

<b>Vorwort</b> .....	<b>XIII</b>
<b>1 Explorative Datenanalyse</b> .....	<b>1</b>
Strukturierte Datentypen .....	2
Weiterführende Literatur .....	4
Tabellarische Daten .....	5
Data Frames und Tabellen .....	6
Nicht tabellarische Datenstrukturen .....	7
Weiterführende Literatur .....	8
Lagemaße .....	8
Mittelwert .....	9
Median und andere robuste Lagemaße .....	11
Beispiel: Lagemaße für Einwohnerzahlen und Mordraten .....	12
Weiterführende Literatur .....	14
Streuungsmaße .....	14
Standardabweichung und ähnliche Maße .....	15
Streuungsmaße auf Basis von Perzentilen .....	17
Beispiel: Streuungsmaße für die Einwohnerzahlen der Bundesstaaten in den USA .....	19
Weiterführende Literatur .....	20
Exploration der Datenverteilung .....	20
Perzentile und Box-Plots .....	21
Häufigkeitstabellen und Histogramme .....	23
Dichtediagramme und -schätzer .....	25
Weiterführende Literatur .....	27
Binäre und kategoriale Daten untersuchen .....	28
Modus .....	30
Erwartungswert .....	30
Wahrscheinlichkeiten .....	31
Weiterführende Literatur .....	32

Korrelation . . . . .	32
Streudiagramme . . . . .	35
Weiterführende Literatur . . . . .	37
Zwei oder mehr Variablen untersuchen . . . . .	37
Hexagonal-Binning- und Konturdiagramme (Diagramme für mehrere numerische Variablen) . . . . .	38
Zwei kategoriale Variablen . . . . .	41
Kategoriale und numerische Variablen . . . . .	42
Mehrere Variablen visualisieren . . . . .	44
Weiterführende Literatur . . . . .	47
Zusammenfassung . . . . .	47
<b>2 Daten- und Stichprobenverteilungen . . . . .</b>	<b>49</b>
Zufallsstichprobenziehung und Stichprobenverzerrung . . . . .	50
Verzerrung . . . . .	52
Zufallsauswahl . . . . .	53
Größe versus Qualität: Wann spielt die Stichproben- größe eine Rolle? . . . . .	54
Unterschied zwischen dem Stichproben- und dem Populations- mittelwert . . . . .	56
Weiterführende Literatur . . . . .	56
Auswahlverzerrung . . . . .	57
Regression zur Mitte . . . . .	58
Weiterführende Literatur . . . . .	60
Stichprobenverteilung einer statistischen Größe . . . . .	60
Zentraler Grenzwertsatz . . . . .	63
Standardfehler . . . . .	64
Weiterführende Literatur . . . . .	65
Bootstrap-Verfahren . . . . .	65
Unterschiede zwischen Resampling und dem Bootstrap- Verfahren . . . . .	69
Weiterführende Literatur . . . . .	69
Konfidenzintervalle . . . . .	70
Weiterführende Literatur . . . . .	72
Normalverteilung . . . . .	73
Standardnormalverteilung und Q-Q-Diagramme . . . . .	75
Verteilungen mit langen Verteilungsenden . . . . .	76
Weiterführende Literatur . . . . .	79
Studentsche t-Verteilung . . . . .	79
Weiterführende Literatur . . . . .	81
Binomialverteilung . . . . .	81
Weiterführende Literatur . . . . .	84

Chi-Quadrat-Verteilung . . . . .	84
Weiterführende Literatur . . . . .	85
F-Verteilung . . . . .	86
Weiterführende Literatur . . . . .	86
Poisson- und verwandte Verteilungen . . . . .	87
Poisson-Verteilung . . . . .	87
Exponentialverteilung . . . . .	88
Die Hazardrate schätzen . . . . .	89
Weibull-Verteilung . . . . .	89
Weiterführende Literatur . . . . .	90
Zusammenfassung . . . . .	90
<b>3 Statistische Versuche und Signifikanztests . . . . .</b>	<b>91</b>
A/B-Test . . . . .	92
Warum eine Kontrollgruppe nutzen? . . . . .	94
Warum lediglich A/B? Warum nicht auch C, D usw.? . . . . .	95
Weiterführende Literatur . . . . .	97
Hypothesentests . . . . .	97
Die Nullhypothese . . . . .	99
Die Alternativhypothese . . . . .	99
Einseitige und zweiseitige Hypothesentests . . . . .	99
Weiterführende Literatur . . . . .	100
Resampling . . . . .	101
Permutationstest . . . . .	101
Beispiel: Die Affinität von Nutzern zu einem Webinhalt messen (Web-Stickiness) . . . . .	102
Exakte und Bootstrap-Permutationstests . . . . .	106
Permutationstests: ein geeigneter Ausgangspunkt in der Data Science . . . . .	107
Weiterführende Literatur . . . . .	108
Statistische Signifikanz und p-Werte . . . . .	108
p-Wert . . . . .	111
Signifikanzniveau . . . . .	112
Fehler 1. und 2. Art . . . . .	113
Data Science und p-Werte . . . . .	114
Weiterführende Literatur . . . . .	114
t-Tests . . . . .	115
Weiterführende Literatur . . . . .	117
Testen mehrerer Hypothesen . . . . .	117
Weiterführende Literatur . . . . .	120
Die Anzahl der Freiheitsgrade . . . . .	121
Weiterführende Literatur . . . . .	122

Varianzanalyse (ANOVA) . . . . .	122
F-Statistik . . . . .	126
Zweifaktorielle Varianzanalyse . . . . .	128
Weiterführende Literatur . . . . .	128
Chi-Quadrat-Test . . . . .	128
Chi-Quadrat-Test: ein Resampling-Ansatz . . . . .	129
Chi-Quadrat-Test: die statistische Theorie . . . . .	131
Exakter Test nach Fisher . . . . .	133
Relevanz in der Data Science . . . . .	135
Weiterführende Literatur . . . . .	136
Mehrmarmige Banditen . . . . .	136
Weiterführende Literatur . . . . .	139
Trennschärfe und Stichprobengröße . . . . .	140
Stichprobengröße . . . . .	142
Weiterführende Literatur . . . . .	144
Zusammenfassung . . . . .	144
<b>4 Regression und Vorhersage . . . . .</b>	<b>145</b>
Lineare Einfachregression . . . . .	145
Die Regressionsgleichung . . . . .	147
Angepasste Werte und Residuen . . . . .	149
Die Methode der kleinsten Quadrate . . . . .	151
Unterschied zwischen Vorhersage- und erklärenden Modellen . . . . .	152
Weiterführende Literatur . . . . .	153
Multiple lineare Regression . . . . .	153
Beispiel: Die King-County-Immobilien­daten . . . . .	154
Das Modell bewerten . . . . .	155
Kreuzvalidierung . . . . .	158
Modellauswahl und schrittweise Regression . . . . .	159
Gewichtete Regression . . . . .	163
Weiterführende Literatur . . . . .	164
Vorhersage mittels Regression . . . . .	164
Risiken bei der Extrapolation . . . . .	165
Konfidenz- und Prognoseintervalle . . . . .	165
Regression mit Faktorvariablen . . . . .	167
Darstellung durch Dummy-Variablen . . . . .	168
Faktorvariablen mit vielen Stufen . . . . .	171
Geordnete Faktorvariablen . . . . .	173
Interpretieren der Regressionsgleichung . . . . .	174
Korrelierte Prädiktorvariablen . . . . .	175
Multikollinearität . . . . .	176

Konfundierende Variablen . . . . .	177
Interaktions- und Haupteffekte . . . . .	178
Regressionsdiagnostik . . . . .	180
Ausreißer. . . . .	181
Einflussreiche Beobachtungen . . . . .	183
Heteroskedastische, nicht normalverteilte und korrelierte Fehler . . . . .	186
Partielle Residuendiagramme und Nichtlinearität . . . . .	190
Polynomiale und Spline-Regression. . . . .	192
Polynome . . . . .	193
Splines. . . . .	195
Verallgemeinerte additive Modelle . . . . .	197
Weiterführende Literatur . . . . .	199
Zusammenfassung . . . . .	199
<b>5 Klassifikation . . . . .</b>	<b>201</b>
Naiver Bayes-Klassifikator . . . . .	203
Warum eine exakte bayessche Klassifikation nicht praktikabel ist . . . . .	204
Die naive Lösung . . . . .	204
Numerische Prädiktorvariablen . . . . .	207
Weiterführende Literatur . . . . .	207
Diskriminanzanalyse . . . . .	208
Kovarianzmatrix . . . . .	209
Lineare Diskriminanzanalyse nach Fisher . . . . .	209
Ein einfaches Beispiel . . . . .	210
Weiterführende Literatur . . . . .	213
Logistische Regression. . . . .	214
Logistische Antwortfunktion und Logit-Funktion . . . . .	215
Logistische Regression und verallgemeinerte lineare Modelle . . . . .	217
Verallgemeinerte lineare Modelle . . . . .	218
Vorhergesagte Werte aus der logistischen Regression . . . . .	219
Interpretation der Koeffizienten und Odds-Ratios . . . . .	220
Lineare und logistische Regression: Gemeinsamkeiten und Unterschiede . . . . .	221
Das Modell prüfen und bewerten . . . . .	223
Weiterführende Literatur . . . . .	226
Klassifikationsmodelle bewerten . . . . .	227
Konfusionsmatrix . . . . .	228
Die Problematik seltener Kategorien . . . . .	230
Relevanz, Sensitivität und Spezifität . . . . .	231
ROC-Kurve . . . . .	232

Fläche unter der ROC-Kurve (AUC) . . . . .	234
Lift . . . . .	236
Weiterführende Literatur . . . . .	238
Strategien bei unausgewogenen Daten . . . . .	238
Undersampling . . . . .	239
Oversampling und Up/Down Weighting . . . . .	240
Generierung von Daten . . . . .	241
Kostenbasierte Klassifikation . . . . .	242
Die Vorhersagen untersuchen . . . . .	243
Weiterführende Literatur . . . . .	244
Zusammenfassung . . . . .	244
<b>6 Statistisches maschinelles Lernen . . . . .</b>	<b>247</b>
K-Nächste-Nachbarn . . . . .	248
Ein kleines Beispiel: Vorhersage von Kreditausfällen . . . . .	249
Distanzmaße . . . . .	252
1-aus-n-Codierung . . . . .	253
Standardisierung (Normierung, z-Werte) . . . . .	254
K festlegen . . . . .	257
KNN zur Merkmalskonstruktion . . . . .	258
Baummodelle . . . . .	260
Ein einfaches Beispiel . . . . .	262
Der Recursive-Partitioning-Algorithmus . . . . .	264
Homogenität und Unreinheit messen . . . . .	266
Den Baum daran hindern, weiterzuwachsen . . . . .	267
Vorhersage eines kontinuierlichen Werts . . . . .	270
Wie Bäume verwendet werden . . . . .	270
Weiterführende Literatur . . . . .	271
Bagging und Random Forests . . . . .	271
Bagging . . . . .	273
Random Forest . . . . .	273
Variablenwichtigkeit . . . . .	278
Hyperparameter . . . . .	281
Boosting . . . . .	282
Der Boosting-Algorithmus . . . . .	283
XGBoost . . . . .	284
Regularisierung: Überanpassung vermeiden . . . . .	287
Hyperparameter und Kreuzvalidierung . . . . .	292
Zusammenfassung . . . . .	295

<b>7 Unüberwachtes Lernen</b> . . . . .	<b>297</b>
Hauptkomponentenanalyse. . . . .	298
Ein einfaches Beispiel . . . . .	299
Die Hauptkomponenten berechnen . . . . .	302
Die Hauptkomponenten interpretieren . . . . .	302
Korrespondenzanalyse . . . . .	305
Weiterführende Literatur . . . . .	307
K-Means-Clustering. . . . .	307
Ein einfaches Beispiel . . . . .	308
Der K-Means-Algorithmus . . . . .	311
Die Cluster interpretieren . . . . .	312
Die Anzahl von Clustern bestimmen . . . . .	314
Hierarchische Clusteranalyse . . . . .	317
Ein einfaches Beispiel . . . . .	317
Das Dendrogramm . . . . .	318
Der agglomerative Algorithmus . . . . .	320
Ähnlichkeitsmaße . . . . .	321
Modellbasierte Clusteranalyse. . . . .	322
Multivariate Normalverteilung . . . . .	323
Zusammengesetzte Normalverteilungen (gaußsche Mischverteilungen) . . . . .	324
Die Anzahl der Cluster bestimmen . . . . .	327
Weiterführende Literatur . . . . .	329
Skalierung und kategoriale Variablen . . . . .	329
Variablen skalieren . . . . .	330
Dominierende Variablen. . . . .	332
Kategoriale Daten und die Gower-Distanz . . . . .	334
Probleme bei der Clusteranalyse mit verschiedenen Datentypen . . . . .	336
Zusammenfassung . . . . .	338
<b>Quellenangaben</b> . . . . .	<b>339</b>
<b>Index</b> . . . . .	<b>341</b>