

Inhalt

Einleitung	XIII
TEIL I: Grundlagen des Unsupervised Learning	1
1 Unsupervised Learning im Ökosystem des maschinellen Lernens	3
Grundbegriffe des maschinellen Lernens	3
Regelbasiertes vs. maschinelles Lernen	4
Supervised vs. Unsupervised	5
Die Stärken und Schwächen des Supervised Learning	6
Die Stärken und Schwächen des Unsupervised Learning	7
Lösungen mit maschinellern Lernen durch Unsupervised Learning verbessern	8
Ein genauerer Blick auf überwachte Algorithmen	11
Lineare Methoden	12
Nachbarschaftsbasierte Methoden	13
Baumbasierte Methoden	14
Support Vector Machines	16
Neuronale Netze	16
Unüberwachte Algorithmen unter der Lupe	17
Reduzierung der Dimensionalität	17
Clustering	20
Feature Extraction	22
Unsupervised Deep Learning	23
Sequenzielle Datenprobleme beim Unsupervised Learning	25
Reinforcement Learning mithilfe von Unsupervised Learning	26
Semi-supervised Learning	27
Erfolgreiche Anwendungen von Unsupervised Learning	27
Anomalieerkennung	28
Zusammenfassung	29

2	Durchgehendes Projekt zum maschinellen Lernen	31
	Die Umgebung einrichten	31
	Versionsverwaltung: Git	31
	Das Git-Repository »handson-unsupervised-learning« klonen	31
	Wissenschaftliche Bibliotheken: Anaconda-Distribution von Python	32
	Neuronale Netze: TensorFlow und Keras	32
	Gradient Boosting, Version 1: XGBoost	33
	Gradient Boosting, Version 2: LightGBM	33
	Clustering-Algorithmen	33
	Interaktive Computerumgebung: Jupyter Notebook	34
	Die Daten im Überblick	34
	Datenvorbereitung	35
	Datenerfassung	35
	Daten erkunden	36
	Featurematrix und Labels-Array generieren	39
	Feature Engineering und Feature Selection	40
	Daten visualisieren	41
	Modellvorbereitung	42
	In Trainings- und Testsets aufteilen	42
	Eine Kostenfunktion auswählen	43
	Sets für k-fache Kreuzvalidierung erzeugen	43
	Modelle des maschinellen Lernens (Teil 1)	44
	Modell #1: Logistische Regression	44
	Kennzahlen bewerten	47
	Wahrheitsmatrix	48
	Präzision/Trefferquote-Diagramm	49
	Operationscharakteristik eines Beobachters	51
	Modelle des maschinellen Lernens (Teil 2)	53
	Modell #2: Random Forests	53
	Modell #3: Gradient Boosting mit XGBoost	56
	Modell #4: Gradient Boosting mit LightGBM	59
	Bewertung der vier Modelle mit dem Testset	62
	Ensembles	66
	Stapeln	66
	Endgültige Modellauswahl	69
	Produktionspipeline	70
	Zusammenfassung	71

TEIL II: Unsupervised Learning mit Scikit-learn	73
3 Dimensionsreduktion	75
Die Motivation zur Dimensionsreduktion	75
Die MNIST-Zifferndatenbank	76
Algorithmen zur Dimensionsreduktion	79
Lineare Projektion vs. Manifold Learning	80
Hauptkomponentenanalyse	80
Hauptkomponentenanalyse, das Konzept	80
PCA in der Praxis	81
Sparse PCA	87
Kernel-PCA	88
Singularwertzerlegung	89
Zufallsprojektion	91
Gaußsche Zufallsprojektion	91
Sparse Zufallsprojektion	92
Isomap	93
Multidimensionale Skalierung	94
Lokal lineare Einbettung	95
Stochastische Nachbarschaftseinbettung mit Student-t-Verteilung	96
Andere Methoden zur Dimensionsreduktion	98
Dictionary Learning	98
Unabhängigkeitsanalyse	99
Zusammenfassung	100
4 Anomalieerkennung	103
Kreditkartenbetrugserkennung	104
Die Daten vorbereiten	104
Anomalie-Score-Funktion definieren	104
Bewertungskennzahlen definieren	105
Eine Diagrammfunktion definieren	107
Anomalieerkennung mit normaler PCA	107
PCA-Komponenten gleich der Anzahl ursprünglicher Dimensionen	108
Suche nach der optimalen Anzahl von Hauptkomponenten	110
Anomalieerkennung mit sparse PCA	112
Anomalieerkennung mit Kernel-PCA	115
Anomalieerkennung mit gaußscher Zufallsprojektion	117
Anomalieerkennung mit sparse Zufallsprojektion	119
Nicht lineare Anomalieerkennung	120

Anomalieerkennung mit Dictionary Learning	121
Anomalieerkennung mit Unabhängigkeitsanalyse	123
Betrugserkennung auf dem Testset.	124
Anomalieerkennung mit normaler PCA auf dem Testset	125
Anomalieerkennung auf dem Testset mit Unabhängigkeitsanalyse.	126
Anomalieerkennung mit Dictionary Learning auf dem Testset	128
Zusammenfassung	129
5 Clustering	131
Das MNIST-Ziffern-Dataset	132
Datenvorbereitung	132
Clustering-Algorithmen	133
k-Means-Algorithmus	134
k-Means-Trägheit.	134
Die Clustering-Ergebnisse bewerten	135
k-Means-Genauigkeit.	137
k-Means und die Anzahl der Hauptkomponenten	138
k-Means auf dem ursprünglichen Dataset	140
Hierarchisches Clustering.	141
Agglomeratives hierarchisches Clustering	142
Das Dendrogramm.	143
Die Clustering-Ergebnisse auswerten.	145
DBSCAN.	147
DBSCAN-Algorithmus.	148
DBSCAN auf unser Dataset anwenden	148
HDBSCAN.	150
Zusammenfassung	151
6 Gruppensegmentierung	153
Lending-Club-Daten.	153
Datenvorbereitung	154
String-Format in numerisches Format überführen	155
Fehlende Werte imputieren	156
Den endgültigen Merkmalssatz auswählen und skalieren	158
Labels für die Bewertung benennen.	158
Güte der Cluster	160
k-Means-Anwendung	162
Anwendung mit hierarchischem Clustering.	164
Anwendung mit HDBSCAN.	168
Zusammenfassung	170

TEIL III: Unsupervised Learning mit TensorFlow und Keras	171
7 Autoencoder	173
Neuronale Netze	174
TensorFlow	175
Keras	177
Autoencoder: der Encoder und der Decoder	177
Untervollständige Autoencoder	178
Übervollständige Autoencoder	179
Dichte vs. sparsame Autoencoder	179
Autoencoder zur Rauschunterdrückung	180
Variational Autoencoder	180
Zusammenfassung	181
8 Praktische Autoencoder	183
Datenvorbereitung	183
Die Bestandteile eines Autoencoders	186
Aktivierungsfunktionen	186
Unser erster Autoencoder	187
Verlustfunktion	188
Optimizer	188
Das Modell trainieren	189
Auf dem Testset bewerten	191
Zweischichtiger unternvollständiger Autoencoder mit linearer	
Aktivierungsfunktion	194
Die Anzahl der Knoten erhöhen	197
Mehr Hidden-Schichten hinzufügen	199
Nicht linearer Autoencoder	200
Übervollständiger Autoencoder mit linearer Aktivierung	202
Übervollständiger Autoencoder mit linearer Aktivierung und	
Drop-out	205
Sparse übervollständiger Autoencoder mit linearer Aktivierung	207
Sparse übervollständiger Autoencoder mit linearer Aktivierung und	
Drop-out	209
Mit verrauschten Datasets arbeiten	211
Rauschreduzierender Autoencoder	211
Zweischichtiger rauschreduzierender unternvollständiger	
Autoencoder mit linearer Aktivierung	212
Zweischichtiger rauschunterdrückender übervollständiger	
Autoencoder mit linearer Aktivierung	215

Zweischichtiger rauschunterdrückender übervollständiger	
Autoencoder mit ReLu-Aktivierung.	217
Zusammenfassung	219
9 Semi-supervised Learning	221
Datenvorbereitung	221
Supervised Modelle.	224
Unsupervised Modelle	226
Semi-supervised Modelle	228
Die Leistung von supervised und unsupervised Modellen	231
Zusammenfassung	231
TEIL IV: Deep Unsupervised Learning mit TensorFlow und Keras	233
10 Empfehlungsdienste mit beschränkten Boltzmann-Maschinen	235
Boltzmann-Maschinen	235
Beschränkte Boltzmann-Maschinen.	236
Empfehlungsdienste	237
Kollaboratives Filtern.	237
Der Netflix Prize.	238
MovieLens-Dataset	238
Datenvorbereitung.	238
Die Kostenfunktion definieren: mittlere quadratische	
Abweichung	242
Baseline-Experimente.	243
Matrixfaktorisierung.	244
Ein latenter Faktor	244
Drei latente Faktoren	246
Fünf latente Faktoren.	246
Kollaboratives Filtern mit RBMs.	247
Die Architektur des neuronalen Netzes von RBMs	248
Die Komponenten der RBM-Klasse erstellen.	249
Das RBM-Empfehlungssystem trainieren	251
Zusammenfassung	253
11 Featureerkennung mit Deep Belief Networks	255
Deep Belief Networks im Detail	255
MNIST-Bildklassifizierung	256
Beschränkte Boltzmann-Maschinen	257
Die Komponenten der RBM-Klasse erstellen.	258
Mit dem RBM-Modell Bilder generieren	260

Die Featuredetektoren der Zwischenstufen anzeigen	261
Die drei RBMs für das DBN trainieren	262
Featuredetektoren untersuchen	264
Generierte Bilder betrachten	264
Das vollständige DBN	267
Wie das Training eines DBN funktioniert	271
Das DBN trainieren	271
Wie Unsupervised Learning das Supervised Learning unterstützt	272
Bilder generieren, um eine bessere Bildklassifizierung zu erstellen	273
Bildklassifizierung mit LightGBM	277
Rein supervised Lösung	277
Unsupervised und supervised Lösung	279
Zusammenfassung	280
12 Generative Adversarial Networks	281
GANs – das Konzept	281
Die Stärke von GANs	282
Deep Convolutional GANs	282
Convolutional Neural Networks	283
Noch einmal: DCGANs	287
Generator des DCGAN	288
Diskriminator des DCGAN	289
Diskriminator- und gegnerische Modelle	290
DCGAN für das MNIST-Dataset	291
MNIST-DCGAN in Aktion	292
Synthetische Bilder generieren	293
Zusammenfassung	294
13 Zeitreihen-Clustering	297
EKG-Daten	298
Ansatz für Zeitreihen-Clustering	298
k-Shape	298
Zeitreihen-Clustering mit k-Shape auf ECGFiveDays	299
Datenvorbereitung	299
Training und Bewertung	304
Zeitreihen-Clustering mit k-Shape auf ECG5000	305
Datenvorbereitung	305
Training und Bewertung	308
Zeitreihen-Clustering mit k-Means auf ECG5000	310
Zeitreihen-Clustering mit hierarchischem DBSCAN auf ECG5000	311
Die Zeitreihen-Clustering-Algorithmen vergleichen	312

Vollständiger Lauf mit k-Shape	312
Vollständiger Lauf mit k-Means.	314
Vollständiger Lauf mit HDBSCAN	315
Alle drei Zeitreihen-Clustering-Ansätze vergleichen	316
Zusammenfassung	318
14 Zum Schluss	319
Supervised Learning	320
Unsupervised Learning.	320
Scikit-learn	321
TensorFlow und Keras.	321
Reinforcement Learning	322
Die vielversprechendsten Bereiche des Unsupervised Learning	323
Die Zukunft des Unsupervised Learning	324
Schlusswort.	326
Index	327