

Inhaltsverzeichnis

Einführung	1
-------------------------	---

Teil I Grundlagen

1 Elemente der Datenorganisation	11
1.1 Konzeptionelle Datenmodellierung	12
1.1.1 Entity-Relationship-Modell	12
1.2 Logische Datenmodellierung	14
1.2.1 Relationales Datenmodell	14
1.2.2 Graphbasierte Datenmodelle	17
1.2.3 Hierarchische Datenmodelle	20
1.3 Datenqualität	22
1.3.1 Datenqualitätsmerkmale	23
1.4 Datenbereinigung	24
1.4.1 Validierung	24
1.4.2 Normierung	25
1.4.3 Imputation	26
1.4.4 Augmentation	28
1.4.5 Deduplikation	28
Quellen	35
2 Deskriptive Statistik	37
2.1 Stichprobe und Merkmale	38
2.2 Diagramme	40
2.2.1 Säulendiagramme und Histogramme	40
2.2.2 Streudiagramme	42
2.2.3 Weitere Diagramme	44
2.3 Lageparameter	45
2.3.1 Arithmetisches Mittel und empirischer Median	48
2.3.2 Quantile	50
2.3.3 Geometrisches und harmonisches Mittel	51

2.4	Streuungsparameter	53
2.4.1	Abweichung von Mittelwert oder Median	53
2.4.2	Shannon-Entropie	54
2.5	Assoziationsparameter	56
2.5.1	Empirische Kovarianz und Korrelation	57
2.5.2	Rangkorrelationskoeffizienten	59
2.5.3	Transinformation und Jaccard-Koeffizient	62
	Quellen	66

Teil II Stochastik

3	Wahrscheinlichkeitstheorie	69
3.1	Wahrscheinlichkeitsmaße	70
3.1.1	Bedingte Wahrscheinlichkeit	74
3.1.2	Der Satz von Bayes	78
3.2	Zufallsvariablen	81
3.2.1	Diskrete und stetige Zufallsvariablen	82
3.2.2	Massen- und Dichtefunktionen	85
3.2.3	Transformation von Zufallsvariablen	88
3.3	Gemeinsame Verteilung von Zufallsvariablen	91
3.3.1	Gemeinsame Verteilungs-, Masse- und Dichtefunktionen ..	91
3.3.2	Bedingte Masse- und Dichtefunktionen	93
3.3.3	Unabhängige Zufallsvariablen	94
3.4	Kennzahlen von Zufallsvariablen	96
3.4.1	Median, Erwartungswert und Varianz	96
3.4.2	Kovarianz und Korrelation	101
3.4.3	Die Tschebyscheff'sche Ungleichung	104
3.5	Summen und Produkte von Zufallsvariablen	106
3.5.1	Chi-Quadrat- und Student'sche t-Verteilung	110
	Quellen	113
4	Inferenzstatistik	115
4.1	Statistische Modelle	116
4.1.1	Modelle diskreter Zufallsvariablen	116
4.1.2	Modelle stetiger Zufallsvariablen	120
4.2	Gesetze der großen Zahlen	123
4.2.1	Bernoulli'sches Gesetz der großen Zahlen	126
4.2.2	Tschebyscheff'sches Gesetz der großen Zahlen	128
4.2.3	Varianzschätzung und Bessel-Korrektur	130
4.2.4	Zentraler Grenzwertsatz von Lindeberg-Lévy	132
4.3	Statistische Schätz- und Testverfahren	133
4.3.1	Intervallschätzung	133
4.3.2	Gauß-Test	137
4.3.3	Student'sche Vertrauensintervalle	139
4.3.4	Effektstärke	142
4.4	Parameter- und Dichteschätzung	143

4.4.1	Maximum-Likelihood-Schätzung	145
4.4.2	Bayes'sche Parameterschätzung	150
4.4.3	Kerndichteschätzung	154
4.5	Regressionsanalyse	156
4.5.1	Einfache lineare Regression	156
4.5.2	Theil-Sen-Verfahren	161
4.5.3	Einfache logistische Regression	162
	Quellen	164
5	Multivariate Statistik	165
5.1	Datenmatrizen	165
5.2	Abstands- und Ähnlichkeitsmaße	167
5.2.1	Metrische Abstands- und Ähnlichkeitsmaße	167
5.2.2	Kategoriale und binäre Abstands- und Ähnlichkeitsmaße	170
5.2.3	Abstands- und Ähnlichkeitsmatrizen	172
5.3	Multivariate Lage- und Streuungsparameter	175
5.3.1	Geometrischer Schwerpunkt und Median, Medoid	176
5.3.2	Empirische Kovarianz- und Korrelationsmatrix	178
5.4	Zufallsvektoren und -matrizen	179
5.4.1	Erwartungswertvektor und Kovarianzmatrix	179
5.4.2	Multivariate Normalverteilung	181
5.4.3	Multinomialverteilung	184
	Quellen	186

Teil III Maschinelles Lernen

6	Überwachtes maschinelles Lernen	189
6.1	Elemente des überwachten Lernens	191
6.1.1	Verlustfunktionen und empirisches Risiko	194
6.1.2	Überanpassung und Unteranpassung	196
6.1.3	Training, Modellauswahl und Test	200
6.1.4	Numerische Optimierung	204
6.2	Regressionsverfahren	208
6.2.1	Lineare Regression	209
6.2.2	Gauß-Prozess-Regression	215
6.3	Klassifikationsverfahren	218
6.3.1	Logistische Regression	218
6.3.2	Nächste-Nachbarn-Klassifikation	223
6.3.3	Bayes'sche Klassifikationsverfahren	227
6.4	Künstliche neuronale Netzwerke	232
6.4.1	Regression und Klassifikation mittels neuronaler Netzwerke	235
6.4.2	Training neuronaler Netzwerke durch Fehlerrückführung	239
6.4.3	Convolutional Neural Networks	245
	Quellen	251

7	Unüberwachtes maschinelles Lernen	255
7.1	Elemente des unüberwachten Lernens	255
7.1.1	Intrinsische Dimension von Daten	256
7.1.2	Topologische Merkmale von Daten	257
7.2	Dimensionsreduktion	259
7.2.1	Hauptkomponentenanalyse	261
7.2.2	Autoencoder	264
7.2.3	Multidimensionale Skalierung	265
7.2.4	T-distributed Stochastic Neighbor Embedding (t-SNE)	269
7.3	Clusteranalyse	272
7.3.1	K-Means-Verfahren	272
7.3.2	Hierarchische Clusteranalyse	276
	Quellen	281
8	Maschinelles Lernen in der Anwendung	283
8.1	Anwendungsbeispiele für überwachtes Lernen	283
8.1.1	MNIST: Handschrifterkennung	284
8.1.2	CIFAR-10: Objekterkennung	286
8.1.3	Large Movie Review Dataset: Sentimentanalyse	288
8.2	Anwendungsbeispiele für unüberwachtes Lernen	293
8.2.1	Textanalyse: Themenmodellierung	293
8.2.2	Netzwerkanalyse: Gemeinschaftsstrukturen	294
	Quellen	299
	Ergänzende Literatur	303
	Sachverzeichnis	307