

Inhaltsverzeichnis

1	Daten als Grundlage von Modellen	1
1.1	Datenbasierte Modellierung	2
1.1.1	Der Begriff Modell	2
1.1.2	White-Box-, Grey-Box- und Black-Box-Modelle	4
1.1.3	Kritik an der datenbasierten Modellierung	4
1.2	Klassifikation von Daten	5
1.2.1	Geschichtliche Einordnung des Begriffs Data	5
1.2.2	Daten in unserer heutigen Zeit	6
1.2.3	Strukturelle Klassifikation	6
1.2.4	Quantitative Kategorisierung von Daten	8
1.2.5	Qualitative Kategorisierung von Daten	8
1.2.6	Zeitreihen	10
1.2.7	Skalen und das Skalenniveau	11
1.3	Data Mining als systematischer Prozess	14
1.3.1	Was ist Data Mining?	14
1.3.2	Schritte, um Data Mining durchzuführen	16
1.3.3	Schlüsse aus Daten ziehen	17
1.3.4	Industrielles Data Mining und der CRISP-DM	18
1.4	Praktischer Umgang mit Daten in Python	21
1.4.1	Programme in diesem Buch	21
1.4.2	Bibliotheken für Python	21
1.4.3	Erklärbare Daten in Python	22
	Literatur	25
2	Mathematische Beschreibung von Daten	27
2.1	Grundlagen der Stochastik	28
2.1.1	Wahrscheinlichkeit	28
2.1.2	Mengen	29
2.1.3	Wahrscheinlichkeitsdefinition nach Laplace	30
2.1.4	Ereignis-, Ergebnis- und Wahrscheinlichkeitsräume	31
2.1.5	Axiome von Kolmogorov	32

2.1.6	Bedingte Wahrscheinlichkeit und der Satz von Bayes	34
2.1.7	Stochastischer Prozess	36
2.1.8	Berücksichtigung der Wahrscheinlichkeitsdichte in den Daten	38
2.2	Stochastik von Daten	39
2.2.1	Stochastik des Abtastprozesses	39
2.2.2	Das Nyquist-Theorem	40
2.2.3	Unterscheidung von Unsicherheiten	42
2.3	Statistische Werkzeuge zum Umgang mit Daten	43
2.3.1	Vektorielle Darstellung einer einzelnen Datenreihe	43
2.3.2	Erwartungswert	43
2.3.3	Varianz	45
2.3.4	Empirische Varianz	46
2.3.5	Momente einer Wahrscheinlichkeitsverteilung	46
2.3.6	Ausgesuchte Wahrscheinlichkeitsverteilungen	47
2.3.7	Matrixdarstellung mehrerer Messreihen	49
2.3.8	Kovarianz und Kovarianzmatrix	50
2.3.9	Korrelation und Korrelationsmatrix	51
	Literatur	54
3	Datenvorverarbeitung	57
3.1	Ziele der Vorverarbeitung	58
3.2	Datenbereinigung	59
3.2.1	Entfernen von fehlerhaften Datenpunkten	59
3.2.2	Fehlende Daten	62
3.2.3	Aufspüren und Eliminieren von Duplikaten	62
3.3	Normalisierung	63
3.3.1	Gründe für die Normalisierung von Daten	63
3.3.2	Typen von Normalisierungen	63
3.3.3	Anwendung der Normalisierung	64
3.4	Filterung von Daten	68
3.4.1	Gleitender Mittelwert	68
3.4.2	Faltung	69
3.4.3	Der Median Filter	70
3.5	Triggerung	71
3.5.1	Zeitreihen und repetitive Daten	71
3.5.2	Implementation des Triggers	72
3.6	Transformationen	74
3.6.1	Differentiation der Daten	74
3.6.2	Funktionale Transformation	76
3.6.3	Fourier-Transformation	77
3.6.4	Wavelet-Transformation	81

3.7	Quantifikation der stochastischen Eigenschaften.	84
3.7.1	Histogramme	84
3.7.2	Identifikation der Wahrscheinlichkeitsverteilung mittels Kerndichteschätzer.	85
3.8	Kenngößen aus Daten ermitteln	88
3.8.1	Statistische Kenngößen.	88
3.8.2	Prozessorientierte Kenngößen	88
	Literatur.	91
4	Überwachtes Lernen	93
4.1	Lernen	93
4.1.1	Was bedeutet überhaupt Lernen?	93
4.1.2	Verschiedene Formen des maschinellen Lernens.	95
4.2	Strategie des überwachten Lernens	95
4.2.1	Algorithmische Vorüberlegung	95
4.2.2	Klassifikation und Regression	97
4.2.3	Trainingsdaten und Testdaten	98
4.3	Evolutionäres Lernen	99
4.3.1	Idee der evolutionären Veränderung	99
4.3.2	Implementation eines einfachen, evolutionären Algorithmus	100
4.3.3	Zwischenspeichern der besten Veränderung	101
4.4	LMS-adaptiver Filter	103
4.4.1	Theorie des LMS-Algorithmus für das Lernen auf einen Zielwert	103
4.4.2	Implementation des LMS-Algorithmus.	105
4.4.3	NLMS-Algorithmus durch Normalisierung der Eingangsvariablen	106
4.4.4	Training	106
4.4.5	Anwendungen und Eigenschaften des adaptiven Filters.	107
4.4.6	Hyperparameter des LMS-adaptiven Filters	107
4.5	Neuronale Netze.	108
4.5.1	Das Neuron	108
4.5.2	Vernetzung von Neuronen	109
4.5.3	Aktivierungsfunktionen	111
4.5.4	Training neuronaler Netze	113
4.5.5	Optimierung als Hyperparameter.	114
4.5.6	Einfaches Klassifikationsnetz mit Scikit-Learn.	115
4.5.7	Klassifikationsnetz in Keras.	116
4.5.8	Regressionsnetz in Keras	119
4.5.9	Überanpassung und Kreuzvalidierung.	122
4.6	Rekurrente Neuronale Netze	124
4.6.1	Rückkopplung in neuronalen Netzstrukturen	124

4.6.2	Gates zur Steuerung des Informationsflusses innerhalb des Neurons	126
4.6.3	Neuronales Netz mit langem Kurzzeitgedächtnis (LSTM)	127
4.6.4	Implementation eines LSTM-Netzes	128
4.7	Entscheidungsbäume	131
4.7.1	Grundlegende Idee des Entscheidungsbaums	131
4.7.2	Entropie und Information Gain	133
4.7.3	Klassifikations- und Regressionsbäume	139
4.7.4	Anwendungsbeispiel mit Scikit-Learn	140
4.7.5	Implementation eines einfachen Entscheidungsbaums	140
4.7.6	Der Gini-Koeffizient für Entscheidungsbäume	145
	Literatur.	148
5	Unüberwachtes Lernen.	151
5.1	Unüberwachte Lernparadigmen.	152
5.2	Hauptkomponentenanalyse, Principal Component Analysis (PCA)	153
5.2.1	Eigenschaften der Kovarianzmatrix	153
5.2.2	Eigenraum	154
5.2.3	Eigenraum der Kovarianzmatrix	155
5.2.4	Eigenraumreduktion.	156
5.2.5	Implementation über Scikit-Learn	157
5.2.6	Diskussion der PCA	159
5.3	K-Means-Clusterverfahren.	161
5.3.1	Finden von Clusterzentren	161
5.3.2	Implementation mit Scikit-Learn	162
5.3.3	Batch K-Means	163
5.4	Der t-Distributed-Stochastic-Neighbour-Embedding- Algorithmus (t-SNE)	164
5.4.1	Idee hinter t-SNE	164
5.4.2	Anwendung der t-SNE Implementation in Scikit-Learn	165
5.4.3	Vor- und Nachteile von t-SNE	167
5.5	Autoencoder	168
5.5.1	Topologie eines Autoencoders	168
5.5.2	Latenter Raum	169
5.5.3	Anomalieerkennung	170
5.5.4	Implementation in Keras	170
5.5.5	Klassifizierung im latenten Raum	173
5.5.6	Unsicherheit im AE	175
5.5.7	Erkennung einer unbekannten Anomalie	175
5.5.8	Kombination von Verfahren	180
	Literatur.	182

6	Physikalisch-informiertes Lernen	185
6.1	Einführung	185
6.1.1	Was ist physikalisch-informiertes Lernen?	186
6.1.2	Ein einfaches, motivierendes Beispiel	187
6.1.3	Kontinuierliche Veränderung in der Prozesskette	189
6.1.4	Statistische Balance	190
6.1.5	Gründe für physikalisch-informiertes, maschinelles Lernen.	190
6.2	Datenanreicherung	191
6.2.1	Optimierte Wahl der Vorverarbeitung	191
6.2.2	Expertenwissen über Datenobjekte	192
6.2.3	Automatische Optimierung der Datenanreicherung bei binärer Klassifizierung.	193
6.3	Einbettung von analytischen Ausdrücken in Neuronale Netze	194
6.3.1	Automatische Ableitung	195
6.3.2	Einbinden einer Funktion als Optimierungsziel	196
6.3.3	Integration einer Funktion	198
6.3.4	Integration einer gewöhnlichen Differentialgleichung erster Ordnung	202
6.4	Stochastische Methoden zur Integration von Unsicherheit in den Lernprozess	204
6.4.1	Berücksichtigung von Unsicherheit in den Eingangsdaten	204
6.4.2	Quantifikation der Unsicherheit aus den gelernten Ergebnissen	207
6.4.3	Stochastische Anreicherung von Zielvariablen	209
6.4.4	Mixture-Density-Netze (MDN)	212
6.5	Prozesskorridore	216
6.5.1	Wahrscheinlichkeitsdichten durch 2D-Histogramme	216
	Literatur	222
7	Erklärbarkeit	223
7.1	Semantische Ordnungsstrukturen zur Digitalisierung von Bedeutung	223
7.1.1	Semantik	224
7.1.2	Übersicht über die semantischen Konzepte	225
7.1.3	Alphabete und Wörter	226
7.1.4	Vokabular	227
7.1.5	Synonymer Ring	227
7.1.6	Taxonomien	229
7.1.7	Ontologien	232
7.2	Sensitivitätsanalyse	235
7.2.1	Störungstheoretischer Ansatz	236
7.2.2	Störung eines Entscheidungsbaum	237

7.3	Erklärbares maschinelles Lernen	238
7.3.1	Ebenen der Erklärbarkeit	239
7.3.2	Praktische Umsetzung von Erklärbarkeit.	241
7.3.3	Kausalität in der technischen Prozesskette	245
7.3.4	Geschäftsverständnis	247
	Literatur.	248
	Anhang A: Grundlagen in Python.	249
	Stichwortverzeichnis.	257