

Stichwortverzeichnis

A

Abbildung
 Aussagekraft 207
 Erkenntnisse 208
 Relevanz 207
Abgeleitetes Attribut 186
Abstandsmatrix 119
Ad-hoc-Abfrage 163
Ähnliche Elemente 110
Ähnlichkeitswert 110
Agent 130
Aktienmarkt 103
Aktivierungsfunktion 149
Aktualisierung 98
Algorithmus 29, 115
 Auswahl 196
 Bayes-Klassifikator 144
 Clusteralgorithmus 209
 DBSCAN 263
 Dimensionalitätsverringereungs-
 algorithmus 238, 255
 Entscheidungsbaum 139
 Flock by Leader-Algorithmus
 126, 217
 k-Means-Algorithmus 115,
 253
 Nächste-Nachbarn-
 Algorithmus 119
 Neuronale Netze 148
 Regressionsalgorithmus 245
 Support Vector Machine 141
Amazon 37, 41, 53, 161, 179
Ameisenkolonie 128
American Cancer Society Surveil-
 lance Research 102
Annahmen
 Einschränkung 308
Anpassen eines Modells 304
Anpassung von Kurven 306
Anpassungsgüte 283
Apache Hadoop 319
Apache Mahout 322
Application Programming
 Interface 124, 325
Artikelbasiertes kollaboratives Fil-
 tern 44
Assoziationsregeln 121
Assoziatives Modell 100

Attribut 29, 84, 110, 233
 abgeleitetes 186
Ausgabeschicht 148
Ausreißer 60, 156, 211, 266, 299
 Berücksichtigung 301
 durch äußere Einflüsse 299
 durch Systemfehler 300
Auto-MPG-Datensatz 278

B

Balkendiagramm 86
Batmans Rückkehr 123
Bayes-Klassifikator 144
Bellaachia, Abdelghani 102, 126
Berührungspunkt 167
Beschreibende Statistik 73
Betreffzeile 65
Betriebskosten 336
Bewertung 165
 Tabelle 166
Bias 149, 200
Big Data 27, 36, 77
 Sammlung 162
 und Business Intelligence 83
 und Predictive Analytics 315
Bildererkennung 141
Bildzerlegung 112
Börsengang 161
Boolscher Operator 275
Brustkrebs 102
Budget 171
Business Intelligence 67, 83

C

Centers for Disease Control and
 Prevention 101
Churn-Modellierung 168
Cluster-Schwerpunkt 115
Cluster-Vertreter 110, 115
Clustering 99, 109 f., 115, 208
 biologisch inspiriert 123
 DBSCAN 263
 k-Means-Algorithmus 115, 253
 Motivation 111
 Nächste-Nachbarn-
 Algorithmus 119

 und Klassifizierung 132
 Visualisierung 208, 255
comparison 284
concatenate 274
CRM-Datenbank 166
Customer Lifetime Value 167
Cyberattacke 319

D

Dashboard 163
Data Frame 276
Data-Mining 27, 74
 Herausforderungen 296
Data-Profiling-Verfahren 185
Data-Science-Technologie 34
Data-Warehouse 318
Daten 35
 Anforderungen 332
 Attribute 84
 Aufbereitung 181, 232, 279,
 288, 344
 Auswahl 345
 Beschreibung 100
 Data Frame 276
 demografisch 70
 Dimensionalität 187
 doppelte 186
 Einstellungsdaten 68
 Entfernung doppelter Einträge
 176
 Faktor 277
 fehlende 185
 fließend 66
 Geschwindigkeit 80
 Glättung 302
 Grenzen 297
 Kategorien 67
 Kauf von Dritten 177
 linear trennbar 253
 logische 275
 Mangel an 176
 Modellierung 95
 numerische 275
 Organisation 333
 Reduzierung 84
 Reinigung 185
 Schulungsdaten 138

- statisch 66
 - strukturiert 64
 - Strukturierung 188
 - Testdaten 138, 189
 - unstrukturiert 64
 - Validität 78
 - Verarbeitung 183
 - verbinden 317
 - verflachen 187
 - Verhaltensdaten 67
 - Volumen 80
 - Vorbereitung 194
 - Zeichenkette 275
 - Datenelemente 110
 - Korrelationen 121
 - Datenerfassung 51
 - Datengesteuerte Analyse 71
 - und nutzergesteuerte Analyse 73
 - Datenmatrix 110
 - Datenpaket 276
 - Datensatz 109
 - Teilung 189
 - Datenstruktur 276
 - Datentypen 63, 298
 - R 275
 - Datenunternehmen 162
 - Datenverbund 173
 - Datenvielfalt 79
 - Datum 298
 - DBSCAN 263 f.
 - Demografische Daten 70
 - Dimensionalität 187
 - Verringerung 187
 - Dimensionalitätsverringereungs-
algorithmus 238, 255
 - Distributed File System 320
 - DJIA 103
 - Document-Term Matrix 113
 - Dokument 113
 - Dow-Jones-Index 103
 - Downdrill 205
 - Duhigg, Charles 104
- E**
- Einfacher Bayes-Klassifizierer 143
 - Algorithmus 144
 - Grundbegriffe 144
 - Einflussnehmer 40
 - Eingabeschicht 148
 - Einstein, Albert 31
 - Einstellungsdaten 68
 - und Verhaltensdaten 69
 - Empfehlungsdienst 30, 41
 - Personalisierung 43
 - Realisierung 43
 - Endknoten 292
 - Ensemble-Modell 156, 197
 - Entdeckungsphase 135
 - Entität 84
 - Entscheidungsbaum 102, 139,
211, 326
 - Algorithmus 140
 - Erwartungswert 140
 - Programmpaket 291
 - Entscheidungsfunktion 241
 - Entscheidungsmodell 99
 - Erdbebenvorhersage 105
 - Ereignis 143
 - Erfolg
 - Definition 171
 - Erwartungswert 140
 - ETL-Prozess 188
 - Euklidischer Abstand 117, 120
 - Evidenz 144, 147
 - Explizite Datenerfassung 51
 - Externe Daten 136
 - Extraktionsschritt 188
 - Extremwert 299
- F**
- Facebook 79, 124, 161, 315
 - Faktor 277
 - Fehler 156
 - Fehlersuche 200
 - FICO Score 60, 132, 165
 - Filtern
 - inhaltsbasiert 51
 - kollaborativ 37, 43
 - Fisher, Ronald 233
 - Fließende Daten 66
 - Flock by Leader-Algorithmus 126,
217
 - Fraud-Detection 185
 - Freiform-Eingabefeld 52
 - Funktion
 - Aufruf 277
- G**
- Gefährten 124
 - Genauigkeitsgrad 51
 - Genexpression 113
 - Genklassifikationsanalyse 196
 - vorrangig 182
 - zweitrangig 182
 - Geschäftskompetenz 342
 - Geschäftsziele 182, 343
 - Geschwindigkeit von Daten 80
 - Gesundheitswesen 101
 - Gewichtung 114
 - Ginsberg, Jeremy 101
 - Glättung von Daten 302
 - Glaubwürdigkeit 52
 - Gleitender Mittelwert 186, 304
 - Gleitkommazahl 275
 - Gmail 79
 - Google 79, 315
 - Google AdWords 56
 - Google Profile of Mood States 103
 - GPOMS 103
 - GPS 315
 - Granger Causality Analysis 103
 - Graph 88
 - Graphentheorie 88
 - Grenzwert 119
 - Grippeepidemie 101
- H**
- Hadoop 319
 - Hadoop Distributed File System
320
 - Handelsmodell 29
 - Hauptkomponentenanalyse 238
 - HDFS 320
 - head 279
 - Hidden Layer 148
 - Hidden Markov-Modell 153
 - Hybrid-Empfehlungsdienst 55
- I**
- IBM 82
 - IDE 270
 - Implizite Datenerfassung 51
 - Inferenzstatistik 74
 - Information Retrieval 113
 - Inhaltbasiertes Filtern 51
 - Instanz 84, 132
 - Integrierte
 - Entwicklungsumgebung 270
 - Intelligente Lösung 97
 - Interaktive Datenuisualisierung
205
 - Interne Daten 136
 - Interpolation 306
 - Interpretierte Sprache 270
 - Intervallattribut 84

Iris-Datensatz 233
 Clusteralgorithmen 252
 DBSCAN 264
 k-Means-Algorithmus 253
 logistische Regression 245
 unüberwachtes Lernen 251
 Iterative Methode 347
 Iterativer Prozess 33, 117

K

K (algebraisch) 115
 k-Means-Algorithmus 115
 k-Nearest-Neighbor-Methode 43, 119
 k-Wert
 Variation 261
 Kaltstartproblem 47, 50 f., 55
 Kante 292
 Kapitalrendite 336
 Kartografie 203
 Kernel-Funktion 143
 Klassifikationsalgorithmus 141
 Klassifikationsmodell 99
 Klassifikator 131
 Klassifizierer 133, 234
 Klassifizierung von Daten 131, 287
 Ablauf 137
 Algorithmen 139
 Anwendungen 132
 Bayes-Klassifizierer 143
 Gesundheit 134
 Implementierung 135
 Marketing 133
 Markov-Modell 150
 Neuronale Netze 148
 Support Vector Machine 141
 und Clustering 132
 Zukünftige Anwendungen 135
 Klassifizierungsmodell 131
 Klickvergütung 56
 Knoten 148, 292
 Knotenpunkt 140
 Kolakowski, Nick 107
 Kollaboratives Filtern 37, 43
 artikelbasiert 44
 nutzerbasiert 48
 Vergleich 50
 Konfidenzniveau 284
 Konfidenzwert 122
 Konfusionsmatrix 243
 Kreditrisiko 132

Kreuzvalidierung 199, 311
 Kunde
 Klassifizierung 136
 Kundenabwanderung 165
 Kundenbeibehaltung 167
 Kundenbetreuungsteam 172
 Kundenertragswert 167
 Kurven
 Anpassung 306
 Kurvendiagramm 90

L

Ladeschritt 188
 Legende 208
 Leitvogel 126
 Lernphase 137
 Level 277
 Lineare Regression 155
 LinkedIn 41, 161, 179
 Logistische Funktion 149
 Logistische Regression 244

M

Mahout 322
 k-Means-Algorithmen 322
 MAPE 285
 MapReduce 320
 Marketing
 zielgerichtet 56
 Marketingtrends 165
 Markov-Annahme 150, 154
 Markov-Kette 150
 Markov-Modell 150
 Markov-Vorhersage 151
 Marktsegmentierung 111
 Maschinelles Lernen 30, 74
 scikit 224
 Masterknoten 320
 matplotlib
 Installation 231
 Matrix 276
 Mean Absolute Percent Error 285
 Means 115
 Meinungs-Mining 103
 Merkmalsgewinnung 238
 Metadaten 82
 metrics 243
 Metrik 33
 Microarray Gene Expression Data 113
 Mining Association Rules 121

Mittelwert 115
 Modell 28, 95
 Aktualisierung 98, 350
 assoziatives 100
 Beurteilung 200, 311
 Definition 96
 Einsatz 201, 349
 Entscheidungsmodell 99
 Entwicklung 197
 Kategorien 98
 Skalierbarkeit 312
 Testen 198
 Überprüfung 350
 Überwachung 202
 und Simulation 96
 Verfeinerung 97
 vorhersagendes 98
 Modellerstellung 96, 191, 282, 290
 Algorithmenwahl 196
 Einstieg 191
 Entwickeln und Testen 197
 Geschäftsziele 193
 Multiclass-SVM 141

N

Nächste-Nachbarn-Algorithmus 119, 326
 National Cancer Institute 102
 Nearest Neighbors Algorithm 119, 326
 Netflix 41, 53, 161, 179
 Netizen 215
 Neuron 148
 Neuronales Netz 148
 Schichten 148
 Nominalattribut 84
 Nullhypothese 284
 Numerische Daten 275
 numpy
 Installation 229
 Nutzerbasiertes kollaboratives Filtern 48
 Nutzergesteuerte Analyse 72
 und datengesteuerte Analyse 73

O

Obama, Barack 60, 107
 Objekt 84, 132
 Ockham, William von 309

- Ockhams Rasiermesser 309
- OLAP 83
- Online Analytical Processing 83
- Online-Marketing 41
- Online-Werbe-Netzwerk 56
- Ontologie 82
- Opentable.com 79
- Opinion Finder 103
- Opinion-Mining 317
- Ordnungsattribut 84

- P**
- Parameterliste 277
- party 291
- Pheromone 128
- Pilotprojekt 180
- Prädiktor 97
- Prädiktorvariable 283, 290
- Präzision 32, 53 f.
 - und Recall 55
- predict 262, 286
- Predictive Analytics
 - als Service 316
 - Definition 28, 178
 - Disziplinen 73
 - Prototyp 177, 323
 - und Big Data 315
 - Vorteile 164
- Programmierschnittstelle 124
- Prototyp 177, 323
- Public Stream 124
- Python 223
 - Installation 223
 - Zusatzdateien 226

- R**
- R 269
 - Community 270
 - Datenstrukturen 276
 - Datentypen 275
 - Einführung 273
 - Funktionen 277
 - Installation 271
 - Operatoren 274
 - Programmierung 270
 - Regressionsanalyse 278
 - Zuordnung von Variablen 273
- Rabattorientierte Kunden 136
- Random Forest 213
- Random Walk 304
- random_state 255
- randomForest 291
- RapidMiner 113
- Rauschen 196, 264, 302
 - Reduzierung 303
- Reaktionsrate 168
- Recall 54
 - und Präzision 55
- Rechtsfragen 137
- Reduzierung 84
- Regression
 - lineare 155
 - logistische 244
- Regularisierungsparameter 245
- Reinigung von Daten 185
- Responsemodellierung 57
- Restbetrag 156
- Risikomodellierung 131
- Rohdaten 113, 183, 188
- rpart 291
- RStudio 270
 - Benutzeroberfläche 271
 - Installation 271

- S**
- Saisonabhängige Kunden 136
- Saisonabhängigkeit 297
- Satz von Bayes 144
- Schlagwortwolke 89
- Schlüssel-Wert-Paar 321
- Schlüsselwörter 81, 106
- Schulungsdaten 102, 138
- Schwangerschaftsvorhersage 104
- Schwarmintelligenz 90
- Schwarmregeln 215
- Schwarmverhalten 123, 215
- scikit-learn 223
 - Installation 227
 - Installation prüfen 231
 - maschinelles Lernen 224
- scipy
 - Installation 230
- Score 165
- SEER 102
- Semantische Suche 81
- Sentiment Detection 103, 107
- Sentimentanalyse 147, 173, 317
- Sentimentextraktor 318
- Sigmoide Funktion 149
- Signal-Rausch-Verhältnis 303
- Simulation 215
 - und Modell 96
- Site Stream 125
- Skalierbarkeit 312
- Slaveknoten 320
- Smart Data 77
- Society of Industrial and Applied Mathematics 102
- Softwaretool 175
- SourceForge 224
- Spärlichkeit 51 f.
- Spam-Erkennung 112
- Statische Daten 66
- Statistik 73
- Strafverfolgung 113
- Streudiagramm 239, 258
- String 275
- Strukturierte Daten 64
 - und unstrukturierte Daten 65
- Stützvektor 235
- Suche
 - semantische 81
- summary 279
- Supervised Training 150
- Support Vector Machine 106, 141, 235
- SVM 141
- svmClassifier 238

- T**
- Tabelle 85
- Tableau 84
- Tag 51
- tail 279
- Taleb, Nasim 40
- Target Corporation 41
- Target Marketing 56
- Target Store 104, 137
- Team
 - datenwissenschaftliches 174
- Technologische Trends 316
- Term 113
- Term Frequency 113
- Testdaten 102, 138, 189, 198, 237
- Textanalyse 61
- Tortendiagramm 86
- Toy Story 44
- Trainingsdaten 190, 198, 237
- Transformationsschritt 188
- tree 291
- Trennlinie 241
- Treue Kunden 136
- True Lift Model 59
- Twitter 105, 124, 138, 161, 315, 325
- Typumwandlung 275

U

Überanpassung 143, 190, 196,
306
Vermeidung 307
Überwachte Analyse 310
Überwachtes Lernen 235, 251
UIMA 82
Unstructured Information Manage-
ment Architecture 82
Unstrukturierte Daten 64
und strukturierte Daten 65
Unsupervised Training 150
Unteranpassung 195
Unternehmenskenntnisse 33
Unternehmensvorteile 163
Unüberwachtes Lernen 251
Genauigkeit 258
Uplift-Modell 58
Upselling 197
User Stream 125

V

Validierungsdatensatz 198
Validität von Daten 78
Variable 184
Varianz 200

Vektor 129, 276
Verarbeitung von Daten 183
Verborgene Schicht 148
Verbrechensbekämpfung 60
Verfeinerung 97
Verflachung von Daten 187
Verhaltensdaten 67, 69
und Einstellungsdaten 69
Verkaufstrends 164
Verzweigungspunkt 140
Vielfalt von Daten 79
Visualisierung 84, 203
als vorhersagendes Element
203
Ausreißer 211
Balkendiagramm 86
Bedeutung 204
Bewertung 207
Graphen 88
Kurvendiagramm 90
Medium 208
Schlagwortwolke 89
Tabelle 85
Tortendiagramm 86
Vielschichtigkeit 206
Vogelschwarm 91
Vorteile 205
Weitere Arten 215

Vogelschwarm 123
Volumen von Daten 80
Vorhersagemodell 98
Vorhersagende Variable 308
Vorhersagephase 137
Vorkommenshäufigkeit 113

W

Wahrscheinlichkeit 143
Walmart 161
Warenkorb 122
Was-ist-wenn-Szenario 97
Weizen-Datensatz 287
Wettbewerbsvorteile 339
Widersprüchlichkeit 52
Wurzelknoten 292

Z

Zeichenkette 275
Zeitreihe 103
Zielgruppengerichtetes Marketing
56
Zufallsbewegung 304
Zufallsgenerator 283, 290

