

Vorwort	IX
1 Die Machine-Learning-Pipeline	1
Daten	1
Aufgaben	1
Modelle	2
Merkmale	3
Modellbewertung	3
2 Trickserien mit einfachen Zahlen	5
Skalare, Vektoren und Räume	7
Der Umgang mit Zählern	8
Binarisierung	9
Quantisierung oder Klasseneinteilung	10
Die Logarithmustransformation	15
Die Logarithmustransformation am Werk	19
Potenztransformationen als verallgemeinerte Logarithmustransformation	23
Merkmalskalierung oder -normierung	28
Min-Max-Skalierung	28
Standardisierung (Varianzskalierung)	29
ℓ^2 -Normierung	30
Kreuzmerkmale	33
Merkmalsauswahl	35
Zusammenfassung	36
Literatur	37
3 Textdaten: Einebnen, Filtern und Wortgruppensuche	39
Bag-of-X: von natürlichem Text zu flachen Vektoren	40
Bag-of-Words	40
Bag-of-n-Grams	43

Reinere Merkmale durch Filtern	46
Stoppwörter	46
Filtern nach Häufigkeit	46
Stemming	49
Bedeutungseinheiten: von Wörtern über n-Gramme zu Phrasen	50
Parsen und Tokenbildung	51
Anordnungsanalyse zur Phrasenerkennung	51
Zusammenfassung	58
Literatur	59
4 Auswirkungen der Merkmalskalierung: von Bag-of-Words zu TF-IDF	61
TF-IDF: eine kleine Variation von Bag-of-Words	61
Ein Praxistest	63
Erzeugung eines Klassifikationsdatensatzes	64
Bag-of-Words skalieren mit der TF-IDF-Transformation	65
Klassifikation mittels logistischer Regression	66
Abstimmen der logistischen Regression durch Regularisierung	67
Der Sache auf den Grund gegangen: Was geht hier vor?	72
Zusammenfassung	75
Literatur	76
5 Kategoriale Variablen: Eier zählen im Roboterzeitalter	77
Kodierung kategorialer Variablen	78
Die One-Hot-Kodierung	78
Die Dummy-Kodierung	79
Die Wirkungskodierung	81
Vor- und Nachteile der Kodierungen kategorialer Variablen	82
Große kategoriale Variablen	83
Merkmals-Hashing	84
Klassenzählung	87
Zusammenfassung	95
Literatur	96
6 Dimensionsreduktion: Mit dem Hauptkomponentenverfahren die Datenwolke flach drücken	99
Die Grundidee	99
Herleitung	101
Lineare Projektion	102
Varianz und empirische Varianz	103
Hauptkomponenten: Erste Schreibweise	104
Hauptkomponenten: Matrix-Vektor-Schreibweise	104
Allgemeine Lösung für die Hauptkomponenten	104

Transformation der Merkmale	105
Implementierung des Hauptkomponentenverfahrens	105
Das Hauptkomponentenverfahren am Werk	106
Weißes und Nullphasenverfahren	108
Bedingungen und Grenzen des Hauptkomponentenverfahrens	109
Anwendungsfälle	111
Zusammenfassung	113
Literatur	114
7 Nichtlineare Merkmalsgewinnung mittels k-Means-Modellstapelung	115
Clustern mit k-Means	117
Clustern als Flächenzerlegung	119
Merkmalsgewinnung mit k-Means zur Klassifikation	122
Die Alternative: Merkmalsgewinnung aus dicht besetzten Daten	127
Vorteile, Nachteile und Stolperfallen	128
Zusammenfassung	131
Literatur	131
8 Automatisierte Merkmalsgewinnung: Bildmerkmale und Deep Learning	133
Die einfachsten Bildmerkmale (und der Grund, warum sie nicht funktionieren)	134
Manuelle Merkmalsgewinnung: SIFT und HOG	135
Bildgradienten	135
Histogramme von Gradientenrichtungen	139
Die Architektur des SIFT-Verfahrens	143
Erlernen von Bildmerkmalen mit tiefen neuronalen Netzen	144
Vollständig verbundene Schichten	144
Konvolutionsschichten	146
Der lineare Gleichrichter (ReLU)	150
Antwortnormierungsschichten	151
Pooling-Schichten	152
Struktur von AlexNet	153
Zusammenfassung	156
Literatur	157
9 Die fabelhafte Welt der Merkmale: ein Empfehlungsalgorithmus für akademische Aufsätze	159
Artikelbezogenes kollaboratives Filtern	159
Erster Durchgang: Datenimport, Säuberung und Merkmalsgewinnung	161
Empfehlungsalgorithmus für akademische Aufsätze: naiver Ansatz	161

Zweiter Durchgang: Mehr Konstruktion und ein intelligenteres Modell	167
Empfehlungsalgorithmus für akademische Aufsätze:	
zweiter Anlauf	167
Dritter Durchgang: Mehr Merkmale bedeuten mehr Information	172
Empfehlungsalgorithmus für akademische Aufsätze:	
dritter Anlauf	173
Zusammenfassung	175
Literatur	176
Anhang: Lineare Modellierung und Grundlagen der linearen Algebra	177
Index	191