8.5.4.8	ADAm.....	228
8.5.4.9	Was tun, wenn ...?	229
8.6	Regularisierung und Dropouts.....	230
8.6.1	Regularisierung.....	230
8.6.2	Dropouts	230
8.7	Aufgaben	231

9	H2O	233
9.1	Das Unternehmen H2O	234
9.2	Installation und erste Schritte	234
9.3	Univariate lineare Regression	236
9.3.1	Generalisierte lineare Modelle	237
9.3.1.1	Lambda-Suche	239
9.3.1.2	Grid-Suche	239
9.4	Entscheidungsbäume, Random Forests und Gradient Boosting	240
9.5	Neuronale Netze	245
9.6	Der Boston-Datensatz	249
9.6.1	AutoML	254
9.6.2	Explain	256
9.6.2.1	Residuenanalyse	257
9.6.2.2	Variablenwichtigkeit	257
9.6.2.3	Heatmap der Variablenwichtigkeit	258
9.6.2.4	Modell-Korrelation	258
9.6.2.5	Partielles Abhängigkeitsdiagramm	260
9.7	Iris	260
9.7.1	Stacked Ensemble	261
9.8	MNIST	266
9.8.1	Der Standarddatensatz	266
9.8.2	Bewertung	270
9.9	Aufgaben	271
10	Keras/Tensorflow	273
10.1	Einrichtung und Nutzung von Keras	274
10.1.1	Dimensionen und Tensoren	274
10.1.2	Normalisierung	276
10.2	Boston	276
10.2.1	Normalisierung	276
10.2.2	Modelldefinition	277
10.2.3	Compilierung	278
10.2.4	Fit	279
10.3	Diagnosemöglichkeiten und Optimierung	282
10.3.1	Eine kleine Regression	282
10.3.2	Speichern und Laden des Modells und der Gewichte	285
10.3.3	Auslesen des Modells	286
10.3.4	Callbacks	287

10.3.5	TensorBoard	287
10.4	Convolutional Networks	291
10.4.1	Faltung und Feature Learning	291
10.4.1.1	Diskrete 2D-Faltung	292
10.4.1.2	Aufbau von Mustern	294
10.4.1.3	Pooling Layers	296
10.4.2	Komposition der Layer	297
10.4.3	MNIST	298
10.4.4	ImageNet	300
10.5	Transferlernen	304
10.5.1	Modelldefinition	304
10.5.2	Datenbereitstellung	306
10.5.3	Augmentation	309
10.5.4	Lernvorgänge	311
10.6	Recurrent Networks	314
10.6.1	Simple Recurrent Networks	314
10.6.2	LSTM	316
10.6.3	Wettervorhersage	318
10.7	Aufgaben	326
Teil V Anhang		329
A Basiswissen Statistik		331
A.1	Beschreibende Statistik	331
A.1.1	Mittelwert und Median	331
A.1.2	Varianz	333
A.1.3	Momente höherer Ordnung	334
A.1.4	Histogramme, Summenhäufigkeiten und Quantile	334
A.1.5	Box-Plots	338
A.2	Schließende Statistik	338
A.2.1	Normalverteilung	339
A.2.2	t-(Student-)Verteilung	342
A.2.3	Chi-Quadrat-Verteilung	344
A.2.4	F-Verteilung	346
Literatur		349
Stichwortverzeichnis		361