

Vorwort	XI
1 Einführung	1
Der Aufstieg der Daten	1
Was ist Data Science?	1
Ein motivierendes Szenario: DataSciencester	3
2 Ein Crashkurs in Python	15
Grundlagen	15
Über die Grundlagen hinaus	28
Weiterführendes Material	37
3 Daten visualisieren	39
matplotlib	39
Balkendiagramme	41
Liniendiagramme	45
Scatterplots	46
Weiterführendes Material	49
4 Lineare Algebra	51
Vektoren	51
Matrizen	55
Weiterführendes Material	58
5 Statistik	59
Einen einzelnen Datensatz beschreiben	59

Korrelation	64
Das Simpson-Paradoxon	67
Weitere Fallstricke von Korrelationen	68
Korrelation und Kausalität	69
Weiterführendes Material	70
6 Wahrscheinlichkeit	71
Abhängigkeit und Unabhängigkeit	71
Bedingte Wahrscheinlichkeit	72
Der Satz von Bayes	74
Zufallsvariablen	75
Kontinuierliche Wahrscheinlichkeitsverteilungen	76
Die Normalverteilung	77
Der zentrale Grenzwertsatz	81
Weiterführendes Material	83
7 Hypothesen und Schlussfolgerungen	85
Testen statistischer Hypothesen	85
Beispiel: Münzwürfe	85
p-Werte	88
Konfidenzintervalle	90
P-Hacking	91
Beispiel: Durchführen eines A/B-Tests	92
Bayessche Inferenz	93
Weiterführendes Material	97
8 Die Gradientenmethode	99
Die Idee hinter der Gradientenmethode	99
Abschätzen des Gradienten	100
Den Gradienten verwenden	103
Auswahl der richtigen Schrittweite	104
Anwendungsbeispiel	104
Stochastische Gradientenmethode	106
Weiterführendes Material	107

9	Daten sammeln	109
	stdin und stdout	109
	Einlesen von Dateien	111
	Auslesen von Webseiten	114
	Verwenden von APIs	121
	Beispiel: Verwenden der Twitter-APIs	124
	Weiterführendes Material	127
10	Arbeiten mit Daten	129
	Erkunden Ihrer Daten	129
	Bereinigen und Umformen	135
	Manipulieren von Daten	137
	Umskalieren	141
	Hauptkomponentenanalyse	142
	Weiterführendes Material	148
11	Maschinelles Lernen	149
	Modellieren	149
	Was ist maschinelles Lernen?	150
	Overfitting und Underfitting	151
	Genauigkeit	153
	Der Kompromiss zwischen Bias und Varianz	156
	Extraktion und Auswahl von Eigenschaften	157
	Weiterführendes Material	159
12	k-Nächste-Nachbarn	161
	Das Modell	161
	Beispiel: bevorzugte Programmiersprachen	163
	Der Fluch der Dimensionalität	167
	Weiterführendes Material	173
13	Naive Bayes-Klassifikatoren	175
	Ein wirklich primitiver Spam-Filter	175
	Ein anspruchsvollerer Spam-Filter	176
	Implementierung	178

Testen des Modells	179
Weiterführendes Material	182
14 Einfache lineare Regression	183
Das Modell	183
Anwenden des Gradientenverfahrens	186
Maximum-Likelihood-Methode	187
Weiterführendes Material	187
15 Multiple Regression	189
Das Modell	189
Weitere Annahmen bei der Methode der kleinsten Quadrate	190
Anpassen des Modells	191
Interpretation des Modells	192
Anpassungsgüte	193
Exkurs: Bootstrapping	194
Standardfehler von Regressionskoeffizienten	195
Regularisierung	197
Weiterführendes Material	199
16 Logistische Regression	201
Die Aufgabe	201
Die logistische Funktion	204
Anwendung des Modells	206
Anpassungsgüte	207
Support Vector Machines	208
Weiterführendes Material	212
17 Entscheidungsbäume	213
Was ist ein Entscheidungsbaum?	213
Entropie	215
Die Entropie einer Partition	217
Einen Entscheidungsbaum erzeugen	218
Verallgemeinerung des Verfahrens	221
Random Forests	223
Weiterführendes Material	224

18 Neuronale Netzwerke	225
Perzeptrons	225
Feed-forward-Netze	227
Backpropagation	230
Beispiel: Bezwingen eines CAPTCHA	231
Weiterführendes Material	236
19 Clustering	237
Die Idee	237
Das Modell	238
Beispiel: Meetups	239
Die Auswahl von k	242
Beispiel: Clustern von Farben	243
Agglomeratives hierarchisches Clustering	245
Weiterführendes Material	250
20 Linguistische Datenverarbeitung	251
Wortwolken	251
N-Gramm-Modelle	253
Grammatiken	256
Exkurs: Gibbs-Sampling	258
Themenmodellierung	260
Weiterführendes Material	265
21 Graphenanalyse	267
Betweenness-Zentralität	267
Eigenvektor-Zentralität	272
Gerichtete Graphen und PageRank	276
Weiterführendes Material	279
22 Empfehlungssysteme	281
Manuelle Pflege	282
Empfehlen, was beliebt ist	282
Nutzerbasiertes kollaboratives Filtern	283
Gegenstands-basiertes kollaboratives Filtern	286
Weiterführendes Material	288

23 Datenbanken und SQL	289
CREATE TABLE und INSERT	289
UPDATE	291
DELETE	292
SELECT	292
GROUP BY	294
ORDER BY	296
JOIN	297
Subqueries	300
Indexstrukturen	300
Optimierung von Anfragen	301
NoSQL	301
Weiterführendes Material	302
24 MapReduce	303
Beispiel: Wörter zählen	303
Warum MapReduce?	305
MapReduce verallgemeinert	306
Beispiel: Statusmeldungen analysieren	307
Beispiel: Matrizenmultiplikation	308
Eine Randbemerkung: Combiners	310
Weiterführendes Material	310
25 Gehet hin und praktiziert Data Science	313
IPython	313
Mathematik	314
Nicht bei null starten	314
Finden Sie Daten	316
Data Science in der Praxis	317
Index	321