

Stichwortverzeichnis

A

AdaBoostClassifier 397
Add-On 59
add_edge() 164, 213, 215
add_edges_from() 213, 215
add_node() 213, 215
add_nodes_from() 213, 215
add_path() 215
Adjazenzmatrix 163
Aggregation 147
Algorithmus
 k-Nearest Neighbors (kNN)
 334
 lineare Regression 324
 Naiver Bayes 329, 332
alpha-Parameter 364, 399
Amazon.com-Datensatz 415
American Standard Code for Information Interchange (ASCII)
 152
Anaconda 45, 58, 217
 auf Linux 64
 auf Windows 60
Analyse
 bivariate 263
 Daten 32, 256
Anmerkung 190 f.
annotate() 191
Anomalie 312
ANOVA (Analyse der Varianz) 265
Anpassung eines Modells 338
 Klassifikationsevaluierungsmaße 341
 Regressionsevaluierungsmaße 341
 Trainings- und Testsatz trennen 343
 Trend und Varianz 339
Ansatz
 multivariater 316
 univariater 312
Ansatz zur Datendichte 294
Anweisung 93
Anwendungsprogrammierschnittstellen (APIs) 121
append() 98
Archiv 68
Argument 88

Array 174
as_matrix() 209
AskSam 118
Aspirational Data Scientist 407
Ast
 Entscheidungsbaum 388
Ausdruck
 regulärer 155
Ausreißer 172, 310
 Anomalien und neue Daten 312
 aufspüren 310
 DBScan 318
 Einfluss auf maschinelle Lernalgorithmen 311
 Gauß-Verteilung 314
 lineare Regression 326
 multivariater Ansatz 316
 SVM (Support-Vector-Maschinen) 318
 univariater Ansatz 312, 315
Aussage
 bedingte 44
Average-Verknüpfungsmethode 303

B

Bag of Words 158, 242
Bagging-Klasse 393
Balkendiagramm 196 f.
BaseEstimator (Scikit-learn) 238
Basemap-Toolkit 212
Baum-Ensemble 140
Beautiful-Soup-Bibliothek 55
Beispielcode 66
Benchmarking
 timeit 247
BernoulliNB 332
Bibliothek 54
Big Data
 Clustering 301 f.
 Datenmenge bestimmen 367
 Definition 277
 Stochastischer Gradientenabstieg 367
Binning 127

Bivariate Analyse 263
Blatt
 Entscheidungsbaum 388
bool() 150
Boolescher Wert 79
Boosting
 Entscheidungsbäume 397
Boston-Datensatz 239, 325
 Probenschichtung für komplexe Daten 347
 Trainings- und Testsatz trennen 343
 Variablentransformation 358
Boxplot 259
 Ausreißer 313
 T-Test 265
 Tukey 313
 untersuchen 264
boxplot() 201

C

C-Parameter 364, 375, 379 f.
Caller 87
Canadian Institute for Advanced Research (CIFAR) 414
Chebyshev-Ungleichung 314
Checkpoint 230
Chi-Quadrat 273, 349, 351
CIFAR (Canadian Institute for Advanced Research) 414
class_weight-Parameter 377
Clear Screen (cls) 47
Cluster-Analyse
 Ausreißer 318
Clustering 293
 Big Data 299, 301
 Bilddatenbeispiel 297
 DBScan 306
 hierarchisches (agglomeratives) 302
 K-means-Algorithmus 294
 optimale Lösungen 298
 Partitions-Clustering-Techniken 294
 Technikarten 293
 Trägheit 299
Codeblock 88

Codierung 55, 152
 Common Crawl 2012 Web Corpus 416
 compile() 158
 Complete-Verknüpfungsmethode 303
 Component Object Model (COM) 121
 concatenate() 201
 Conductrics 406
 count_vect.fit_transform() 160
 CountVectorizer 155, 246
 cumsum-Funktion (NumPy) 280
 cycle_graph() 164

D

Data Munging 237
 Data Science 29
 Definition 31
 Ressourcensammlungen 403 f.
 Data Science London + Scikit-learn 410
 Data Science Weekly 404
 Data Scientist 35
 Kernkompetenzen 32
 Data Wrangling 237
 Data-Science-Central 405
 Dataframe 127
 kategorialer 261
 DataFrame-Objekt 358
 DataFrame.to_sql() 119
 Datei
 Komma-separierte 110
 Daten
 kategoriale 261
 kombinieren 143
 Validierung 127
 verketten 143
 Datenanalyse 32
 Algorithmus 51
 beschreibende Statistik für numerische Daten 257
 Chi-Quadrat 273
 EDA-Ansatz 256
 Häufigkeiten 262
 kategoriale Daten 261
 Kontingenztafeln 263
 Normalitätsmaß 260
 parallele Koordinaten 266
 Perzentile 259
 Varianz und Spannweite 258
 Datenbank
 relationale 105, 118

Datenbankmanagementsystem (DBMS) 118
 Datenerfassung 32
 Datenfehler
 Formen 309
 Datenform 125
 Datenmanagementsystem 105
 Datenplan 130
 Datensatz 74
 Amazon.com 415
 Canadian Institute for Advanced Research (CIFAR) 414
 Common Crawl 2012 Web Corpus 416
 Data Science London + Scikit-learn 410
 Größe 51
 Kaggle-Wettbewerb 411
 Madelon-Datensatz 411
 Mixed National Institute of Standards and Technology (MNIST) 413
 MovieLens 412
 Spambase 413
 Datentyp 84
 DBScan 306
 Ausreißer 318
 DecisionTreeClassifier-Klasse 389
 DecisionTreeRegressor-Klasse 389, 392
 Deque 98
 describe() 130
 Deterministische Selektion 245
 Diabetes-Datensatz 312
 Diagrammtyp 180
 Dictionary 97, 101
 DiGraph() 215
 Dimensionalität 277
 Faktoranalysen 282
 Filme empfehlen 289
 Gesichter erkennen 284
 PCA (Hauptkomponentenanalyse) 283, 297
 SVD (Singulärwertzerlegung) 278
 Themen mit NMF extrahieren 287
 verringern 277
 dir() 52
 Diskretisierung 172
 DisplayMulti() 91
 DisplaySum() 91

Distanz
 euklidische 296
 Distanzmetrik 303
 DoSum() 89
 dot-Funktion 279
 draw_networkx() 215
 drawcoastlines() 211
 drawcountries() 211
 drop_duplicates() 129
 dropna() 139
 dual-Parameter, LinearSVC 384
 Dünnbesetzte Matrix 245
 Duplikat 127

E

EDA (Exploratory Data Analysis) 256
 angewandte Visualisierung 263
 Datenverteilungen modifizieren 274
 Z-Score-Standardisierung 275
 Eingabeaufforderung 45, 152
 Einzigartige Varianz 281
 Elasticnet-Regularisierung 366 f.
 elif-Anweisung 93
 else-Anweisung 93
 Encapsulated PostScript (EPS) 183
 Ensemble maschineller Lernmethoden 252
 Ensemble von Modellen 387, 392
 Enthought Canopy Express 59
 Entscheidungsbaum
 Äste 388
 Blätter 388
 Levels 389
 Überblick 388
 Vorhersagen stärken 397
 Entwicklungsumgebung 48
 enumerate() 109
 eps-Parameter 306, 318
 Epsilon 382
 epsilon-Support-Vector-Regression (SVR) 382
 Euklidischer Abstand 335
 Evaluierungsmetrik
 Algorithmen 340
 extend() 98

F

F score 360
 F1-Wert 342
 Faktor 282

Faktoranalyse 281 f.
 Fall 106, 167
 Fehlende Daten
 lineare Regression 326
 Fehlerfunktion 311
 fetch_20newsgroups
 (subset='train') 75
 fetch_olivetti_faces() 75
 fillcontinents() 211
 fillna() 139
 Film empfehlen 289
 Filtern
 kollaboratives 289
 fit() 140
 fit(X,y) 239
 Fläche unter Grenzwertoptimie-
 rungskurve (ROC AUC) 342
 Flatfile 105, 110
 float() 84
 for-Schleifenanweisung 94, 108
 Funktion 87
 magische 224
 Funktionalität 36

G

gamma-Parameter 319, 375, 380,
 382
 GaussianNB 332
 Gauß-Verteilung 274
 Ausreißer 314
 GBM (Gradient Boosting
 Machine) 398 f.
 Hyperparameter 400
 Genauigkeitsfehlermaß
 Klassifikation 341
 General Public License (GPL) 60
 Gesichtserkennung 284
 Gini-Index 389
 GitHub 408
 Gleitkommazahl 79
 Glockenkurve 199
 GradientBoostingClassifier 399
 GradientBoostingRegressor 399
 Graph 212
 gerichtet 195, 212
 ungerichtet 195, 212
 Greedy-Selektion 350
 Greifpunkt 183
 Grid searching 352
 GridSearch-Klasse 354
 GridSearchCV 356, 379
 Ground Truth 298
 groupby() 130, 148

H

Hadoop-Cluster 168
 Häufigkeit 261 f.
 Handgeschriebene Daten 413
 Handle 183
 hash() 243
 Hash-Funktion 242
 Hashing-Trick 242, 246 f.
 HashingVectorizer 246 f.
 Haupteffektmodell 360
 Hauptkomponentenanalyse 281,
 316
 help() 221
 Hilfe
 interaktive 221, 223
 Hilfe-Modus 221 f.
 Histogramm 199, 267
 HTML 149, 227
 Hyperebene 373
 Hyperparameter 351
 GBM (Gradient Boosting
 Machine) 400
 Scikit-learn 239
 Hypothese 241

I

IDA (Initial Data Analysis) 256
 Identitätsoperator 84
 if-Anweisung 44, 92
 Imperativ 36
 Import 85, 96
 Imputer 140
 imread() 116
 imshow() 116
 Indikatorvariable 173
 Information Retrieval (IR) 159
 Informationsredundanz 271
 input() 93
 Instanziierung
 Scikit-learn 239
 int() 84, 93
 Integer 78
 Integrated Development Environ-
 ment (IDE) 60
 Interface
 Schätzfunktion (Scikit-learn) 239
 IPython Notebook 39, 67 f., 217,
 226
 IPython-Hilfe 223
 IPython-Konsole 217, 221
 IQR (Interquartilsabstand) 259,
 264, 313

Iris-Datensatz
 logistische Regression 327
 Matrix von Streudiagrammen
 269
 Übersicht 257
 versteckte Faktoren 282
 isin() 135
 isnull() 133
 Iterator 100
 ix[] 135, 207

J

jQuery 121

K

K-means-Algorithmus 294
 K-neighbors-Klassifizierer 353
 Kaggle-Wettbewerb 410
 Kategoriale Daten 261
 KDnuggets 404
 Kernel 228
 Kernel-Spezifikation 375
 key() 52
 KFold
 Kreuzvalidierung 346
 Klasse
 Scikit-learn 238
 Klasseneinteilung 172, 261
 Klassenszenario
 unausgeglichenes 377
 Klassifikation 74
 SVC 375, 377
 SVM 376
 unüberwachte 293
 Klassifikationsevaluierungsmaß 341
 Klassifizierer 158, 162
 Knäuel 163
 kNN (k-Nearest Neighbors) 334
 Knoten 163
 Komma-separierte Datei (CSV) 110
 Komponente 283
 Kontextualisierung 168
 Kontingenztafel 263
 Koordinate
 parallele 266
 Korrelation 271
 multivariate 171
 nichtparametrische 273
 nutzen 272
 Pearson 273
 Quadrat 272
 Spearman 273

Kosinus-Distanz 303
 Kovarianz 270 f.
 Kreisdiagramm 196
 Kreuzvalidierung 345
 allgemein 252
 KFolds 346
 cross_val_score-Funktion 346, 378
 csc_matrix 245
 Hyperparameter 351, 356
 Probenschichtung für komplexe Daten 347
 Kurtosis *siehe* Wölbung

L

L1-Regularisierung (Lasso) 364, 366
 l1_ratio Parameter 366
 L2-Regularisierung (Ridge) 364
 Label 190
 Lade-Technik 51
 last_valid_index() 144
 Latent Semantic Indexing (LSI) 281
 Leerzeichen-Regel 78
 legend() 192
 Legende 190, 192, 209
 Levels
 Entscheidungsbaum 389
 Lineare Regression
 Formel 324, 327
 Limitierungen und Probleme 326
 Regressionsevaluierungsmaße 341
 Scikit-learn 239
 Variablen 324
 Lineares Modell
 regularisieren 364
 LinearSVC-Klasse 374, 383 ff.
 Liniendiagramm 182
 Liniensstil 186
 load_boston() 74
 load_diabetes() 74
 load_digits([n_class]) 75
 load_iris() 74
 Logistische Regression
 Algorithmen 327
 Los Alamos National Laboratory
 Stability of Unstable Learning Algorithms 413
 loss-Parameter, LinearSVC 384
 LSTAT
 nichtlineare Transformation 359

M

Madelon-Datensatz 411
 Magische Funktion 224
 make_classification 385
 Manhattan (manhattan oder 11) Distanz 303, 335
 map() 151
 Marker 188
 Maschinelles Lernen
 allgemein 337
 Kreuzvalidierung 345
 No-free-lunch-Theorem 340
 problematische Aspekte 338
 Rastersuche 352
 Trend und Varianz 339
 Zufallssuche 356
 Maschinencode 41
 MathJax-Bibliothek 68
 MATLAB 179
 Matplotlib 55, 179
 Matrix 174, 176
 Dimensionalitätsverringern 279
 dünnbesetzte 245
 Streudiagramme 269
 max_depth-Parameter 390
 max_samples-Parameter 393
 max_features-Parameter 393
 mean_absolute_error-Evaluierungsmaß 341
 mean_squared_error-Evaluierungsmaß 341
 Median 258
 Mehrkern-Verarbeitung 251
 Menge 96
 Merkmal 106, 167
 Merkmalerstellung 171
 Microservice 120
 min_sample-Parameter 306, 318
 MiniBatchKMeans 301 f.
 Miniconda Installer 59
 Mittelwert 258
 Mixed National Institute of Standards and Technology (MNIST) 413
 model Interface, Scikit-learn 238
 Model-Objektklasse
 Scikit-learn 241
 Modeling for Optimal Probability Prediction 413
 Modell
 Anpassung 338
 Modul
 Import 85

MongoDB 120
 MovieLens 412
 MultiDiGraph() 215
 MultiGraph() 215
 Multimarker-Vorhersage 252
 MultinomialNB 331
 Multiprocessing 251
 Multivariate Korrelation 171
 Mustersuche 33, 155

N

N-Gramm 160
 n_estimators 397
 n_iter-Parameter 356
 Naiver Bayes-Algorithmus 329
 Formel 330
 Textklassifizierung 332
 National Institute of Standards and Technology (NIST) Datensatz 413
 Natural Language Processing (NLP) 159
 Natural Language Toolkit (NLTK) 154
 ndarray, NumPy 289, 358
 NetworkX 163, 213
 20-NewsGroups-Datensatz 158, 287
 Nicht-SQL-(NoSQL-)-Datenbank 120
 Nichtlineare Transformationen 357
 NIPS-2003-MerkmalAuswahl-Herausforderung 411
 NMF (Nicht-negative Matrix-Faktorisierung-Zerlegungsklasse) 287, 289
 No-free-lunch-Theorem 340
 Normalverteilung 274
 Notebook
 exportieren 72
 löschen 72
 neu anlegen 70
 Notebook Converter (NBConverter) 228
 nu-Parameter 319
 NumPy 54, 126
 cumsum 280
 Iris-Datensatz 257
 Kovarianz und Korrelation 270
 linalg 279
 logspace 379
 ndarray 312, 358

O

objectify.parse() 123
 Objektorientiert 36, 225
 Olivetti-Gesichtsdatensatz 284
 One versus one 328
 One versus rest 328
 One-Hot-Codierung 243
 OneVsOneClassifier-Klasse 328
 OneVsRestClassifier-Klasse 328
 Online Policy Adaptation for Ensemble Classifiers 413
 open() 107
 Open-Source-Data-Science-Master (OSDSM) 406
 Operator 81
 != 82
 % 80
 & 81
 * 80
 ** 80
 **= 80
 *= 80
 + 80 f.
 += 80
 - 80 f.
 / 80
 // 80
 //= 80
 /= 80
 < 82
 << 81
 <= 82
 > 82
 >> 81
 >= 82
 ^ 81
 ~ 81
 = 80
 == 82
 %= 80
 -= 80
 and 82
 arithmetischer 80
 Bitoperator 81
 Identitätsoperator 84
 in 84
 is 85
 is not 85
 not 82
 not in 84
 or 82
 Rangfolge 83
 relationaler 82
 unärer 81

Zugehörigkeitsoperator 84
 zuweisender 80
 ord() 84

P

Paket 58
 Pandas-Bibliothek 54, 126
 DataFrame 257
 kategoriale Daten 261
 Kovarianzmatrix 271
 pandas.crosstab-Funktion 263
 parallele Koordinaten 266
 parse() 115
 Parsen 115, 152
 PCA (Hauptkomponentenanalyse) 283
 Ausreißer 316
 Bilddaten 297
 Gesichter erkennen 284
 pd.DataFrame.duplicated() 128
 Pearson-Korrelation 270, 272
 penalty-Parameter, LinearSVC 367
 Performance 246
 Perzentil 259
 Platzhalter 97
 plot.show() 180
 polyfit() 209
 polyld() 205
 Portable Document Format (PDF) 183
 Positionsargument 90
 Prädiktor-Klasse 241
 Precision, Fehlermaß bei der Klassifikation 342
 predictor Interface, Scikit-learn 238
 preferred installer program (pip) 250
 print() 44, 52, 57, 90
 Probenschichtung
 komplexe Daten 347
 Programmierstil 51
 Prototyping 49 f.
 Psychometrie 282
 Python 30
 Einrückungen 44
 Versionen 42
 python -? 45
 Python 2.7.x 57, 77
 Python 3.4.x 57
 Python Enhancement Proposals 43
 Python-Hilfe 221

Python-Interpreter 45
 Python-Konsole 217
 Python-Version
 Unterschiede 57

Q

QtConsole 47
 Quartil 200
 quit() 45
 Quora 406

R

R 30
 R2 (R Quadrat) 361
 random() 109
 Random Forest
 Classifier 394
 optimieren 396
 Regressor 395
 Überblick 392
 RandomizedLasso-Klasse 366
 RandomizedLogistic-Klasse 366
 RandomizedPCA-Klasse 285
 range() 213
 Rangfolge
 Operatoren 83
 Rastersuche 352
 Raw NBConvert 228
 rbf (radial basis function) 380
 read() 107
 read_csv() 113
 read_table() 111
 read_sql() 119
 read_sql_query() 119
 read_sql_table() 119
 Recall, Fehlermaß in der Klassifikation 342
 Regression 205
 lineare 323
 lineare (Scikit-learn) 239
 mit SVR 382
 Regressionsanalyse 74
 Regularisierung
 L1 (Lasso) 364, 366
 L2 (Ridge) 364
 lineare Modelle 364
 nutzen 366
 Regularisierung|ElasticNet 366
 remove() 98
 reset_index() 144
 Ressource 169

Ressourcensammlung
 AnalyticBridge 408
 Aspirational Data Scientist 407
 Conductrics 406
 Data Science Central 405
 Data Science Weekly 404
 GitHub 408
 KDnuggets 404
 Quora 406
 U Climb Higher 404
 Überblick 403
 Rohdaten 33

S

sample_weight-Parameter 377
 Scalable Vector Graphics (SVG) 183
 Schiefe 260
 Schlüsselwort 94
 Schweizer-Rollen-Datensatz 306
 Scikit-learn-Bibliothek 55, 105
 SciPy 54
 search() 158
 SecretNumber() 23
 Selektion
 deterministische 245
 Sequenz 97
 set_xlim() 184
 set_xticks() 184
 set_ylim() 184
 set_yticks() 184
 sets-Modul 96
 show() 116
 Skalierung
 SVM (Support-Vector-Maschinen) 378
 Skewness *siehe* Schiefe
 Skriptsprache 43
 Spam-Detektor 331
 Spambase-Datensatz 413
 Spannweite 259
 Spearman-Korrelation 273
 Speicher-Profiler 249
 Spyder 48
 Standardabweichung 259
 Stapel 97
 last in/first out (LIFO) 97
 Statistische Businessanalyse (SAS) 30
 Stemming 153
 Stochastische Lösung
 SVM 383

Stochastischer Gradientabstiegs-Klassifizierer (SGDClassifier) 367, 385
 Stochastischer Gradientabstiegs-regressor (SGDRegressor) 385
 Stoppwort 153, 333
 str() 84, 90, 136
 StratifiedKFold-Klasse 348
 Streamen 107
 Streudiagramm 202
 zeichnen 268
 strftime() 136
 String 84
 Structured Query Language (SQL) 30, 119
 subsample-Parameter 398
 SVC (Support Vector Classifier) 252
 Klassifizieren 375
 SVD (Singular Value Decomposition) 278
 Faktor- und Hauptkomponentenanalyse 281
 Themen extrahieren 287
 Unsichtbares messen 280
 SVM (Support-Vector-Maschinen)
 Ausreißer 318
 klassifizieren mit SVC 377
 komplexe Daten 373
 neue Parameter festlegen 373
 nichtlineare Funktionen 380
 Rand 372
 stochastische Lösungen 383
 Übersicht 370
 Vorteile und Nachteile 370
 SVR (epsilon-Support-Vector-Regression) 382

T

T-Test
 nach Boxplots 265
 Tabelle
 Chi-Quadrat 273
 Term Frequency times Inverse Document Frequency-(TF-IDF)-Transformation 161
 Text extrahieren 157
 Textdatei 152
 tfidf.transform() 162
 TfidfTransformer() 162
 time() 85
 timedelta() 137
 %timeit 247

%timeit 247
 Tokenisieren 153, 158
 tolower() 151
 Trägheit 299
 transform() 140, 148
 Transformationen
 nichtlineare 357
 Trend 339
 Trendlinie 208
 TruncatedSVD-Klasse 290
 Tukey-Boxplot 313
 Tupel 97 f.

U

U Climb Higher 404
 Unausgeglichenes
 Klassenszenario 377
 Univariater Ansatz 312
 Universal Transformation Format 8-Bit (UTF-8) 152
 unstack() 131
 Unterpassung, SVM-Modell 375

V

validation_curve-Klasse 354
 Validierung
 Daten 127
 Validierungskurve 355
 values() 101
 Variable 79, 106
 Indikatorvariablen 173
 kategoriale 131 f.
 Variablentransformation 358
 Varianz 339
 einzigartige 281
 Vektor-Iterator 161
 Vektorisierung 161, 174
 Verfahrensorientiert 36
 Verknüpfung 163
 Verknüpfungsmethode
 Average 303
 Complete 303
 Ward 303
 Verlaufsprotokoll 49
 Verschachtelung 93
 Version
 Python 42
 version() 222
 Verteilung 173
 Histogramm 267
 modifizieren 274
 Normalverteilung 274

Visualization and Data Mining in
an 3D Immersive Environment
Summer Project 413
Vorhersageklasse 241

W

Wahrscheinlichkeit
Naiver Bayes 329
Ward-Verknüpfungsmethode 303
warm_start-Parameter 399
Warteschlange 97
first in/first out (FIFO) 97
Web-Service 120
Wert
boolescher 79

while-Anweisung 95
Whisker 200
WinPython 60
Winsorisieren 315
Wölbung 260
Wort
vordefiniertes 78

X

XML-Datei 150
XPath 149

Y

ylabel() 190

Z

Z-Score-Standardisierung 275
Zahl
Gleitkommazahl 79
Integer 78
komplexe 79
Zeichenkette 84
Zeitreihe 206
Zelle 71
zip() 101
Zufallssuche 356

