
Inhaltsverzeichnis

1	Big Data und Data Science	1
1.1	Einführung in Big Data	1
1.1.1	Volumen	2
1.1.2	Velocity/Geschwindigkeit	3
1.1.3	Variety/Vielfalt	4
1.1.4	Innovative Informationsverarbeitungsmethoden	6
1.1.5	Wissen generieren, Entscheidungen treffen, Prozesse automatisieren	7
1.1.6	Noch mehr Vs	7
1.2	Einführung in Data Science	7
1.2.1	Was gehört zu Data Science?	8
1.2.2	Beispielanwendungen	10
1.3	Fähigkeiten von Data Scientists	11
2	Der Prozess von Data-Science-Projekten	13
2.1	Der generische Data-Science-Prozess	14
2.1.1	Discovery	15
2.1.2	Datenvorbereitung	18
2.1.3	Modellplanung	21
2.1.4	Modellerstellung	24
2.1.5	Kommunikation der Ergebnisse	25
2.1.6	Operationalisierung	25
2.2	Rollen in Data-Science-Projekten	26
2.2.1	Anwenderin	27
2.2.2	Projektsponsorin	27
2.2.3	Projektmanagerin	28
2.2.4	Dateningenieurin	28
2.2.5	Datenbankadministratorin	29
2.2.6	Data Scientist	29

2.3	Deliverables	29
2.3.1	Sponsorenpräsentation	30
2.3.2	Analystenpräsentation	30
2.3.3	Quelltext	31
2.3.4	Technische Spezifikation	31
2.3.5	Daten	31
3	Allgemeines zur Datenanalyse	33
3.1	Das No-free-Lunch-Theorem	33
3.2	Definition von maschinellem Lernen	34
3.3	Merkmale	35
3.4	Trainings- und Testdaten	38
3.5	Kategorien von Algorithmen	41
3.6	Übung	42
4	Erkunden der Daten	45
4.1	Texteditoren und die Kommandozeile	45
4.2	Deskriptive Statistik	47
4.2.1	Lagemaße	47
4.2.2	Variabilität	50
4.2.3	Datenbereich	52
4.3	Visualisierung	53
4.3.1	Anscombes Quartett	55
4.3.2	Einzelne Merkmale	57
4.3.3	Beziehungen zwischen Merkmalen	69
4.3.4	Scatterplots für hochdimensionale Daten	77
4.3.5	Zeitliche Trends	79
4.4	Übung	82
5	Assoziationsregeln	83
5.1	Der Apriori-Algorithmus	85
5.1.1	Support und Frequent Itemsets	85
5.1.2	Ableiten von Regeln	87
5.1.3	Confidence, Lift und Leverage	87
5.1.4	Exponentielles Wachstum	90
5.1.5	Die Apriori-Eigenschaft	91
5.1.6	Einschränkungen für Regeln	93
5.2	Bewertung von Assoziationsregeln	93
5.3	Übung	94

6	Clusteranalyse	95
6.1	Ähnlichkeitsmaße	96
6.2	Städte und Häuser	98
6.3	k -Means-Algorithmus	98
6.3.1	Der Algorithmus	100
6.3.2	Bestimmen von k	102
6.3.3	Probleme des k -Means-Algorithmus	106
6.4	EM-Clustering	107
6.4.1	Der Algorithmus	110
6.4.2	Bestimmen von k	113
6.4.3	Probleme des EM-Clustering	117
6.5	DBSCAN	118
6.5.1	Der Algorithmus	119
6.5.2	Bestimmen von ε und $minPts$	123
6.5.3	Probleme bei DBSCAN	127
6.6	Single Linkage Clustering	128
6.6.1	Der SLINK-Algorithmus	129
6.6.2	Dendrogramme	130
6.6.3	Probleme bei SLINK	132
6.7	Vergleich der Algorithmen	134
6.7.1	Clusterformen	134
6.7.2	Anzahl der Cluster	137
6.7.3	Ausführungszeit	137
6.7.4	Interpretierbarkeit und Darstellung	139
6.7.5	Kategorische Merkmale	139
6.7.6	Fehlende Merkmale	139
6.7.7	Korrelierte Merkmale	140
6.7.8	Zusammenfassung des Vergleichs	140
6.8	Übung	141
7	Klassifikation	143
7.1	Binäre Klassifikation und Grenzwerte	145
7.2	Gütemaße	148
7.2.1	Die Confusion Matrix	148
7.2.2	Die binäre Confusion Matrix	149
7.2.3	Binäre Gütemaße	150
7.2.4	Die Receiver Operator Characteristic (ROC)	152
7.2.5	Area Under the Curve (AUC)	154
7.2.6	Micro und Macro Averages	156
7.2.7	Jenseits der Confusion Matrix	157

7.3	Decision Surfaces	158
7.4	k -Nearest Neighbor	160
7.5	Entscheidungsbäume	164
7.6	Random Forests	169
7.7	Logistische Regression	174
7.8	Naive Bayes	177
7.9	Support Vector Machines (SVMs)	179
7.10	Neuronale Netzwerke	185
	7.10.1 Exkurs: CNNs zum Erkennen von Zahlen	192
7.11	Vergleich der Klassifikationsalgorithmen	197
	7.11.1 Grundidee	197
	7.11.2 Decision Surfaces	197
	7.11.3 Ausführungszeit	206
	7.11.4 Interpretierbarkeit und Darstellung	209
	7.11.5 Scoring	209
	7.11.6 Kategorische Merkmale	210
	7.11.7 Fehlende Merkmale	210
	7.11.8 Korrelierte Merkmale	210
	7.11.9 Zusammenfassung des Vergleichs	211
7.12	Übung	212
8	Regression	215
8.1	Güte von Regressionen	216
	8.1.1 Visuelle Bewertung der Güte	218
	8.1.2 Gütemaße	221
8.2	Lineare Regression	223
	8.2.1 Ordinary Least Squares (OLS)	224
	8.2.2 Ridge	225
	8.2.3 Lasso	225
	8.2.4 Elastic Net	226
	8.2.5 Auswirkung der Regularisierung	226
8.3	Jenseits von linearer Regression	231
8.4	Übung	231
9	Zeitreihenanalyse	233
9.1	Box-Jenkins-Verfahren	235
9.2	Trends und saisonale Effekte	236
	9.2.1 Regression und das saisonale Mittel	236
	9.2.2 Differencing	240
	9.2.3 Vergleich der Ansätze	242

9.3	Autokorrelationen mit ARMA	242
9.3.1	Autokorrelation und partielle Autokorrelation	242
9.3.2	AR, MA und ARMA	246
9.3.3	Auswahl von p und q	247
9.3.4	ARIMA	248
9.4	Jenseits von Box-Jenkins	249
9.5	Übung	249
10	Text Mining	251
10.1	Preprocessing	253
10.1.1	Erstellung eines Korpus	253
10.1.2	Relevanter Inhalt	253
10.1.3	Zeichensetzung und Großschreibung	255
10.1.4	Stoppwörter	256
10.1.5	Stemming und Lemmatisierung	257
10.1.6	Visualisierung des Preprocessing	259
10.1.7	Bag-of-Words	261
10.1.8	Inverse Document Frequency	262
10.1.9	Jenseits des Bag-of-Words	264
10.2	Herausforderungen des Text Mining	265
10.2.1	Dimensionalität	265
10.2.2	Mehrdeutigkeiten	265
10.2.3	Weitere Probleme	266
10.3	Übung	266
11	Statistik	267
11.1	Hypothesentests	268
11.1.1	t -Test	269
11.1.2	Das Signifikanzniveau	272
11.1.3	Wichtige Hypothesentests	272
11.1.4	Anwendung der Tests	274
11.1.5	Übliche Fehler bei Hypothesentests	275
11.2	Effektstärke	277
11.3	Konfidenzintervalle	279
11.4	Gute Beschreibung von Ergebnissen	282
11.5	Übung	283
12	Big Data Processing	285
12.1	Parallelisierung	285
12.2	Verteiltes Rechnen zur Datenanalyse	286
12.3	Datenlokalität	288

12.4	MapReduce	288
12.4.1	map()	289
12.4.2	shuffle()	290
12.4.3	reduce()	290
12.4.4	Worthäufigkeiten mit MapReduce	290
12.4.5	Parallelisierung	291
12.5	Apache Hadoop	292
12.5.1	HDFS	293
12.5.2	YARN	295
12.5.3	MapReduce mit Hadoop	297
12.5.4	Streaming Mode	302
12.5.5	Weitere Komponenten von Hadoop	304
12.5.6	Grenzen von Hadoop	305
12.6	Apache Spark	305
12.6.1	Architektur	305
12.6.2	Datenstrukturen	306
12.6.3	Infrastruktur	307
12.6.4	Worthäufigkeiten mit Spark	307
12.7	Jenseits von Hadoop und Spark	308
13	Weiterführende Konzepte	309
Anhang		311
<hr/>		
A	Selbst ausführen	313
B	Notationen	315
C	Abkürzungen	319
D	Literatur	321
	Index	323