

Vorwort zur 2. Auflage	XI
Vorwort zur 1. Auflage	XV
1 Einführung	1
Der Aufstieg der Daten	1
Was ist Data Science?	1
Ein motivierendes Szenario: DataSciencester	3
Finden von Schlüsselpersonen	3
Data Scientists, die Sie kennen könnten	6
Gehälter und Erfahrung	8
Bezahlte Nutzerkonten	11
Interessante Themen	11
Weiter geht's!	13
2 Ein Crashkurs in Python	15
Zen und Python	15
Python installieren	16
Virtuelle Umgebungen	16
Formatieren durch Leerzeichen	18
Module	19
Funktionen	20
Strings	21
Exceptions	22
Listen	22
Tupel	23
Dictionaries	24
defaultdict	25
Counter	26
Sets	27

Kontrollfluss	27
Wahrheitswerte	28
Sortieren	29
List Comprehensions	30
Automatisierte Tests und assert	31
Objektorientierte Programmierung	31
Iterables und Generatoren	33
Zufall	35
Reguläre Ausdrücke	36
Funktionale Programmierung	37
zip und Entpacken von Argumenten	37
args und kwargs	37
Type Annotations	39
Wie man Type Annotations schreibt	41
Willkommen bei DataSciencester!	43
Weiterführendes Material	43
3 Daten visualisieren	45
matplotlib	45
Balkendiagramme	47
Liniendiagramme	50
Scatterplots	51
Weiterführendes Material	53
4 Lineare Algebra	55
Vektoren	55
Matrizen	59
Weiterführendes Material	62
5 Statistik	63
Einen einzelnen Datensatz beschreiben	63
Lagemaße	65
Streuung	67
Korrelation	68
Das Simpson-Paradoxon	71
Weitere Fallstricke von Korrelationen	72
Korrelation und Kausalität	73
Weiterführendes Material	74
6 Wahrscheinlichkeit	75
Abhängigkeit und Unabhängigkeit	75
Bedingte Wahrscheinlichkeit	76

Der Satz von Bayes	78
Zufallsvariablen	79
Kontinuierliche Wahrscheinlichkeitsverteilungen	80
Die Normalverteilung	81
Der zentrale Grenzwertsatz	84
Weiterführendes Material	86
7 Hypothesen und Schlussfolgerungen	87
Testen statistischer Hypothesen	87
Beispiel: Münzwürfe	87
p-Werte	90
Konfidenzintervalle	92
p-Hacking	93
Beispiel: Durchführen eines A/B-Tests	94
Bayessche Inferenz	96
Weiterführendes Material	99
8 Die Gradientenmethode	101
Die Idee hinter der Gradientenmethode	101
Abschätzen des Gradienten	102
Den Gradienten verwenden	105
Auswahl der richtigen Schrittweite	106
Mit der Gradientenmethode Modelle anpassen	106
Minibatch und stochastische Gradientenmethode	108
Weiterführendes Material	109
9 Daten sammeln	111
stdin und stdout	111
Einlesen von Dateien	113
Grundlagen von Textdateien	113
Dateien mit Feldtrennern	114
Auslesen von Webseiten	116
Parsen von HTML-Dokumenten	116
Beispiel: Den Kongress im Auge behalten	118
Verwenden von APIs	121
JSON und XML	121
Eine nicht authentifizierte API verwenden	122
APIs finden	123
Beispiel: Verwenden der Twitter-APIs	124
Zugriff auf die APIs erhalten	124
Weiterführendes Material	128

10	Arbeiten mit Daten	129
	Erkunden Ihrer Daten	129
	Erkunden eindimensionaler Daten	129
	Zwei Dimensionen	132
	Mehrere Dimensionen	133
	NamedTuples	134
	Datenklassen	136
	Bereinigen und Umformen	137
	Manipulieren von Daten	139
	Umskalieren	141
	Exkurs: tqdm	143
	Hauptkomponentenanalyse	144
	Weiterführendes Material	150
11	Maschinelles Lernen	151
	Modellieren	151
	Was ist maschinelles Lernen?	152
	Overfitting und Underfitting	153
	Genauigkeit	155
	Der Kompromiss zwischen Bias und Varianz	158
	Extraktion und Auswahl von Eigenschaften	159
	Weiterführendes Material	160
12	k-Nächste-Nachbarn	161
	Das Modell	161
	Beispiel: Der Iris-Datensatz	163
	Der Fluch der Dimensionalität	166
	Weiterführendes Material	170
13	Naive Bayes-Klassifikatoren	171
	Ein wirklich primitiver Spam-Filter	171
	Ein anspruchsvollerer Spam-Filter	172
	Implementierung	174
	Das Modell testen	176
	Das Modell verwenden	177
	Weiterführendes Material	180
14	Einfache lineare Regression	181
	Das Modell	181
	Anwenden des Gradientenverfahrens	185
	Maximum-Likelihood-Methode	186
	Weiterführendes Material	186

15 Multiple Regression	187
Das Modell	187
Weitere Annahmen bei der Methode der kleinsten Quadrate.	188
Anpassen des Modells	189
Interpretation des Modells	191
Anpassungsgüte	192
Exkurs: Bootstrapping	192
Standardfehler von Regressionskoeffizienten	194
Regularisierung	196
Weiterführendes Material	198
16 Logistische Regression	199
Die Aufgabe	199
Die logistische Funktion	202
Anwendung des Modells	204
Anpassungsgüte	205
Support Vector Machines	207
Weiterführendes Material	209
17 Entscheidungsbäume	211
Was ist ein Entscheidungsbaum?	211
Entropie	213
Die Entropie einer Partition	215
Einen Entscheidungsbaum erzeugen	216
Verallgemeinerung des Verfahrens	219
Random Forests	221
Weiterführendes Material	222
18 Neuronale Netzwerke	223
Perzeptrons	223
Feed-forward-Netze	226
Backpropagation	228
Beispiel: Fizz Buzz	231
Weiterführendes Material	234
19 Deep Learning	235
Der Tensor	235
Die Layer-Abstrahierung	238
Der lineare Layer	240
Neuronale Netzwerke als Abfolge von Layern	242
Verlust und Optimierung	243
Beispiel XOR überarbeitet	246

Andere Aktivierungsfunktionen	247
Beispiel: Fizz Buzz überarbeitet	248
Softmaxes und Kreuz-Entropie	249
Dropout	251
Beispiel: MNIST	252
Modelle sichern und laden	257
Weiterführendes Material.	258
20 Clustering	259
Die Idee.	259
Das Modell	260
Beispiel: Meet-ups.	262
Die Auswahl von k	264
Beispiel: Clustern von Farben	265
Agglomeratives hierarchisches Clustering	267
Weiterführendes Material.	272
21 Linguistische Datenverarbeitung	273
Wortwolken	273
N-Gramm-Sprachmodelle.	275
Grammatiken	278
Exkurs: Gibbs-Sampling.	280
Themenmodellierung	282
Wortvektoren	287
Rekurrente neuronale Netzwerke.	296
Beispiel: Ein RNN auf Zeichenebene verwenden	299
Weiterführendes Material.	302
22 Graphenanalyse	303
Betweenness-Zentralität	303
Eigenvektor-Zentralität.	308
Matrizenmultiplikation	308
Zentralität	310
Gerichtete Graphen und PageRank	312
Weiterführendes Material.	314
23 Empfehlungssysteme	315
Manuelle Pflege.	316
Empfehlen, was beliebt ist	316
Nutzerbasiertes kollaboratives Filtern	317

Gegenstands-basiertes kollaboratives Filtern	320
Matrixfaktorisierung	322
Weiterführendes Material	326
24 Datenbanken und SQL	327
CREATE TABLE und INSERT	327
UPDATE	330
DELETE	331
SELECT	332
GROUP BY	334
ORDER BY	337
JOIN	337
Subqueries	340
Indexstrukturen	340
Optimierung von Anfragen	341
NoSQL.	341
Weiterführendes Material	342
25 MapReduce	343
Beispiel: Wörter zählen	343
Warum MapReduce?	345
MapReduce verallgemeinert	346
Beispiel: Statusmeldungen analysieren	347
Beispiel: Matrizenmultiplikation	349
Eine Randbemerkung: Combiners	351
Weiterführendes Material	351
26 Datenethik	353
Was ist Datenethik?	353
Jetzt aber wirklich: Was ist Datenethik?	354
Sollte ich mir über Datenethik Gedanken machen?	354
Schlechte Produkte bauen	355
Genauigkeit und Fairness abwägen	356
Zusammenarbeit	357
Interpretierbarkeit	358
Empfehlungen	358
Tendenziöse Daten	359
Datenschutz.	360
Zusammenfassung	361
Weiterführendes Material	361

27 Gehet hin und praktiziert Data Science	363
IPython	363
Mathematik	364
Nicht bei null starten	364
NumPy	364
pandas	364
scikit-learn	365
Visualisierung	365
R	366
Deep Learning	366
Finden Sie Daten	366
Data Science in der Praxis	367
Hacker News	367
Feuerwehrautos	367
T-Shirts	368
Tweets on a Globe	368
Und Sie?	369
Index	371