Inhalt

Vorwort . . . . . **IX**

1 Einführung **1**
Warum Machine Learning? 1
Welche Probleme kann Machine Learning lösen? 2
Ihre Aufgabe und Ihre Daten kennen 5
Warum Python? 5
scikit-learn 6
Installieren von scikit-learn 6
Grundlegende Bibliotheken und Werkzeuge 7
Jupyter Notebook 8
NumPy 8
SciPy 8
matplotlib 10
pandas 10
mglearn 11
Python 2 versus Python 3 12
In diesem Buch verwendete Versionen 13
Eine erste Anwendung: Klassifizieren von Iris-Spezies 14
Die Daten kennenlernen 15
Erfolg nachweisen: Trainings- und Testdaten 17
Das Wichtigste zuerst: Sichten Sie Ihre Daten 19
Ihr erstes Modell konstruieren: k-nächste-Nachbarn 21
Vorhersagen treffen 22
Evaluieren des Modells 23
Zusammenfassung und Ausblick 23

Andreas C. Müller / Sarah Guido, Einführung in Machine Learning mit Python, O'Reilly, ISBN 978-3-96009-049-6

2 Überwachtes Lernen 27
Klassifikation und Regression 27
Verallgemeinerung, Overfitting und Underfitting 28
Zusammenhang zwischen Modellkomplexität und Größe des Datensatzes 31
Algorithmen zum überwachten Lernen 32
Einige Beispieldatensätze 32
k-nächste-Nachbarn 36
Lineare Modelle 45
Naive Bayes-Klassifikatoren 66
Entscheidungsbäume 68
Ensembles von Entscheidungsbäumen 80
Support Vector Machines mit Kernel 88
Neuronale Netze (Deep Learning) 99
Schätzungen der Unsicherheit von Klassifikatoren 112
Die Entscheidungsfunktion 113
Vorhersagen von Wahrscheinlichkeiten 116
Unsicherheit bei der Klassifikation mehrerer Kategorien 118
Zusammenfassung und Ausblick 120

3 Unüberwachtes Lernen und Vorverarbeitung 123
Arten von unüberwachtem Lernen 123
Herausforderungen beim unüberwachten Lernen 124
Vorverarbeiten und Skalieren 124
Unterschiedliche Möglichkeiten der Vorverarbeitung 125
Anwenden von Datentransformationen 126
Trainings- und Testdaten in gleicher Weise skalieren 128
Die Auswirkungen der Vorverarbeitung auf überwachtes Lernen 130
Dimensionsreduktion, Extraktion von Merkmalen und Manifold Learning 132
Hauptkomponentenzerlegung (PCA) 132
Nicht-negative-Matrix-Faktorisierung (NMF) 147
Manifold Learning mit t-SNE 154
Clusteranalyse 158
k-Means-Clustering 158
Agglomeratives Clustering 169
DBSCAN 174
Vergleichen und Auswerten von Clusteralgorithmen 178
Zusammenfassung der Clustering-Methoden 192
Zusammenfassung und Ausblick 193

4 Repräsentation von Daten und Merkmalsgenerierung ... 195
Kategorische Variablen ... 196
One-Hot-Kodierung (Dummy-Variablen) ... 197
Zahlen können kategorische Daten kodieren ... 202
Binning, Diskretisierung, lineare Modelle und Bäume ... 204
Interaktionen und Polynome ... 208
Univariate nichtlineare Transformation ... 214
Automatische Auswahl von Merkmalen ... 218
Univariate Statistiken ... 218
Modellbasierte Auswahl von Merkmalen ... 221
Iterative Auswahl von Merkmalen ... 222
Berücksichtigen von Expertenwissen ... 224
Zusammenfassung und Ausblick ... 233

5 Evaluierung und Verbesserung von Modellen ... 235
Kreuzvalidierung ... 236
Kreuzvalidierung in scikit-learn ... 237
Vorteile der Kreuzvalidierung ... 238
Stratifizierte k-fache Kreuzvalidierung und andere Strategien ... 238
Gittersuche ... 244
Einfache Gittersuche ... 245
Die Gefahr des Overfittings von Parametern und Validierungsdaten ... 246
Gittersuche mit Kreuzvalidierung ... 248
Evaluationsmetriken ... 260
Das Ziel im Auge behalten ... 260
Metriken zur binären Klassifikation ... 261
Metriken zur Klassifikation mehrerer Kategorien ... 282
Regressionsmetriken ... 284
Verwenden von Metriken zur Modellauswahl ... 285
Zusammenfassung und Ausblick ... 287

6 Verkettete Algorithmen und Pipelines ... 289
Parameterauswahl mit Vorverarbeitung ... 290
Erstellen von Pipelines ... 292
Pipelines zur Gittersuche einsetzen ... 293
Die allgemeine Pipeline-Schnittstelle ... 296
Bequemes Erstellen von Pipelines mit make_pipeline ... 297
Zugriff auf Attribute von Schritten ... 298
Zugriff auf Attribute in einer Pipeline mit Gittersuche ... 299

Andreas C. Müller / Sarah Guido, Einführung in Machine Learning mit Python, O'Reilly, ISBN 978-3-96009-049-6

Gittersuche für Vorverarbeitungsschritte und Modellparameter 300
Gittersuche nach dem richtigen Modell 303
Zusammenfassung und Ausblick 304

7 Verarbeiten von Textdaten 307
Arten von als Strings repräsentierter Daten 307
Anwendungsbeispiel: Meinungsanalyse zu Filmbewertungen 309
Repräsentation von Text als Bag-of-Words 311
Anwenden von Bag-of-Words auf einen einfachen Datensatz 313
Bag-of-Words der Filmbewertungen 314
Stoppwörter 318
Umskalieren der Daten mit tf-idf 319
Untersuchen der Koeffizienten des Modells 322
Bag-of-Words mit mehr als einem Wort (n-Gramme) 323
Fortgeschrittene Tokenisierung, Stemming und Lemmatisierung ... 327
Modellierung von Themen und Clustering von Dokumenten 331
Latent Dirichlet Allocation. 331
Zusammenfassung und Ausblick 338

8 Zusammenfassung und weiterführende Ressourcen 341
Herangehensweise an eine Fragestellung beim maschinellen Lernen . 341
Der menschliche Faktor 342
Vom Prototyp zum Produktivsystem 343
Testen von Produktivsystemen 344
Konstruieren eines eigenen Estimators 344
Wie geht es von hier aus weiter? 345
Theorie 345
Andere Umgebungen und Programmpakete zum maschinellen Lernen 346
Ranking, Empfehlungssysteme und andere Arten von Lernen ... 347
Probabilistische Modellierung, Inferenz und probabilistische Programmierung 347
Neuronale Netze 348
Skalieren auf größere Datensätze 349
Verfeinern Sie Ihre Fähigkeiten 350
Schlussbemerkung 351

Index 353