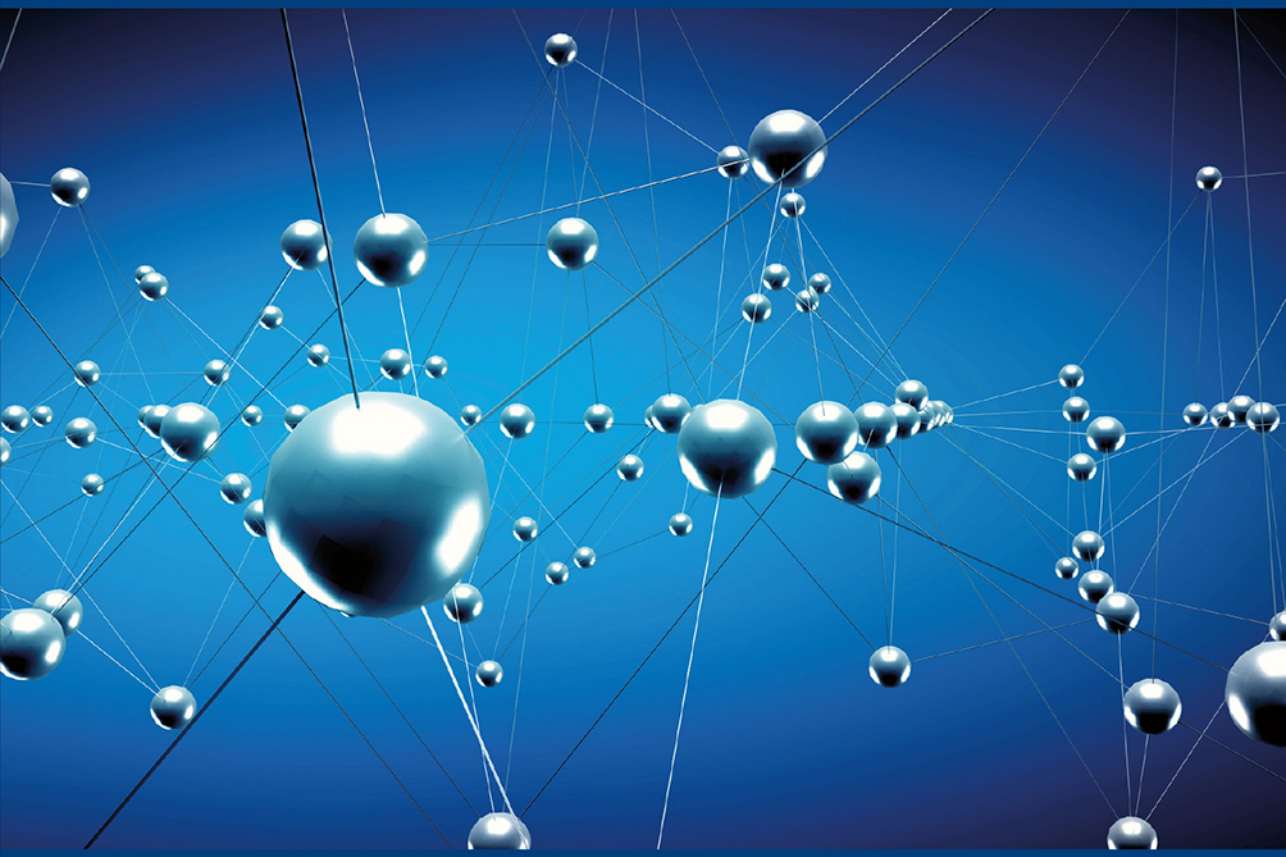
FOSTER PROVOST | TOM FAWCETT



DATA SCIENCE für UNTERNEHMEN

Data Mining und datenanalytisches Denken praktisch anwenden

mitp

Inhaltsverzeichnis

Einleitung 13

Über die Autoren 21

1 Einführung: Datenanalytisches Denken 23
1.1 Allgegenwärtige Datenerfassungsmöglichkeiten 23
1.2 Beispiel: Hurrikan Frances 25
1.3 Beispiel: Vorhersage der Kundenfluktuation 26
1.4 Data Science, Engineering und datengestützte Entscheidungsfindung 27
1.5 Datenverarbeitung und »Big Data« 31
1.6 Von Big Data 1.0 zu Big Data 2.0 32
1.7 Daten und Data-Science-Fähigkeiten als strategisches Gut 33
1.8 Datenanalytische Denkweise 36
1.9 Dieses Buch 38
1.10 Data Mining und Data Science 39
1.11 In der Chemie geht es nicht um Reagenzgläser: Data Science vs. die Aufgabe des Data Scientists 40
1.12 Zusammenfassung 41

2 Geschäftliche Aufgaben und Data-Science-Lösungen 43
2.1 Von geschäftlichen Aufgaben zum Data-Mining-Verfahren 44
2.2 Überwachte vs. unüberwachte Verfahren 49
2.3 Ergebnisse des Data Minings 51
2.4 Der Data-Mining-Prozess 52
2.4.1 Aufgabenverständnis 53
2.4.2 Datenverständnis 54
2.4.3 Datenaufbereitung 56
2.4.4 Modellbildung 57
2.4.5 Beurteilung 57
2.4.6 Einsatz 59
2.5 Auswirkungen auf das Management des Data-Science-Teams 61
2.6 Weitere Analyseverfahren und -Technologien 62
2.6.1 Statistik 63
2.6.2 Datenbankabfragen 65

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

2.6.3 Data Warehouses 66
2.6.4 Regressionsanalyse 67
2.6.5 Machine Learning und Data Mining 67
2.6.6 Geschäftliche Aufgaben durch diese Verfahren lösen 68
2.7 Zusammenfassung 70

3 Einführung in die Vorhersagemodellbildung: Von der Korrelation zur überwachten Segmentierung 71
3.1 Modelle, Induktion und Vorhersage 73
3.2 Überwachte Segmentierung 77
3.2.1 Auswahl informativer Merkmale 78
3.2.2 Beispiel: Merkmalsauswahl anhand des Informationsgewinns 86
3.2.3 Überwachte Segmentierung mit Baumstrukturmodellen . . . 92
3.3 Segmentierungen visualisieren 98
3.4 Bäume als Regelsätze 100
3.5 Wahrscheinlichkeitsabschätzung 101
3.6 Beispiel: Abwanderungsrate per Entscheidungsbaum ermitteln 104
3.7 Zusammenfassung 108

4 Ein Modell an Daten anpassen 111
4.1 Klassifizierung via mathematischer Funktionen 113
4.1.1 Lineare Diskriminanzfunktion 115
4.1.2 Optimieren der Zielfunktion 118
4.1.3 Beispiel: Extraktion einer linearen Diskriminanzfunktion aus Daten 119
4.1.4 Lineare Diskriminanzfunktionen zur Beurteilung und zum Erstellen einer Rangfolge von Instanzen 121
4.1.5 Support Vector Machines kompakt erklärt 122
4.2 Regression via mathematischer Funktionen 125
4.3 Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit und logistische »Regression« 127
4.3.1 * Logistische Regression: Technische Details 131
4.4 Beispiel: Logistische Regression vs. Entscheidungsbaumverfahren 134
4.5 Nichtlineare Funktionen, Support Vector Machines und neuronale Netze 138
4.6 Zusammenfassung 141

* Abschnitte, die im Inhaltsverzeichnis mit einem * versehen sind, enthalten mathematische oder technische Details (siehe Seite 17)

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

5 Überanpassung erkennen und vermeiden 143
5.1 Verallgemeinerungsfähigkeit 143
5.2 Überanpassung 145
5.3 Überanpassung im Detail 146
5.3.1 Zurückgehaltene Daten und Fitfunktionen 146
5.3.2 Überanpassung bei Entscheidungsbaumverfahren 149
5.3.3 Überanpassung bei mathematischen Funktionen 151
5.4 Beispiel: Überanpassung linearer Funktionen 152
5.5 * Beispiel: Nachteile der Überanpassung 156
5.6 Von der Beurteilung durch Testdatenmengen zur Kreuzvalidierung 159
5.7 Abwanderungsdaten 163
5.8 Lernkurven 165
5.9 Überanpassung vermeiden und Steuerung der Komplexität 167
5.9.1 Überanpassung von Entscheidungsbäumen vermeiden 167
5.9.2 Eine allgemeine Methode zur Vermeidung von Überanpassung 168
5.9.3 * Überanpassung bei der Parameteroptimierung vermeiden 171
5.10 Zusammenfassung 175

6 Ähnlichkeit, Nachbarn und Cluster 177
6.1 Ähnlichkeit und Distanz 178
6.2 Nächste-Nachbarn-Methoden 181
6.2.1 Beispiel: Whisky-Analyse 181
6.2.2 Nächste Nachbarn und Vorhersagemodelle 184
6.2.3 Anzahl der Nachbarn und ihre Gewichtung 187
6.2.4 Geometrische Interpretation, Überanpassung und Steuerung der Komplexität 189
6.2.5 Probleme mit Nächste-Nachbarn-Methoden 193
6.3 Ähnlichkeit und Nachbarn: Wichtige technische Details 196
6.3.1 Heterogene Merkmale 196
6.3.2 * Weitere Distanzmaße 197
6.3.3 * Zusammenfassende Funktionen: Scores der Nachbarn berechnen 200
6.4 Clustering 202
6.4.1 Beispiel: Weitere Whisky-Analysen 203
6.4.2 Hierarchisches Clustering 204
6.4.3 Nächste Nachbarn: Clustering um Zentroiden 209

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

6.4.4 Beispiel: Clustering von Wirtschaftsnachrichten 214
6.4.5 Das Ergebnis des Clusterings verstehen 218
6.4.6 * Cluster-Beschreibungen durch überwachtes Lernen erzeugen 220
6.5 Lösen von geschäftlichen Aufgaben vs. Datenerkundung 223
6.6 Zusammenfassung 226

7 Entscheidungsanalyse I: Was ist ein gutes Modell? 227
7.1 Beurteilung von Klassifizierern 228
7.1.1 Korrektklassifizierungsrate und damit verbundene Probleme 229
7.1.2 Die Wahrheitsmatrix 230
7.1.3 Klassifizierungsaufgaben mit unausgewogener Klassenverteilung 230
7.1.4 Klassifizierungsaufgaben mit unausgewogenem Kosten-Nutzen-Verhältnis 233
7.2 Verallgemeinerung über Klassifizierungen hinaus 234
7.3 Ein wichtiges analytisches Tool: Der Erwartungswert 235
7.3.1 Erwartungswerte für Klassifizierer verwenden 236
7.3.2 Erwartungswerte zur Beurteilung von Klassifizierern verwenden 238
7.4 Beurteilung, Leistung und die Folgen für Investitionen in Daten 246
7.5 Zusammenfassung 249

8 Visualisierung der Leistung von Modellen 251
8.1 Rangfolge statt Klassifizierung 252
8.2 Profitkurven 254
8.3 ROC-Diagramme und -Kurven 257
8.4 Die Fläche unter der ROC-Kurve 263
8.5 Kumulative Reaktionskurven und Lift-Kurven 263
8.6 Beispiel: Leistungsanalyse 266
8.7 Zusammenfassung 275

9 Evidenz und Wahrscheinlichkeiten 277
9.1 Beispiel: Gezielte Kundenansprache durch Onlinewerbung 277
9.2 Evidenzen probabilistisch kombinieren 280
9.2.1 Verbundwahrscheinlichkeit und Unabhängigkeit 281
9.2.2 Der Satz von Bayes 282
9.3 Anwendung des Satzes von Bayes in der Data Science 284
9.3.1 Bedingte Unabhängigkeit und naive Bayes-Klassifizierung 286

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

9.3.2 Vor- und Nachteile des naiven Bayes-Klassifizierers 288
9.4 Ein Modell für den Lift der Evidenz 290
9.5 Beispiel: Lifts der Evidenz von Facebooks-Likes 291
9.5.1 Evidenz in Aktion: Gezielte Kundenansprache durch Werbung 293
9.6 Zusammenfassung 294

10 Texte repräsentieren und auswerten 295
10.1 Die Bedeutung von Text 296
10.2 Probleme bei der Auswertung von Text 297
10.3 Repräsentierung 298
10.3.1 Das Bag-of-words-Modell 298
10.3.2 Vorkommenshäufigkeiten 299
10.3.3 Inverse Dokumenthäufigkeit 302
10.3.4 Die Kombination aus Vorkommenshäufigkeit und inverser Dokumenthäufigkeit: TFIDF 303
10.4 Beispiel: Jazzmusiker 304
10.5 * Der Zusammenhang zwischen IDF und Entropie 308
10.6 Jenseits des Bag-of-words-Modells 310
10.6.1 N-Gramme 310
10.6.2 Eigennamenerkennung 311
10.6.3 Topic Models 312
10.7 Beispiel: Auswertung von Wirtschaftsnachrichten zwecks Vorhersage von Börsenkursen 313
10.7.1 Die Aufgabe 314
10.7.2 Die Daten 316
10.7.3 Datenvorverarbeitung 319
10.7.4 Ergebnisse 320
10.8 Zusammenfassung 324

11 Entscheidungsanalyse II: Analytisches Engineering 325
11.1 Auswahl geeigneter Empfänger eines Spendenaufrufs 326
11.1.1 Erwartungswerte: Zerlegung in Teilaufgaben und Kombination der Teilergebnisse 326
11.1.2 Ein kurzer Exkurs zum Thema Auswahleffekte 328
11.2 Eine noch ausgeklügeltere Vorhersage der Kundenabwanderung 329
11.2.1 Erwartungswerte: Strukturierung einer komplizierteren geschäftlichen Aufgabe 330
11.2.2 Den Einfluss des Anreizes beurteilen 331

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

11.2.3 Von der Zerlegung eines Erwartungswerts zur Data-Science-Lösung . . . 333
11.3 Zusammenfassung . . . 336

12 Weitere Verfahren und Methoden der Data Science . . . 339
12.1 Gleichzeitiges Auftreten und Assoziationen: Zueinander passende Objekte finden . . . 340
12.1.1 Unerwartetheit messen: Lift und Leverage . . . 341
12.1.2 Beispiel: Bier und Lotterielose . . . 342
12.1.3 Assoziationen von Facebook-Likes . . . 343
12.2 Profiling: Typisches Verhalten erkennen . . . 347
12.3 Verknüpfungsvorhersagen und Kontaktempfehlungen . . . 352
12.4 Datenreduzierung, latente Informationen und Filmempfehlungen . . . 354
12.5 Bias, Varianz und Ensemblemethoden . . . 358
12.6 Datengestützte Kausalmodelle und ein Beispiel für virales Marketing . . . 362
12.7 Zusammenfassung . . . 363

13 Data Science und Geschäftsstrategie . . . 365
13.1 Datenanalytische Denkweise . . . 365
13.2 Durch Data Science Wettbewerbsvorteile erzielen . . . 368
13.3 Durch Data Science erzielte Wettbewerbsvorteile bewahren . . . 369
13.3.1 Vorteile durch historische Gegebenheiten . . . 370
13.3.2 Einzigartiges geistiges Eigentum . . . 370
13.3.3 Einzigartige immaterielle Werte . . . 371
13.3.4 Überlegene Data Scientists . . . 371
13.3.5 Überlegenes Data-Science-Management . . . 373
13.4 Gewinnung und Förderung von Data Scientists und ihren Teams . . . 375
13.5 Data-Science-Fallstudien . . . 377
13.6 Kreative Ideen von beliebigen Quellen übernehmen . . . 378
13.7 Beurteilung von Vorschlägen für Data-Science-Projekte . . . 379
13.7.1 Beispiel für einen Data-Mining-Projektvorschlag . . . 379
13.7.2 Mängel des Projektvorschlags von Big Red . . . 380
13.8 Ausgereifte Data Science . . . 382

14 Schlussfolgerungen . . . 385
14.1 Die fundamentalen Konzepte der Data Science . . . 385
14.1.1 Anwendung der fundamentalen Konzepte auf eine neue Aufgabe: Auswertung der Daten von Mobilgeräten . . . 388

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

14.1.2 Eine neue Sichtweise auf die Lösung von geschäftlichen Aufgaben . . . 391
14.2 Was Daten nicht leisten können: Der menschliche Faktor . . . 392
14.3 Privatsphäre, Ethik und Auswertung der Daten von Einzelpersonen . . . 396
14.4 Data Science: Steckt noch mehr dahinter? . . . 397
14.5 Ein letztes Beispiel: Vom Crowd-Sourcing zum Cloud-Sourcing . . . 398
14.6 Schlussworte . . . 400

A Leitfaden zur Beurteilung von Projektvorschlägen . . . 401
A.1 Aufgaben- und Datenverständnis . . . 401
A.2 Datenaufbereitung . . . 402
A.3 Modellbildung . . . 403
A.4 Beurteilung und Deployment . . . 403

B Ein weiteres Beispiel für einen Projektvorschlag . . . 405
B.1 Szenario und Projektvorschlag . . . 405
B.2 Mängel des Projektvorschlags von GGC . . . 406

Glossar . . . 409

Quellenverzeichnis . . . 415

Stichwortverzeichnis . . . 423

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Einleitung

»Data Science im Unternehmen« ist für verschiedene Lesergruppen geeignet:

- Führungskräfte und Projektmanager, die mit Data Scientists zusammenarbeiten, Data-Science-orientierte Projekte managen oder in solche Projekte investieren
- Entwickler, die Data-Science-Lösungen implementieren
- angehende Data Scientists

Dies ist weder ein Buch über Algorithmen, noch ist es ein Ersatz für ein solches Buch. Wir vermeiden ganz bewusst einen Ansatz, der sich auf Algorithmen konzentriert, denn wir sind der Meinung, dass es nur einiger weniger grundlegender Konzepte oder Prinzipien bedarf, um aus Daten nützliche Erkenntnisse zu gewinnen. Diese Konzepte dienen als *Grundlage* vieler wohlbekannter Data-Mining-Algorithmen. Sie bilden das Fundament, auf dem die Analyse datenzentrierter, unternehmensrelevanter Probleme, das Erstellen und Bewerten von Data-Science-Lösungen und die Beurteilung allgemeiner Strategien und Lösungsansätze der Data Science beruhen. Dementsprechend orientiert sich die Darstellung an diesen allgemeinen Prinzipien, nicht an bestimmten Algorithmen. Wenn es erforderlich ist, Verfahrensvorschriften detailliert zu beschreiben, verwenden wir statt einer Liste ausführlicher algorithmischer Schritte eine Kombination aus Text und Diagrammen, die unserer Ansicht nach leichter zugänglich ist.

Das Buch setzt keine besonderen mathematischen Kenntnisse voraus. Der Inhalt ist jedoch naturgemäß etwas technisch – Ziel ist es, ein echtes Verständnis von Data Science zu vermitteln, nicht nur einen generellen Überblick zu geben. Wir haben versucht, die Mathematik auf ein Minimum zu beschränken und die Darstellung so »konzeptionell« wie möglich zu gestalten.

Den Aussagen von Branchenkollegen zufolge ist das Buch von unschätzbarem Wert, um eine gute Verständigung zwischen den Managern eines Unternehmens, den Mitarbeitern in Technik/Entwicklung und den Data-Science-Teams zu erzielen. Allerdings stammt diese Beobachtung nur von einer kleinen Gruppe, daher sind wir gespannt, als wie allgemeingültig sich diese Beurteilung tatsächlich erweisen wird (siehe Kapitel 5). Unsere Idealvorstellung sieht so aus, dass jeder Data Scientist seinen Teamkollegen im geschäftlichen Bereich und in der Entwicklung dieses Buch gibt und damit gewissermaßen sagt: Wenn wir wirklich erstklassige Data-Science-Lösungen für unternehmensrelevante Probleme entwickeln und im-

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

plementieren wollen, dann müssen wir zu einem einheitlichen Verständnis dieses Themas gelangen.

Die Kollegen haben uns außerdem mitgeteilt, dass sich das Buch noch in einem ganz unvorhergesehenen Bereich als nützlich erwiesen hat: für die Vorbereitung auf Bewerbungsgespräche mit Data Scientists. Die Nachfrage nach Data-Science-Experten auf dem Arbeitsmarkt ist hoch und nimmt weiter zu. Aus diesem Grund geben sich immer mehr Stellensuchende als Data Scientists aus. Ein Bewerber für einen solchen Job sollte die Grundlagen der Data Science, die in diesem Buch präsentiert werden, unbedingt beherrschen. (Die Branchenkollegen waren erstaunt, bei wie vielen Kandidaten das nicht der Fall ist. Halb im Scherz, halb im Ernst haben wir sogar erwogen, eine Arbeit mit dem Titel »Anmerkungen zu Bewerbungsgesprächen mit Data Scientists« zu veröffentlichen.)

Unser konzeptioneller Zugang zu Data Science

Wir stellen in diesem Buch die wichtigsten grundlegenden Konzepte der Data Science vor. Einige davon dienen als Überschriften für entsprechende Abschnitte, andere ergeben sich bei der Erörterung ganz einfach aus dem Zusammenhang (und sind daher nicht unbedingt als grundlegende Konzepte zu betrachten). Diese Konzepte umfassen die Beschreibung der eigentlichen Aufgabe, den Einsatz von Data Science und die Anwendung der Ergebnisse zur Verbesserung von Entscheidungsfindungen. Sie untermauern außerdem eine Vielzahl anderer geschäftsanalytischer Methoden und Verfahren.

Die Konzepte lassen sich in drei allgemeine Kategorien unterteilen:

1. Konzepte, die zeigen, wie Data Science an die Organisation und die Wettbewerbslandschaft angepasst werden kann, inklusive verschiedener Methoden, Data-Science-Teams aufzubauen, zu strukturieren und zu fördern; wie Data Science zu Wettbewerbsvorteilen führen kann und taktische Konzepte zur praktischen Handhabung von Data-Science-Projekten.
2. Allgemeine Konzepte der Datenanalyse, die dabei helfen, geeignete Daten und angemessene Erfassungsmethoden zu erkennen. Diese Konzepte umfassen den *Data-Mining-Prozess* sowie eine Reihe verschiedener *Aufgaben des High-Level-Data-Minings*.
3. Allgemeine Konzepte zur Wissensextraktion aus Daten, die umfangreiche Data-Science-Verfahren und ihre Algorithmen unterstützen.

Eines der fundamentalen Konzepte ist beispielsweise die Erkennung der Ähnlichkeit zweier Objekte, die durch Daten beschrieben werden. Diese Fähigkeit bildet die Grundlage für verschiedene spezielle Aufgaben. Sie kann etwa direkt dazu genutzt werden, Kunden zu *finden*, die einem vorgegebenen Kunden ähnlich sind. Sie bildet den Kern verschiedener *Vorhersage*-Algorithmen, die einen Zielwert

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

abschätzen, wie z.B. der zu erwartende Ressourcenverbrauch eines Kunden oder die Wahrscheinlichkeit, mit der ein Kunde ein Angebot akzeptiert. Sie bildet außerdem die Grundlage für *Clustering*-Verfahren, bei denen Objekte anhand gemeinsamer Merkmale gruppiert werden, ohne dabei ein festes Ziel zu verfolgen. Ähnlichkeit ist die Grundlage der *Informationsgewinnung*, bei der für eine Suchanfrage relevante Dokumente oder Webseiten abgerufen werden. Und schließlich liegt sie auch vielen gängigen *Empfehlungs*-Algorithmen zugrunde. Ein auf Algorithmen konzentriertes Buch würde all diese Aufgaben womöglich in jeweils eigenen Kapiteln abhandeln, mit unterschiedlichen Bezeichnungen hantieren und die übereinstimmenden Aspekte in den Details von Algorithmen oder mathematischen Sätzen vergraben. In diesem Buch fokussieren wir uns stattdessen auf die vereinheitlichenden Konzepte und stellen bestimmte Aufgaben und Algorithmen als deren natürliche Erscheinungsform vor.

Ein weiteres Beispiel, das bei der Beurteilung der Nützlichkeit eines Musters eine wichtige Rolle spielt, ist der sogenannte *Lift*, der in der Data Science immer wieder auftritt – ein Maß dafür, wie viel verbreiteter ein Muster ist, als man vielleicht erwarten würde. Er dient dazu, völlig verschiedene Muster in unterschiedlichen Kontexten zu beurteilen. Algorithmen für gezielte Werbung werden ausgewertet, indem man den Lift berechnet, den man für die anvisierte Zielgruppe erhält. Der Lift dient zur Beurteilung der Gewichtung von Hinweisen, die für oder gegen eine Schlussfolgerung sprechen. Er gestattet es, zu ermitteln, ob ein gleichzeitiges Auftreten (eine Assoziation) von Daten wirklich von Interesse oder einfach nur auf häufiges Vorkommen zurückzuführen ist.

Wir sind der Ansicht, dass die Erklärung von Data Science anhand dieser grundlegenden Konzepte nicht nur hilfreich für den Leser ist, sondern auch die Kommunikation zwischen geschäftlichen Interessengruppen und Data Scientists vereinfacht. Sie stellt eine gemeinsame Sprache bereit und erleichtert es beiden Seiten, einander besser zu verstehen. Die gemeinsamen Konzepte führen zu intensiveren Diskussionen, die wichtige Themen aufdecken, die anderenfalls vielleicht übersehen würden.

Hinweise für Dozenten

Dieses Buch wurde erfolgreich als Lehrbuch für ein breites Spektrum von Data Science-Lehrgängen eingesetzt. Es entstand ursprünglich durch die Entwicklung von Fosters fachübergreifenden Data-Science-Kursen an der Stern School der New York University (NYU) im Herbst 2005.[1] Der Kurs wurde eigentlich für Betriebswirtschaftler und Wirtschaftsinformatiker konzipiert, wurde aber auch von Studenten vieler anderer Fächer besucht. Es ist kaum erwähnenswert, dass dieser Kurs bei

1 Natürlich ist jeder der beiden Autoren der Meinung, dass er den Großteil der Arbeit an diesem Buch geleistet hat.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Betriebswirtschaftlern und Wirtschaftsinformatikern gut ankam, da er ja eigentlich für sie gedacht war. Wirklich interessant ist, dass auch Studenten, deren Fächer sich mit Machine Learning und anderen technischen Disziplinen befassten, ihn belegten und als sehr wertvoll ansahen. Der Grund dafür scheint zumindest teilweise darin zu liegen, dass in ihren Lehrplänen außer Algorithmen andere fundamentale Prinzipien und weitere diesbezügliche Themen nicht vorhandenen waren.

An der NYU nutzen wir dieses Buch inzwischen für eine Reihe von Kursen, die in irgendeinem Zusammenhang mit Data Science stehen: den ursprünglichen Kursen für Betriebswirtschaftler und Wirtschaftsinformatiker, Grundkursen für Geschäftsanalyse im Grundstudium, den neuen Vorlesungen über Geschäftsanalyse im Hauptstudium und als Einführung für den neuen Studiengang Data Science der NYU. Darüber hinaus wird das Buch von mehr als zwanzig weiteren Universitäten in neun Ländern an Wirtschaftshochschulen, in Informatikkursen und für allgemeine Einführungen in Data Science eingesetzt (das geschah auch schon vor der eigentlichen Veröffentlichung).

Hinweis

Wir führen eine aktuelle Liste der Institute, die das Buch nutzen. Besuchen Sie `http://www.data-science-for-biz.com` und klicken Sie oben auf WHO'S USING IT.

Weitere Kenntnisse und Konzepte

Es gibt eine Vielzahl weiterer Konzepte und Kenntnisse, die einem praktisch tätigen Data Scientist neben den grundlegenden Prinzipien der Data Science bekannt sein sollten. Diese werden in den Kapiteln 1 und 2 vorgestellt. Wir empfehlen dem interessierten Leser, auch die englische Website zum Buch (`http://www.data-science-for-biz.com`) zu besuchen und sich die dortigen Hinweise zu diesen Konzepten und Kenntnissen näher anzusehen. (Dazu gehören z.B. Python-Skripte, Verarbeitung auf der Unix-Kommandozeile, Datendateien, gängige Datenformate, Datenbanken und Datenbankabfragen, Big-Data-Architekturen und Systeme wie MapReduce oder Hadoop, Datenvisualisierung und andere verwandte Themen.) Auf der Website finden Sie außerdem Lehrmaterialien wie Vorlesungsfolien, Themen für mögliche Hausaufgaben, auf dem Buch aufbauende Beispielprojekte, Prüfungsfragen und vieles andere.

Konventionen dieses Buchs

Neben vereinzelten Fußnoten enthält das Buch verschiedene Kästen. Dabei handelt es sich im Wesentlichen um erweiterte Fußnoten, die Material enthalten, das

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

wir für interessant und erwähnenswert halten, das für eine Fußnote jedoch zu umfangreich und für den Fließtext zu abschweifend ist.

Technische Details – Anmerkung zu mit einem Stern gekennzeichneten Abschnitten

Die gelegentlich erscheinenden mathematischen Details werden in separaten Abschnitten erläutert, die mit einem Stern gekennzeichnet sind. Den Überschriften dieser Abschnitte ist ein * vorangestellt, und sie werden durch einen Kasten wie diesen eingeleitet. Sie enthalten ausführlichere mathematische oder technische Details als der übrige Text, und der einleitende Kasten erläutert den Zweck. Das Buch ist so aufgebaut, dass Sie diese Abschnitte überspringen können, ohne den Faden zu verlieren. An einigen wenigen Stellen verweisen wir den Leser jedoch darauf, dass dort wichtige Details zu finden sind.

Angaben wie (Hinz und Kunz, 2003) sind ein Verweis auf einen Eintrag im Quellenverzeichnis (in diesem Fall auf einen Artikel oder ein Buch von Hinz und Kunz aus dem Jahr 2003). »Hinz und Kunz (2003)« bedeutet dasselbe. Das Quellenverzeichnis für das gesamte Buch finden Sie im Anhang.

Wir versuchen, in diesem Buch mit so wenig Mathematik wie möglich auszukommen, und die vorhandene haben wir weitgehend vereinfacht, ohne dass es zu Missverständnissen kommen kann. Für Leser mit technischen Fachkenntnissen sind einige Anmerkungen bezüglich der von uns vorgenommenen Vereinfachungen angebracht:

1. Wir verzichten auf die Sigma-(Σ) und Pi-(Π) Notation, die in Lehrbüchern üblicherweise für Summen und Produkte benutzt wird. Stattdessen verwenden wir einfache Gleichungen mit Auslassungspunkten wie diese:

   $$f(x) = w_1x_1 + w_2x_2 + \cdots + w_nx_n$$

   In den technischen, mit Sternen gekennzeichneten Abschnitten verwenden wir mitunter die Sigma- und Pi-Notation, wenn Auslassungspunkte einfach zu umständlich wären. Wir gehen davon aus, dass die Leser dieser Abschnitte mit der mathematischen Notation besser vertraut sind und dadurch nicht verwirrt werden.
2. In Statistikbüchern wird sorgfältig zwischen einem Wert und seiner Schätzung unterschieden, indem Variablen, die Abschätzungen sind, mit einem Zirkumflex versehen werden. Die tatsächliche Wahrscheinlichkeit wird typischerweise mit p und die Abschätzung mit $\hat{p}$ gekennzeichnet. Da in diesem Buch fast ausschließlich von Abschätzungen die Rede ist, verzichten wir auf diese Notation, da sie die Gleichungen nur verkomplizieren würde. Sie können davon ausge-

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

hen, dass es sich immer um Abschätzungen handelt, sofern wir nicht ausdrücklich auf etwas anderes hinweisen.

3. Wir lassen überflüssige Variablen weg und vereinfachen die Notation, wenn ihre Bedeutung aus dem Kontext heraus klar ist. Wenn wir beispielsweise Klassifizierer mathematisch betrachten, haben wir es technisch gesehen mit Entscheidungsprädikaten und Merkmalsvektoren zu tun. Formal würde das zu einer Gleichung wie dieser führen:

   $$\hat{f}_R(\mathbf{x}) = x_{\text{Alter}} \times -1 + 0.7 \times x_{\text{Saldo}} + 60$$

   Stattdessen verwenden wir die besser lesbare Gleichung:

   $$f(\mathbf{x}) = \text{Alter} \times -1 + 0.7 \times \text{Saldo} + 60$$

   Hier ist **x** ein Vektor und `Alter` und `Saldo` sind dessen Komponenten.

Wir haben uns um einheitliche Typografie bemüht und verwenden für Merkmale und Schlüsselwörter `nicht-proportionale` Schrift. Im Kapitel über Textmining bezieht sich *Ausdruck* beispielsweise auf ein Wort im Dokument, `Ausdruck` hingegen bezeichnet das entsprechende Token in den Daten.

Es gelten die folgenden typografischen Konventionen:

Neue Begriffe, Dateinamen und -erweiterungen werden *kursiv* dargestellt.

Programm-Listings sowie im Fließtext erscheinende Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter werden in `nicht-proportionaler` Schrift gedruckt.

Vom Benutzer einzugebender Text oder kontextabhängige Werte werden in *`kursiver nicht-proportionaler`* Schrift gedruckt.

Auf Websites oder an anderen Stellen auswählbare oder anklickbare Bezeichnungen, wie z.B. Menüpunkte oder Schaltflächen, werden in der Schriftart KAPITÄLCHEN gedruckt.

Im gesamten Buch finden sich Kästen mit Hinweisen und Warnungen, die womöglich unterschiedlich aussehen, je nachdem, ob sie ein auf Papier gedrucktes Buch, ein PDF oder ein E-Book lesen. Sie haben folgende Bedeutung:

Tipp

So werden Tipps und Vorschläge dargestellt.

Hinweis

Dies ist ein allgemeiner Hinweis.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Vorsicht

Warnungen erscheinen in einem Kasten wie diesem. Sie sind wichtiger als Tipps und Hinweise und werden nur sparsam eingesetzt.

Aus diesem Buch zitieren

Das Buch soll nicht nur eine Einführung in Data Science bieten, sondern auch bei der alltäglichen Arbeit nützlich sein. Die Verwendung von Zitaten oder Beispielen aus diesem Buch unter Angabe der Quelle bedarf keiner besonderen Genehmigung. Üblich sind Nennung von Titel, Autor(en), Verlag, Erscheinungsjahr und ISBN, also beispielweise *Data Science im Unternehmen* von Foster Provost und Tom Fawcett (mitp-Verlag 2017), ISBN 978-3-95845-546-7.

Danksagungen

Wir möchten den vielen Kollegen und den Menschen danken, die uns in zahlreichen Diskussionen und bei der Durchsicht der Entwürfe zu diesem Buch wertvolle Ideen, Feedback, Kritik und Vorschläge lieferten und die uns stets ermuntert haben. Trotz des Risikos, jemanden zu vergessen, möchten wird insbesondere den folgenden Personen danken:

Panos Adamopoulos, Manuel Arriaga, Josh Attenberg, Solon Barocas, Ron Bekkerman, Josh Blumenstock, Ohad Brazilay, Aaron Brick, Jessica Clark, Nitesh Chawla, Peter Devito, Vasant Dhar, Jan Ehmke, Theos Evgeniou, Justin Gapper, Tomer Geva, Daniel Gillick, Shawndra Hill, Nidhi Kathuria, Ronny Kohavi, Marios Kokkodis, Tom Lee, Philipp Marek, David Martens, Sophie Mohin, Lauren Moores, Alan Murray, Nick Nishimura, Balaji Padmanabhan, Jason Pan, Claudia Perlich, Gregory Piatetsky-Shapiro, Tom Phillips, Kevin Reilly, Maytal Saar-Tsechansky, Evan Sadler, Galit Shmueli, Roger Stein, Nick Street, Kiril Tsemekhman, Craig Vaughan, Chris Volinsky, Wally Wang, Geoff Webb, Debbie Yuster und Rong Zheng.

Darüber hinaus möchten wir den Studenten danken, die an Fosters Kursen *Data Mining for Business Analytics*, *Practical Data Science*, *Introduction to Data Science* und dem Data-Science-Forschungsseminar teilgenommen haben. Ihre Anregungen und Fragen, die bei der Durchsicht der ersten Entwürfe dieses Manuskripts aufgekommen sind, haben wertvolle Hinweise zur Verbesserung des Buchs geliefert.

Wir danken allen Kollegen, die uns all die Jahre Wissen über Data Science vermittelt haben und ebenso, wie man dieses Wissen weitergeben kann. Besonders möchten wir Maytal Saar-Tsechansky und Claudia Perlich danken. Vor einigen Jahren hat

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Maytal Foster freundlicherweise die Notizen zu ihrem Data-Mining-Kurs zur Verfügung gestellt. Das Beispiel für den Klassifizierungsbaum in Kapitel 3 (vielen Dank besonders für die Visualisierung durch »Körper«) beruht vornehmlich auf ihren Ideen, die auch den Anstoß gaben für die Visualisierung des Vergleichs der Aufteilung des Hypothesenraums durch Bäume und lineare Diskriminanzfunktionen in Kapitel 4. Auch das Beispiel »Wird David das Angebot annehmen?« in Kapitel 6 und vermutlich weitere, längst vergessene Dinge, basieren auf ihrer Arbeit. Claudia hat in den vergangenen Jahren gemeinsam mit Foster begleitende Kurse zu *Data Mining for Business Analytics* und *Introduction to Data Science* geleitet und ihn dabei vieles über Data Science (und darüber hinaus) gelehrt.

Dank an David Stillwell, Thore Graepel und Michal Kosinski für die Bereitstellung der Facebook-Like-Daten für einige der Beispiele. Dank an Nick Street für die Bereitstellung der Zellkerndaten und des Zellkernbilds in Kapitel 4. Dank an David Martens für die Hilfe bei der Visualisierung der Aufenthaltsorte. Dank an Chris Volinsky für die Bereitstellung der Daten seiner Arbeit über den Netflix-Wettbewerb. Dank an Sonny Tambe für den Zugang zu seiner Arbeit über Big-Data-Technologien und Produktivität. Dank an Patrick Perry für den Hinweis auf das Bank-Callcenter (Kapitel 12). Dank an Geoff Webb, dass wir das Assoziationsanalysesystem *Magnus Opus* benutzen durften.

Vor allem danken wir unseren Familien für ihre Zuneigung, Geduld und Ermutigung.

Beim Verfassen des Buchs kam eine Vielzahl von Open-Source-Software zum Einsatz. Die Autoren möchten den Entwicklern und Mitarbeitern folgender Projekte danken:

- Python und Perl
- Scipy, Numpy, Matplotlib und Scikit-Learn
- Weka
- dem *Machine Learning Repository* der University of California in Irvine (Bache und Lichmann, 2013)

Abschließend möchten wir die Leser noch einmal einladen, unsere englische Website `http://www.data-science-for-biz.com` zu besuchen, die Aktualisierungen, Neues, Errata und Ergänzungen zu dem im Buch vorgestellten Material enthält.

Foster Provost und Tom Fawcett

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Über die Autoren

Foster Provost ist Professor und Fakultätsmitglied an der New York University (NYU) Stern School of Business, an der er Business Analytics und Data Science lehrt und Vorlesungen über Betriebswirtschaftslehre hält. Seine preisgekrönten Forschungsarbeiten sind weltweit bekannt und werden häufig zitiert. Bevor er zur NYU wechselte, war er fünf Jahre lang als Data Scientist bei dem Unternehmen tätig, aus dem schließlich Verizon, der größte amerikanische Mobilfunkbetreiber, hervorging. In den letzten zehn Jahren hat Professor Provost verschiedene erfolgreiche Unternehmen mitbegründet, die schwerpunktmäßig Data Science einsetzen.

Tom Fawcett hat einen Doktortitel für Machine Learning und war mehr als zwei Jahrzehnte in verschiedenen Branchen (GTE Laboratories, NYNEX/Verizon Labs, HP Labs usw.) in der Forschung und Entwicklung tätig. Die von ihm veröffentlichten Arbeiten zur Methodologie (wie etwa die Beurteilung von Ergebnissen des Data Minings) und Anwendung von Data Science (z.B. Erkennung von Betrugsfällen und Spamfilter) sind zu Standardwerken geworden.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Kapitel 2

Geschäftliche Aufgaben und Data-Science-Lösungen

Fundamentale Konzepte:

- Verschiedene kanonische Data-Mining-Aufgaben
- Der Data-Mining-Prozess
- Überwachtes und unüberwachtes Data Mining

Zu den wichtigen Prinzipien der Data Science gehört, dass Data Mining ein *Prozess* mit wohlverstandenen Phasen ist. Bei einigen dieser Phasen kommt Informationstechnologie zum Einsatz, wie etwa bei der automatisierten Erkennung und Auswertung von Mustern in den Daten, während bei anderen vor allem die Kreativität des Analysten, Fachkenntnisse und gesunder Menschenverstand gefragt sind. Den gesamten Prozess zu verstehen trägt dazu bei, Data-Mining-Projekte zu strukturieren, damit sie eher einer systematischen Analyse gleichen als einem auf Glück und Scharfsinn gestützten Unterfangen.

Der Data-Mining-Prozess teilt die eigentliche Aufgabe, Muster in den Daten zu finden, in mehrere klar definierte Unteraufgaben auf, und ist deswegen auch für die Strukturierung von Diskussionen über Data Science nützlich. In diesem Buch werden wir diesen Prozess als übergreifenden Handlungsrahmen für unsere Erläuterungen nutzen. Dieses Kapitel ist eine Einleitung in den Data-Mining-Prozess, zunächst aber stellen wir diesen in einen geeigneten Zusammenhang, indem wir einige gängige Data-Mining-Aufgaben betrachten. Dadurch können wir bei den Erläuterungen zum Prozess als Ganzem und bei der Vorstellung weiterer Konzepte in den nachfolgenden Kapiteln konkreter werden.

Am Ende dieses Kapitels kommen wir kurz auf einige wichtige Themen der Geschäftsanalyse zu sprechen, die eigentlich nicht Gegenstand dieses Buchs sind (für die es aber umfangreiche Literatur gibt), wie Datenbanken, Datenhaltung und elementare Statistik.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

2.1 Von geschäftlichen Aufgaben zum Data-Mining-Verfahren

Jede datengestützte Aufgabe zur Entscheidungsfindung in einem Unternehmen ist einzigartig und bildet eine eigene Kombination aus Zielen, Wünschen und Einschränkungen – und manchmal spielen sogar Persönlichkeiten eine Rolle. Wie so oft beim Engineering, sind den geschäftlichen Aufgaben bestimmte Anforderungen gemeinsam. In Zusammenarbeit mit den Projektleitern der jeweiligen Firma zerlegt der Data Scientist die geschäftliche Aufgabe in Teilaufgaben. Die Lösungen der Teilaufgaben können dann schließlich zur Lösung der Hauptaufgabe zusammengesetzt werden. Manche dieser Teilaufgaben sind einzigartig, andere sind beim Data Mining aber durchaus gängig. Die Vorhersage der Kundenabwanderung ist beispielsweise einzigartig für MegaTelCo: Die Aufgabe besitzt Aspekte, die sich von den Vorhersagen der Kundenfluktuation aller anderen Telekommunikationsunternehmen unterscheiden. Die Teilaufgabe, anhand der vorhandenen Daten die Wahrscheinlichkeit abzuschätzen, ob ein bestimmter Kunde nach Ablauf seines Vertrags abwandert, wird wahrscheinlich Bestandteil aller Vorhersagen der Abwanderung sein. Nachdem die spezifischen Daten von MegaTelCo in ein bestimmtes Format (das im nächsten Kapitel erläutert wird) konvertiert worden sind, ist diese Wahrscheinlichkeitsabschätzung bestens für ein sehr gebräuchliches Data-Mining-Verfahren geeignet. Wir wissen sowohl praktisch als auch theoretisch sehr gut Bescheid über die Lösung gängiger Data-Mining-Aufgaben. In den nachfolgenden Kapiteln werden wir Handlungsrahmen vorstellen, die bei der Aufteilung geschäftlicher Aufgaben in Teilaufgaben und der Zusammensetzung der Teillösungen hilfreich sind.

Tipp

In der Data Science ist es wichtig, datenanalytische Aufgaben in Teilaufgaben zerlegen zu können, die sich durch gebräuchliche Data-Mining-Verfahren lösen lassen, für die Tools verfügbar sind. Das Erkennen gängiger Aufgaben verhindert, dass man das Rad neu erfindet und so Zeit und Ressourcen verschwendet. Außerdem wird es dadurch möglich, sich auf interessantere Teile des Prozesses zu konzentrieren, die ein menschliches Eingreifen erforderlich machen – noch nicht automatisierte Teilaufgaben, bei denen menschliche Kreativität und Intelligenz ins Spiel kommen müssen.

Trotz der Vielzahl der Data-Mining-Algorithmen, die im Laufe der Jahre entwickelt wurden, gibt es nur eine Handvoll grundsätzlich unterschiedlicher Aufgabentypen, die sich mit diesen Algorithmen lösen lassen. Es lohnt sich, diese Aufgaben eindeutig zu definieren. In den nächsten Kapiteln werden wir die ersten beiden (Klassifizierung und Regression) nutzen, um einige der fundamentalen Konzepte

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

zu veranschaulichen. Im Folgenden bezeichnet der Begriff »Individuum« sowohl natürliche Personen, wie Kunden oder Verbraucher, als auch juristische Personen, wie z. B. ein Unternehmen. Wir kommen in Kapitel 3 ausführlicher darauf zurück. Bei vielen Analyseprojekten suchen wir nach »Korrelationen« zwischen Variablen, die ein Individuum beschreiben, und anderen Variablen. Wir könnten den Daten beispielsweise entnehmen, welche Kunden abgewandert sind, nachdem die Laufzeit ihres Vertrags endete. Wir könnten dann versuchen herauszufinden, welche anderen Variablen mit einem Kunden korrelieren, der in naher Zukunft abwandern wird. Das Aufspüren solcher Korrelationen ist das einfachste Beispiel für eine Klassifizierung bzw. eine Regression.

1. *Klassifizierung* und *Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit* versuchen, für ein zur Grundgesamtheit gehörendes Individuum vorherzusagen, zu welchen (einigen wenigen) Klassen dieses Individuum gehört. Für gewöhnlich schließt die Zugehörigkeit zu einer Klasse die Zugehörigkeit zu einer anderen aus. Eine Frage, die der Klassifizierung dient, ist beispielsweise: *»Welche der MegaTelCo-Kunden werden ein bestimmtes Angebot wahrscheinlich annehmen?«* In diesem Beispiel könnte man die beiden Klassen als `wird annehmen` und `wird nicht annehmen` bezeichnen.

   Für die Klassifizierung liefert ein Data-Mining-Verfahren ein Modell, das bei Auftauchen eines neuen Individuums ermittelt, zu welcher Klasse dieses Individuum gehört. *Scoring* und *Wahrscheinlichkeitsabschätzungen* der Klassenzugehörigkeit sind sehr ähnliche Aufgaben. Ein auf ein Individuum angewendetes Scoring-Modell ermittelt statt einer Klassenvorhersage eine Punktzahl, die die Wahrscheinlichkeit (oder ein anderes Wahrscheinlichkeitsmaß) dafür angibt, dass ein Individuum zu einer bestimmten Klasse gehört. In unserem Szenario der Angebotsannahme würde ein Scoring-Modell die einzelnen Kunden auswerten und eine Punktzahl liefern, die angibt, wie wahrscheinlich es ist, dass ein Kunde auf das Angebot eingeht. Klassifizierung und Scoring sind eng miteinander verwandt – ein Modell, das die eine Aufgabe erledigen kann, lässt sich für gewöhnlich modifizieren, um auch die andere zu erledigen.

2. Eine *Regression* (Abschätzung des Werts) versucht, für jedes Individuum den numerischen Wert einer Variable abzuschätzen oder vorherzusagen. *»Wie oft wird ein bestimmter Kunde den Dienst nutzen?«* ist ein Beispiel für eine Frage, die durch eine Regression beantwortet werden kann. Die hier vorherzusagende Eigenschaft (die Variable) ist die Dienstnutzung. Man könnte dafür ein Modell entwickeln, indem man die Dienstnutzung vergleichbarer Individuen der Grundgesamtheit betrachtet. Eine Regression liefert ein Modell, das den Wert einer Variable für ein bestimmtes Individuum abschätzt.

   Regression und Klassifizierung sind verwandt, es gibt jedoch Unterschiede. Zwanglos könnte man sagen, dass eine Klassifizierung vorhersagt, *ob* etwas der

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Fall ist, eine Regression hingegen sagt voraus, *in welchem Ausmaß* dies eintritt. Der Unterschied wird im weiteren Verlauf des Buchs noch deutlich werden.

3. Ein *Ähnlichkeitstest* versucht, anhand der bekannten Daten einander ähnliche Individuen zu finden und kann unmittelbar dazu genutzt werden, ähnliche Objekte zu identifizieren. Beispielsweise ist IBM daran interessiert, Firmen zu finden, die ihren besten Geschäftskunden ähnlich sind. So wird gewährleistet, dass sich die Vertriebsmitarbeiter ausschließlich auf diese Kunden, die die besten Verkaufschancen versprechen, konzentrieren können. Die Ähnlichkeitstests werden anhand von Daten durchgeführt, die bestimmte Merkmale von Unternehmen beschreiben. Ähnlichkeitstests bilden die Grundlage einer der verbreitetsten Methoden für Produktempfehlungen (Leute zu finden, die ähnliche Produkte wie Sie mögen oder gekauft haben). Ähnlichkeitstest kommen auch bei vielen anderen Data-Mining-Tasks zum Einsatz, wie bei der Klassifizierung, der Regression und beim Clustering. In Kapitel 6 werden wir Ähnlichkeit und die verschiedenen Anwendungen ausführlich erörtern.
4. Beim *Clustering* wird versucht, die Individuen einer Grundgesamtheit anhand ihrer Ähnlichkeit zu gruppieren, ohne damit ein bestimmtes Ziel zu verfolgen. *»Lassen sich unsere Kunden in natürliche Gruppen oder Geschäftsfelder einteilen?«* ist eine Frage, die durch Clustering beantwortet werden kann. Bei der vorbereitenden Erkundung neuer Geschäftsbereiche erweist sich Clustering als nützlich, um zu ermitteln, ob es natürliche Gruppen gibt, weil diese Gruppen ihrerseits andere Data-Mining-Verfahren oder -Ansätze nahelegen. Clustering dient auch als Grundlage für Entscheidungsfindungen, die Fragen wie die folgenden betreffen: *»Welche Produkte sollten wir anbieten oder entwickeln? Wie sollte unser Kundendienst oder Vertriebsteam aufgebaut sein?«* In Kapitel 6 werden wir Clustering eingehender betrachten.
5. Die *Gruppierung nach gleichzeitigem Auftreten* (manchmal auch als Verbundkaufanalyse, Suche nach Assoziationsregeln oder Warenkorbanalyse bezeichnet) dient dazu, Zusammenhänge zwischen Individuen zu finden, die an denselben Transaktionen beteiligt sind. Die dazugehörige Frage könnte lauten: *»Welche Produkte werden häufig zusammen gekauft?«* Beim Clustering wird anhand der Merkmale von Objekten nach Ähnlichkeiten gesucht, bei der Gruppierung nach gleichzeitigem Auftreten dagegen anhand des gemeinsamen Vorkommens in Transaktionen. So könnte beispielsweise die Analyse der Einkäufe in einem Supermarkt ergeben, dass Hackfleisch und Chilisoße viel häufiger zusammen gekauft werden als angenommen. Um zu entscheiden, wie man auf diese Entdeckung reagieren sollte, ist ein wenig Kreativität erforderlich. Vielleicht könnte man eine gezielte Werbeaktion veranstalten, eine gemeinsame Warenpräsentation beider Produkte oder ein Kombiangebot anbieten. Das gleichzeitige Auftreten von Produkten bei Käufen ist eine gängige Art der Gruppierung und wird auch als Warenkorbanalyse bezeichnet. Einige Empfehlungssysteme führen auch eine Art Affinitätsgruppierung durch, indem sie beispielsweise nach

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Büchern suchen, die oft zusammen gekauft werden (»*Kunden, die dieses Buch gekauft haben, kauften auch …*«).

Das Ergebnis der Gruppierung nach gleichzeitigem Auftreten ist eine Beschreibung der Objekte, die gemeinsam vorkommen. Sie enthält für gewöhnlich eine Statistik über die Häufigkeit gleichzeitigen Auftretens und eine Einschätzung darüber, wie wahrscheinlich dieses ist.

6. Das *Profiling* (auch Beschreibung des Marktverhaltens genannt) versucht, das typische Verhalten eines Individuums, einer Gruppe oder der Grundgesamtheit zu charakterisieren. Ein Beispiel für eine Profiling-Frage wäre etwa: »*Wie häufig wird das Mobiltelefon von dieser Kundengruppe genutzt?*« Das Verhalten lässt sich nicht immer leicht beschreiben – ein Nutzungsprofil zu erstellen, könnte eine komplizierte Protokollierung der Nutzung des Mobilfunknetzes in der Nacht, an Wochenenden oder im Ausland, von Roaming-Gebühren und SMS usw. erforderlich machen. Das Verhalten kann im Allgemeinen für die Grundgesamtheit, für einzelne Gruppen oder sogar für einzelne Individuen beschrieben werden.

   Profiling kommt häufig zum Einsatz, um Verhaltensnormen für Anwendungen zur Erkennung von Anomalien aufzustellen, wie z. B. Betrügereien oder das Eindringen in Computersysteme (wenn sich beispielsweise jemand in Ihr iTunes-Konto hackt). Wenn etwa bekannt ist, welche Art Einkäufe jemand für gewöhnlich per Kreditkarte bezahlt, können wir feststellen, ob eine bestimmte Abbuchung zu dem dazugehörigen Profil passt oder nicht. Wir können das Ausmaß der Abweichung als Bewertung des Verdachts nutzen und gegebenenfalls Alarm schlagen, wenn der Wert zu hoch ist.

7. Die *Verknüpfungsvorhersage* versucht, Verbindungen zwischen Datenobjekten zu entdecken und gibt bei Vorhandensein einer Verbindung für gewöhnlich ein Maß für deren Stärke an. Verknüpfungsvorhersagen sind in sozialen Netzwerken üblich: »*Sie und Karen haben 10 gemeinsame Freunde. Möchten Sie Karens Freund werden?*« Die Verknüpfungsvorhersage kann außerdem die Stärke einer Verbindung abschätzen. Um beispielsweise Kunden Filme zu empfehlen, könnte man einen Graphen mit Kunden und den von ihnen gesehenen oder bewerteten Filmen verwenden. Im Graphen suchen wir dann nach *nicht* vorhandenen Verbindungen zwischen Kunden und Filmen, die der Vorhersage nach jedoch vorhanden und stark sein sollten. Diese Verknüpfungen bilden die Grundlage für Filmempfehlungen.

8. Bei der *Datenreduzierung* versucht man, eine größere Datenmenge durch eine kleinere zu ersetzen, die jedoch die wichtigsten Informationen der größeren Datenmenge enthält. Die kleinere Datenmenge lässt sich oft einfacher handhaben oder leichter verarbeiten. Zudem sind die wichtigen Informationen in einer kleinen Datenmenge womöglich besser erkennbar. Eine umfangreiche Datenmenge mit den Filmvorlieben von Zuschauern könnte beispielsweise auf eine

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

sehr viel kleinere Datenmenge reduziert werden und so die in den Daten latent vorhandenen Filmvorlieben (z. B. für verschiedene Filmgenres) offenbaren. Mit einer Datenreduzierung geht für gewöhnlich ein Informationsverlust einher – entscheidend ist hier, dass Sie dafür einen besseren Einblick in die Daten bekommen.

9. Mithilfe der *Kausalmodellbildung* versuchen wir zu verstehen, welche Ereignisse oder Aktionen andere *beeinflussen*. Nehmen wir beispielsweise an, dass wir ein Vorhersagemodell einsetzen, um Kunden gezielt durch Werbung anzusprechen und feststellen, dass die Zielgruppe tatsächlich häufiger kauft als zuvor. Hat die Werbung die Kunden tatsächlich zum Kauf veranlasst? Oder hat das Vorhersagemodell einfach nur genau die richtigen Kunden ausgewählt, die ohnehin gekauft hätten? Zu den bei der Kausalmodellbildung eingesetzten Verfahren gehören auch solche, die beträchtliche Investitionen in die Daten erfordern, wie etwa kontrollierte randomisierte Experimente (sogenannte A/B-Tests) sowie ausgeklügelte Methoden, um aus den Beobachtungsdaten auf ursächliche Zusammenhänge zu schließen. Sowohl die experimentellen als auch die beobachtenden Methoden der Kausalmodellbildung sind im Allgemeinen als »kontrafaktische« Analyse zu betrachten: Sie versuchen, den Unterschied zwischen zwei Situationen zu erfassen – die nicht gleichzeitig eintreten können –, in denen ein Ereignis (z.B. einer bestimmten Person eine Werbung zu zeigen) stattgefunden oder eben nicht stattgefunden hat.

   Ein sorgfältiger Data Scientist sollte bei jeder kausalen Schlussfolgerung immer auch die genauen Voraussetzungen berücksichtigen, die gegeben sein müssen, damit die Schlussfolgerung auch wirklich zutrifft (solche Voraussetzungen gibt es immer – fragen Sie danach). Ein Unternehmen, das Kausalmodelle einsetzt, muss einen Kompromiss finden zwischen den höheren Investitionen, die zur Reduzierung der zu Grunde gelegten Voraussetzungen erforderlich sind, und der Entscheidung, ob die Schlussfolgerungen in Anbetracht der gegebenen Voraussetzungen hinreichend verlässlich sind. Auch bei äußerst sorgfältig durchgeführten randomisierten kontrollierten Experimenten können Annahmen einfließen, die zu ungültigen Schlussfolgerungen führen. Die Entdeckung des medizinischen Placeboeffekts ist ein bekanntes Beispiel dafür, wie leicht bei sorgfältig geplanten randomisierten Experimenten Voraussetzungen übersehen werden können.

Eine ausführliche Erörterung all dieser Aufgaben würde mehrere Bücher füllen. In diesem Buch stellen wir die wichtigsten Prinzipien der Data Science vor – Prinzipien, die all diesen Aufgaben gemeinsam zugrunde liegen. Wir werden vornehmlich Klassifizierung, Regression, Ähnlichkeitstests und Clustering verwenden, um diese Prinzipien zu veranschaulichen. Gegen Ende des Buchs kommen weitere Verfahren zur Sprache, die wichtige Aspekte der grundlegenden Prinzipien verdeutlichen.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Welche dieser Aufgabentypen erscheint Ihnen für die Vorhersage der Kundenabwanderung geeignet? In der Praxis wird die Vorhersage der Kundenabwanderung oft als die Aufgabe betrachtet, *Kundensegmente* zu finden, in denen eine Abwanderung mehr oder weniger wahrscheinlich ist. Diese Segmentierungsaufgabe sieht doch sehr nach einer Klassifizierung aus, vielleicht wären auch Regression oder sogar Clustering zur Lösung geeignet. Um den besten Ansatz zu finden, müssen wir zunächst einige wichtige Unterscheidungen vorstellen.

2.2 Überwachte vs. unüberwachte Verfahren

Betrachten Sie zwei ganz ähnliche Fragen, die wir über die Grundgesamtheit der Kunden stellen könnten. Die erste lautet: *»Lassen sich die Kunden auf naheliegende Weise in verschiedene Gruppen einteilen?«* Für eine solche Gruppierung ist kein bestimmter Zweck oder kein *Ziel* vorgegeben. Wenn es ein solches Ziel nicht gibt, spricht man von einer *unüberwachten* Data-Mining-Aufgabe. Vergleichen Sie das mit einer leicht abgeänderten Frage: *»Gibt es Kundengruppen, die nach Ablauf ihrer Verträge mit einer besonders hohen Wahrscheinlichkeit abwandern?«* Hier ist ein Ziel angegeben: Wird ein Kunde nach Ablauf seines Vertrags abwandern? In diesem Fall wird die Segmentierung aus einem bestimmten Grund durchgeführt: um anhand der Abwanderungswahrscheinlichkeit etwas unternehmen zu können. Hierbei handelt es sich um eine *überwachte* Data-Mining-Aufgabe.

Hinweis

Eine Anmerkung zu den Begriffen *überwachtes* und *unüberwachtes Lernen*

Die Begriffe *überwacht* und *unüberwacht* entstammen dem Fachgebiet des Machine Learnings. Im übertragenen Sinn »überwacht« ein Lehrer den Lernenden, indem er ihm vorsichtige Hinweise auf das Lernziel sowie einige Beispiele gibt. Beim unüberwachten Lernen könnten dieselben Beispiele vorhanden sein, es gibt jedoch keine Hinweise auf das Lernziel. Der Lernende wird nicht über den Zweck des Lernens informiert, sondern muss seine eigenen Schlüsse ziehen, was die Beispiele gemeinsam haben.

Der Unterschied zwischen den beiden Fragen ist zwar nicht groß, aber bedeutsam. Wenn ein Ziel angegeben ist, kann die Aufgabe als überwachter Task formuliert werden. Überwachte Tasks erfordern andere Verfahren als unüberwachte und liefern oft sehr viel nützlichere Ergebnisse. Einem überwachten Verfahren wird ein bestimmter Zweck der Gruppierung vorgegeben – die Vorhersage des Ziels. Clustering ist ein unüberwachtes Verfahren und nimmt eine Gruppierung anhand von Ähnlichkeiten vor, dabei ist jedoch nicht gewährleistet, dass diese Ähnlichkeiten aussagekräftig oder für irgendeinen Zweck nützlich sind.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Rein technisch betrachtet muss für überwachtes Data Mining eine weitere Bedingung erfüllt sein: Es müssen Daten über das Ziel verfügbar sein. Es reicht nicht aus, dass Informationen über das Ziel prinzipiell existieren – sie müssen in den Daten auch vorhanden sein. Es könnte beispielsweise nützlich sein, zu wissen, ob ein bestimmter Kunde mindestens sechs Monate bleibt. Wenn diese Information jedoch in den Verlaufsdaten nicht oder nur unvollständig (z.B. weil die Daten nur zwei Monate lang gespeichert werden) vorhanden ist, kann der Zielwert nicht ermittelt werden. Daten über das Ziel zu sammeln, ist oft eine entscheidende Investition in der Data Science. Der Wert oder Inhalt der Zielvariablen eines Individuums wird auch als sein *Label* bezeichnet, wodurch hervorgehoben wird, dass häufig (nicht immer) Aufwand zur Kennzeichnung der Daten erforderlich ist.

Klassifizierung, Regression und Kausalmodellbildung werden im Allgemeinen durch überwachte Verfahren realisiert. Für Ähnlichkeitstests, Verknüpfungsvorhersagen und Datenreduzierung können beide Verfahren eingesetzt werden. Bei Clustering, Gruppierung nach gleichzeitigem Auftreten und Profiling werden üblicherweise unüberwachte Verfahren durchgeführt. All diesen Verfahren liegen die fundamentalen Prinzipien des Data Minings zugrunde, die wir vorstellen werden.

Die beiden wichtigsten Unterklassen des überwachten Data Minings, Klassifizierung und Regression, unterscheiden sich durch die Art des Ziels. Bei der Regression ist die Zielvariable numerisch, bei der Klassifizierung hingegen geht es um die Zuordnung zu einem (oft binären) kategorialen Ziel. Betrachten Sie die folgenden drei ähnlichen Fragen, die wir mithilfe des überwachten Data Minings in Angriff nehmen könnten:

- *Wird der Kunde auf ein Angebot für den Dienst D1 eingehen, wenn er den Anreiz A erhält?*

  Hierbei handelt es sich um eine Klassifizierung, weil das Ziel binär ist (entweder geht der Kunde auf das Angebot ein oder nicht).

- *Welches Dienstpaket (D1, D2 oder keines) wird der Kunde wahrscheinlich kaufen, wenn er den Anreiz A erhält?*

  Hierbei handelt es sich ebenfalls um eine Klassifizierung, allerdings mit einer Zielvariablen, die drei Werte annehmen kann.

- *Wie häufig wird dieser Kunde den Dienst nutzen?*

  Dies ist ein klassisches Beispiel für eine Regression, denn die Zielvariable ist numerisch.

Bei diesen Fragen gibt es einige Feinheiten, auf die wir hinweisen müssen. Bei geschäftlichen Anwendungen ist oft eine numerische *Vorhersage* für kategoriale Ziele erwünscht. In unserem Beispiel der Kundenabwanderung ist eine einfache

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Ja/Nein-Vorhersage, ob ein Kunde bleibt, womöglich nicht ausreichend. Wir möchten vielmehr die *Wahrscheinlichkeit* ermitteln, dass der Kunde bleibt. Ein solches Modell wird nach wie vor als Klassifizierung eingestuft, nicht als Regression, denn die Zielvariable ist kategorial. Dort, wo es der Deutlichkeit halber erforderlich ist, bezeichnen wir dieses Modell als »Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit«.

Zu den ersten Schritten eines Data-Mining-Prozesses gehört zum einen, festzulegen, ob das Verfahren überwacht oder unüberwacht durchgeführt wird, und zum anderen – sofern überwacht –, die Zielvariable präzise zu definieren. Bei dieser Variable muss es sich um einen bestimmten Wert handeln, der im Fokus des Data Minings steht (und für den wir anhand einiger Beispieldaten Werte erhalten können). Wir kommen in Kapitel 3 auf dieses Thema zurück.

2.3 Ergebnisse des Data Minings

Beim Data Mining ist eine weitere wichtige Unterscheidung zu beachten: die Trennung zwischen der *Untersuchung* der Daten, um Muster aufzuspüren und Modelle zu entwickeln, und der *Anwendung* der Ergebnisse des Data Minings. Die Data-Science-Studenten verwechseln diese beiden Prozesse oft miteinander, und Manager bringen sie durcheinander, wenn sie sich über Business Analytics unterhalten. Die Anwendung der Ergebnisse des Data Minings sollte auf den eigentlichen Data-Mining-Prozess Einfluss nehmen und ihm zugute kommen, die beiden Prozesse sollten aber klar voneinander getrennt werden.

Betrachten Sie bei unserem Abwanderungsbeispiel ein Anwendungsszenario, in dem die Ergebnisse genutzt werden. Wir möchten unser Modell einsetzen, um vorherzusagen, welche Kunden abwandern werden. Nehmen Sie für diesen Fall an, dass mithilfe des Data Minings ein Modell M zur Wahrscheinlichkeitsabschätzung der Klassenzugehörigkeit entwickelt wurde. Jeder Kunde wird mithilfe bestimmter charakteristischer Merkmale beschrieben. M nimmt diese verschiedenen charakteristischen Merkmale als Dateneingabe entgegen und erzeugt eine Punktzahl oder eine Abschätzung der Abwanderungswahrscheinlichkeit. Hierbei handelt es sich um die *Anwendung* der Ergebnisse des Data Minings. Das Data Mining erzeugt das Modell M anhand anderer, oft älterer Daten.

In Abbildung 2.1 sind diese beiden Phasen dargestellt. Das Data Mining liefert, wie in der oberen Hälfte der Abbildung dargestellt, ein Modell für die Wahrscheinlichkeitsabschätzung. In der Anwendungsphase (untere Hälfte) wird das Modell auf einen neuen, unbekannten Fall angewendet und erzeugt eine Wahrscheinlichkeitsabschätzung dafür.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

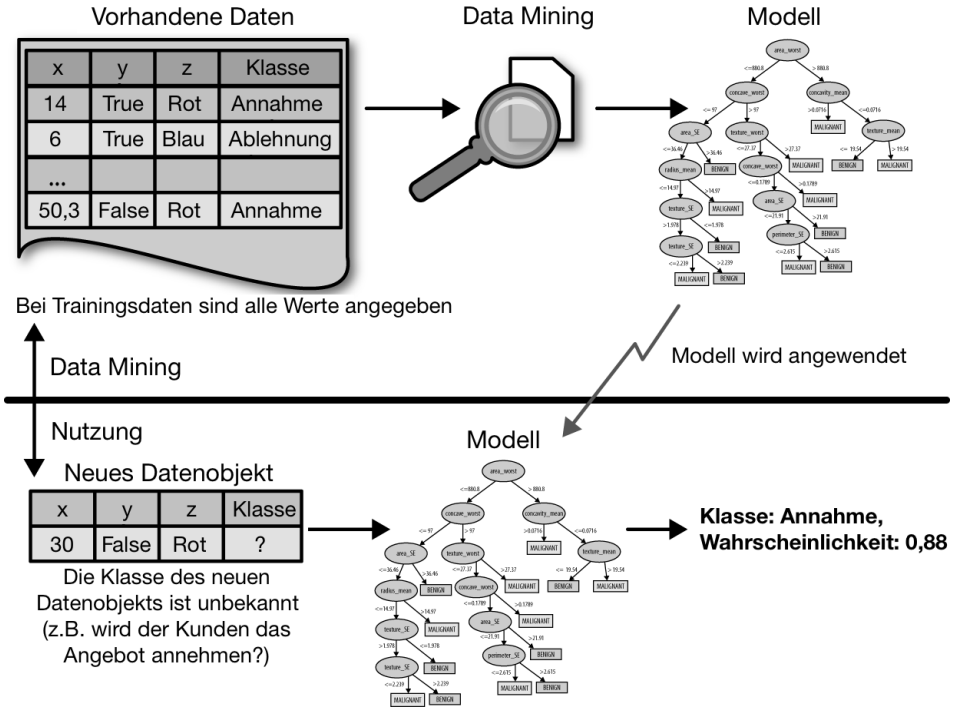


Abb. 2.1: Data Mining und das Ergebnis des Data Minings. In der oberen Hälfte der Abbildung ist das Data Mining vorhandener Daten zur Erzeugung des Modells dargestellt. Hier ist wichtig zu beachten, dass für die vorhandenen Daten ein Ziel (»Klasse«) angegeben wurde. Die untere Hälfte zeigt die Anwendung des Resultats des Data Minings. Das Modell wird auf neue Daten angewendet, deren Klassenzugehörigkeit unbekannt ist. Das Modell sagt sowohl die Klassenzugehörigkeit als auch die Wahrscheinlichkeit dafür voraus, dass die Klassenvariable diesen Wert annimmt.

2.4 Der Data-Mining-Prozess

Data Mining ist eine wahre Kunstfertigkeit. Es erfordert die Anwendung beträchtlicher wissenschaftlicher und technologischer Kenntnisse, jedoch auch ein gewisses künstlerisches Geschick. Ebenso wie in anderen ausgereiften Disziplinen verfügt das Data Mining über ein wohlverstandenes Verfahren für die Strukturierung von Aufgaben, das Einheitlichkeit, Reproduzierbarkeit und Objektivität ermöglicht. Eine nützliche Festschreibung ist das Standard-Prozess-Modell für Data Mining (*Cross Industry Standard Process for Data Mining*, CRISP-DM, Shearer, 2000), das in Abbildung 2.2 dargestellt ist.[1]

1 Siehe auch den englischsprachigen Wikipedia-Eintrag zum Thema CRISP-DM (`https://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining`)

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

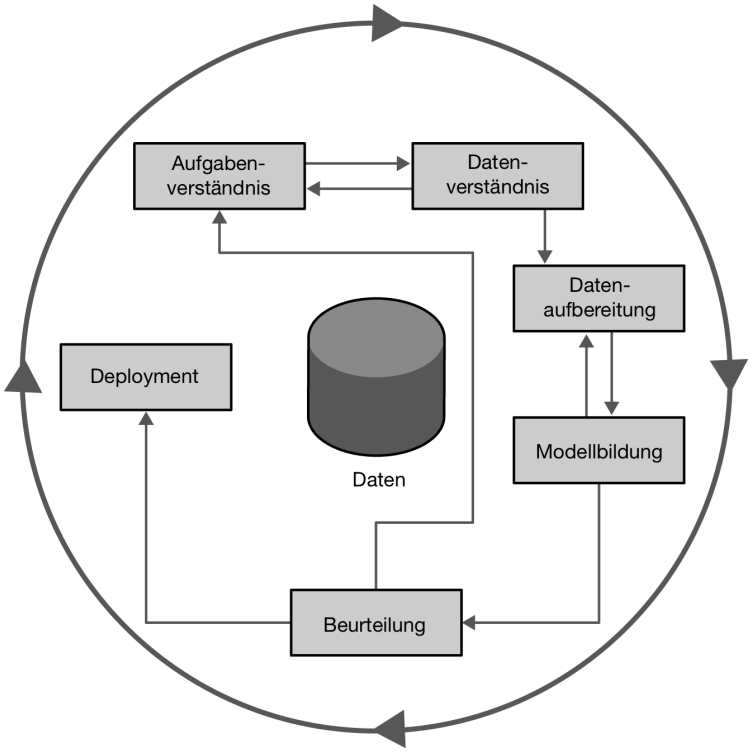


Abb. 2.2: Der CRISP-Data-Mining-Prozess

Das Diagramm verdeutlicht die Tatsache, dass Iteration keine Ausnahme, sondern die Regel ist. Den Prozess einmalig zu durchlaufen, ohne dabei die Aufgabe erledigt zu haben, ist also im Allgemeinen kein Fehlschlag. Oftmals dient der gesamte Prozess der Erkundung der Daten: Nach dem ersten Durchlauf weiß das Data-Science-Team erheblich mehr und kann bei der zweiten Iteration gezielter vorgehen. Sehen wir uns die einzelnen Schritte etwas genauer an.

2.4.1 Aufgabenverständnis

Zunächst einmal muss man wissen, welche Aufgabe gelöst werden soll. Man könnte meinen, eine Aufgabe sei sicher offensichtlich, aber bei geschäftlichen Aufgaben ist nur selten eine greifbare, klare und eindeutige Data-Mining-Aufgabenstellung zu erkennen. Oftmals muss die Aufgabenstellung umformuliert werden, sodass die Entwicklung einer Lösung zu einem iterativen Entdeckungsvorgang wird. Das Diagramm (Abbildung 2.2) stellt diesen Vorgang als einen Kreislauf innerhalb eines Kreislaufs dar, nicht als einfachen linearen Prozess. Die anfängliche Formulierung der Aufgabenstellung ist womöglich unvollständig oder nicht optimal, sodass mehrere Iterationen erforderlich sein können, bis sich eine akzeptable Formulierung ergibt.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Bei der Formulierung der Aufgabe spielt die Kreativität des Analysten eine große Rolle. Die Data Science kann dazu ihren Beitrag leisten, wie wir noch erläutern werden. Doch der Schlüssel zum Erfolg ist oftmals eine kreative Formulierung der Anforderungen durch einen Analysten, der dabei berücksichtigt, wie die Aufgabenstellung auf eine oder mehrere Data-Science-Aufgaben verteilt werden kann. Kreativen Analysten, die hervorragende Kenntnisse der Grundlagen besitzen, fällt es leichter, neue Ansätze zu entdecken.

Uns stehen leistungsfähige Tools für die Lösung bestimmter Data-Mining-Tasks zur Verfügung, die wir in Abschnitt 2.1 *Von geschäftlichen Aufgaben zum Data-Mining-Verfahren* beschrieben haben. Bei den ersten Versuchen wird typischerweise ein Lösungsansatz entwickelt, der sich diese Tools zunutze macht. Das kann beinhalten, dass die Aufgabenstellung so strukturiert (bzw. organisiert) wird, dass Teilaufgaben durch Klassifizierung, Regression, Wahrscheinlichkeitsabschätzung usw. gelöst werden.

In dieser ersten Stufe muss *das Entwicklungsteam sorgfältig über die zu lösende Aufgabe und die denkbaren Anwendungsfälle nachdenken*. Hierbei handelt es sich um eines der wichtigsten fundamentalen Prinzipien der Data Science, dem wir zwei vollständige Kapitel gewidmet haben (Kapitel 7 und Kapitel 11). Was genau ist das Ziel? Und wie genau wollen wir es erreichen? Welche Teile des Anwendungsszenarios sind für Data-Mining-Modelle geeignet? Um das ausführlich erörtern zu können, verwenden wir zunächst ein vereinfachtes Szenario, das bei der allmählichen Weiterentwicklung nach und nach angepasst wird, um die tatsächlichen geschäftlichen Aufgaben besser widerzuspiegeln. Wir werden einige konzeptuelle Tools vorstellen, die uns bei unseren Überlegungen unterstützen, beispielsweise kann die Formulierung einer Aufgabe in Form von Erwartungswerten es ermöglichen, sie systematisch in Data-Mining-Aufgaben zu zerlegen.

2.4.2 Datenverständnis

Wenn das Ziel darin besteht, geschäftliche Aufgaben zu lösen, stellen die verfügbaren Daten das Rohmaterial dar, auf dem die Lösung aufbaut. Die Kenntnis und das Verständnis der Stärken und Beschränkungen der Daten ist wichtig, weil sie nur in den seltensten Fällen exakt auf die zu lösende Aufgabe zugeschnitten sind. Die vorhandenen Daten wurden oft gesammelt, ohne ein bestimmtes Ziel zu verfolgen, manchmal wurden sie auch für andere Zwecke gesammelt, die in keinem Zusammenhang mit der vorliegenden Aufgabe stehen. Eine Kundendatenbank, eine Transaktionsdatenbank und eine Vertriebsdatenbank enthalten unterschiedliche Informationen, decken verschiedene Kundengruppen ab, und die Verlässlichkeit der Daten kann schwanken.

Auch die *Kosten* für die Daten variieren häufig. Manche Daten sind praktisch kostenlos verfügbar, bei anderen ist die Beschaffung mit hohem Aufwand verbunden oder sie müssen gekauft werden. Wieder andere Daten existieren noch gar nicht,

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

und es ist erforderlich, sie in einem zusätzlichen Projekt eigens zu beschaffen. Bei der Beurteilung der Daten ist von entscheidender Bedeutung, die Kosten und den Nutzen verschiedener Datenquellen abzuschätzen und zu prüfen, ob sie weitere Investitionen rechtfertigen. Nachdem alle Daten beschafft wurden, kann durch den Datenabgleich weiterer Aufwand entstehen. So sind beispielsweise Kundendatensätze und Produktkennzeichnungen dafür berüchtigt, »verrauscht« zu sein und sich ständig zu ändern. Kundendatensätze zu bereinigen und abzugleichen, um zu gewährleisten, dass es für jeden Kunden nur einen Eintrag gibt, ist schon für sich genommen ein kompliziertes Analyseproblem (Hernández und Stolfo, 1995; Elmagarmid, Ipeirotis und Verykios, 2007).

Wenn die Kenntnisse über die Daten allmählich zunehmen, ändern sich dadurch möglicherweise die Lösungswege, und auch die Leistungen des Teams können in verschiedene Richtungen auseinanderdriften. Die Betrugserkennung ist hierfür ein schönes Beispiel. Bei dieser wird Data Mining auf breiter Front eingesetzt, und viele damit einhergehende Aufgabenstellungen verwenden klassische überwachte Data-Mining-Tasks. Betrachten Sie z.B. die Aufgabe, Kreditkartenbetrug aufzudecken. Die Abbuchungen erscheinen auf dem Konto des Kunden, betrügerische Abbuchungen werden also für gewöhnlich erkannt – wenn nicht direkt vom Kreditkartenunternehmen, dann später vom Kunden, der seine Kontoumsätze überprüft. Wir dürfen annehmen, dass nahezu alle Betrügereien aufgedeckt und zuverlässig als solche gekennzeichnet werden, denn der ehrliche Kunde und der Betrüger sind verschiedene Personen und verfolgen unterschiedliche Ziele. Kreditkartentransaktionen sind daher verlässlich durch Label (*betrügerisch* und *rechtmäßig*) gekennzeichnet, die als Ziele für überwachte Verfahren dienen können.

Betrachten Sie nun die ähnliche Aufgabe, Medicare-Betrügereien zu erkennen. (Medicare ist eine öffentliche Krankenversicherung des US-Gesundheitssystems für ältere oder behinderte Bürger.) In den USA stellen diese Betrugsfälle ein enormes Problem dar, das jährlich Schäden in Höhe von mehreren Milliarden Dollar verursacht. Auf den ersten Blick scheint es sich hier um eine konventionelle Betrugserkennungsaufgabe zu handeln, doch wenn man die Beziehung zwischen der geschäftlichen Aufgabe und den Daten berücksichtigt, wird klar, dass die Aufgabe hier eine völlig andere ist. Die Betrüger – medizinische Dienstleister und manchmal auch deren Patienten, die unberechtigte Forderungen erheben – sind gleichzeitig auch rechtmäßige Dienstanbieter und Nutzer des Abrechnungssystems. Die Betrüger sind also eine Teilmenge der ehrlichen Nutzer – zusätzliche unabhängige Beteiligte, die festlegen, welche Forderungen berechtigt sind, gibt es nicht. Dementsprechend enthalten die Medicare-Abrechnungsdaten auch keine zuverlässige Zielvariable, die einen Betrug kenntlich macht, und ein überwachter Lernansatz, der bei Kreditkartenbetrug funktionieren würde, ist nicht anwendbar. Für Aufgaben wie diese sind üblicherweise unüberwachte Verfahren wie Profiling, Clustering, Anomalie-Erkennung und Gruppierung nach gleichzeitigem Auftreten erforderlich.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Die Tatsache, dass es sich in beiden Fällen um Betrugserkennung handelt, ist eine tatsächlich irreführende oberflächliche Ähnlichkeit. Wir müssen einen Blick unter die Oberfläche werfen, um die Struktur der geschäftlichen Aufgabe und die verfügbaren Daten zu entdecken und dann geeignete, wissenschaftlich und technologisch fundierte Data-Mining-Verfahren darauf anwenden. Es ist nicht ungewöhnlich, dass für die Erfüllung geschäftlicher Aufgaben mehrere Data-Mining-Verfahren, oft verschiedenen Typs, und eine Kombination ihrer Ergebnisse notwendig sind (siehe Kapitel 11).

2.4.3 Datenaufbereitung

Die Analysemethoden, die wir einsetzen können, sind zwar leistungsfähig, erfordern aber auch, dass die verwendeten Daten gewisse Voraussetzungen erfüllen. Daten müssen oftmals in einer bestimmten Form vorliegen, die sich von derjenigen unterscheidet, in der sie üblicherweise bereitgestellt werden, daher ist eine Konvertierung notwendig. Die Phase der Datenaufbereitung, in der Daten verarbeitet und in Formen konvertiert werden, die bessere Resultate liefern, geht daher mit dem Datenverständnis einher.

Das Konvertieren in Tabellenformate, die Entfernung oder Ergänzung fehlender Werte und die Umwandlung in andere Datenformate sind typische Beispiele für Datenaufbereitungen. Manche Data-Mining-Verfahren sind für symbolische und kategoriale Daten ausgelegt, andere hingegen können nur numerische Werte verarbeiten. Darüber hinaus müssen numerische Werte oft normalisiert oder skaliert werden, damit sie vergleichbar sind. Für derartige Konvertierungen sind Standardverfahren und Faustregeln verfügbar. In Kapitel 3 stellen wir die beim Data Mining gebräuchlichsten Formate ausführlicher vor.

Wir werden uns in diesem Buch im Allgemeinen jedoch nicht eingehender mit Datenaufbereitungsverfahren befassen, die genügend Stoff für ein eigenes Buch bieten (Pyle, 1999). In den folgenden Kapiteln werden wir elementare Datenformate definieren und uns ansonsten nur mit Datenaufbereitungsdetails befassen, wenn sie fundamentale Prinzipien der Data Science erhellen oder um konkrete Beispiele vorzustellen.

Hinweis

Im Allgemeinen müssen Data Scientists anfangs beträchtliche Zeit dafür aufwenden, die Variablen zu definieren, die später verwendet werden. Gerade hier kommen die menschliche Kreativität, der gesunde Menschenverstand und das Fachwissen ins Spiel. Die Qualität einer Data-Mining-Lösung beruht oft darauf, wie gut die Analysten die Aufgabenstellung strukturieren und die Variablen gestalten (und manchmal ist es verblüffend schwierig für sie, das zuzugeben).

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Während der Datenaufbereitungsphase ist ein sehr allgemeiner und wichtiger Punkt zu beachten: Hüten Sie sich vor sogenannten »Lecks« (Kaufman et al., 2012). Man spricht von einem Leck, wenn eine Variable in älteren Daten Auskunft über die Zielvariable gibt – die Informationen also nur in älteren Daten enthalten sind, jedoch nicht zum Zeitpunkt der Entscheidungsfindung zur Verfügung stehen. Ein Beispiel: Wenn man zu einem bestimmten Zeitpunkt versucht vorherzusagen, ob ein Besucher einer Website seine Sitzung beendet oder aber weitersurft, ist die Variable »Gesamtzahl der in dieser Sitzung besuchten Webseiten« eine Vorhersage. Allerdings steht die Gesamtzahl der in der Sitzung besuchten Webseiten ja erst fest, wenn die Sitzung tatsächlich beendet wurde (Kohavi et al., 2000) – erst zu diesem Zeitpunkt wüsste man also den Wert der Zielvariablen! Betrachten Sie als weiteres anschauliches Beispiel die Vorhersage, ob ein Kunde viel Geld *ausgeben wird* – die Kategorien, zu denen die gekauften Artikel gehören (oder sogar die dabei gezahlte Mehrwertsteuer), zu kennen wäre sehr praktisch. Sie sind jedoch zum Zeitpunkt der Entscheidungsfindung unbekannt (Kohavi und Perekh, 2003). Solche Lecks müssen bei der Datenaufbereitung sorgfältig berücksichtigt werden, da diese typischerweise erst im Nachhinein stattfindet – mit bereits vorhanden Daten. In Kapitel 14 stellen wir ein Beispiel für ein Leck aus der Praxis ausführlicher vor, das wirklich schwer zu finden war.

2.4.4 Modellbildung

Die folgenden Kapitel haben die Modellbildung zum Inhalt, und wir werden an dieser Stelle nicht näher darauf eingehen. Es sei jedoch erwähnt, dass das Ergebnis der Modellbildung eine Art Modell oder Muster ist, das Gesetzmäßigkeiten in den Daten erfasst.

Data-Mining-Verfahren werden vornehmlich während der Modellbildungsphase auf die Daten angewendet. Hier ist es wichtig, über grundlegende Kenntnisse des Data Minings zu verfügen, inklusive der schon vorhandenen Verfahren und Algorithmen, denn in diesem Bereich kommen die wissenschaftlichen Erkenntnisse und die Technologie am stärksten zum Tragen.

2.4.5 Beurteilung

Das Ziel der Beurteilung ist es, die Ergebnisse des Data Minings gründlich zu bewerten und Vertrauen darin zu gewinnen, dass sie gültig und verlässlich sind, bevor man fortfährt. In jeder Datenmenge lassen sich Muster erkennen, wenn man nur will, sie werden einer gründlichen Untersuchung aber vermutlich nicht standhalten. Wir möchten darauf vertrauen können, dass die aus den Daten gewonnenen Modelle und Muster echte Gesetzmäßigkeiten darstellen und nicht nur Eigenheiten oder Abweichungen vom Muster sind. Es ist zwar machbar, die Ergebnisse sofort nach dem Data Mining anzuwenden, aber nicht zu empfehlen. Für gewöhnlich ist es viel einfacher, weniger aufwendig, schneller und sicherer, ein Modell zunächst einmal unter kontrollierten Laborbedingungen zu testen.

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Stichwortverzeichnis

0-1-Verlustfunktion 125
1-NN-Klassifizierer 190

A

A priori 409
A/B-Test 48
Abdeckung 409
Abfrage (Datenbank) 65
Abfragetool 65
Abhängige Variable 75
Abhängigkeit 291
Absoluter Fehler 126
Abstimmungsverfahren
    gewichtetes 188
Abwanderung 26
Abwanderungsmuster 144
Abwanderungsrate 104, 163
ACM SIGKDD International Conference on Data Mining and Knowledge Discovery 398
Ähnlichkeit 14, 177, 178, 387, 390
Ähnlichkeitsgewichtung 188, 201
Ähnlichkeitsmaß 179, 353
Ähnlichkeitsscore 355
Ähnlichkeitstest 46
Algorithmen 13
Amazon 24, 32, 36, 178, 193, 366, 368, 370
Analyseverfahren 62, 337
Analytische Fähigkeiten 62
Analytisches Engineering 325, 339
Anbieterwechsel 36
Änderung der Entropie 82
Annahme der Unabhängigkeit 291
Annahmen
    vereinfachende 112
Anonymisierung 397
Anpassung 152
Anpassung an die Daten 141, 155
Anschauliche Visualisierung 263
Anwendungen 367
Anwendungsphase 51
Apollo 13 378
A-posteriori-Wahrscheinlichkeit 284
Apple 215
A-priori-Verteilung 242
A-priori-Wahrscheinlichkeit 284
Arbeitsblatt 75, 293
Arbeitsmarktökonomie 398
Art des Ziels 50
Assoziation 15, 340
Assoziationsanalyse 341, 343
Assoziationsmining 409
Assoziationsregel 46
Attribut 409
Attribut Siehe Merkmal
AUC-Wert 270
Aufgabe 330
    Definition 401
Aufgaben 363
Aufgabentypen 44
Aufgabenvereinfachung 314
Aufgabenverständnis 53, 61, 326, 333, 335
Aufspüren informativer Variablen 248
Aufspüren von Assoziationen 344
Aufteilung der Datenmenge 148
Ausreißer 154, 190
Auswahl einer Teilmenge 72
Auswahleffekt 329, 395
Auswahlverfahren 329
automatisierte Entscheidungsfindung 30

B

Bag-of-words-Modell 298
Basisrate 148, 257
Baum
    Komplexität 149
    stutzen 168
Baumstruktur 92
Baumstrukturmodell 92, 149
Bayes
    Satz von 282
Bayes-Fehlerrate 360
Bayes-Klassifizierer
    naiver 271, 287, 288
    Varianten 289
Bayessche Klassifizierungsregel 284
Bayessche Methoden 283, 294

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Bedingte Unabhängigkeit 286
Bedingte Wahrscheinlichkeit 242, 280
Berechnung der Ähnlichkeit 390
Beschreibung
    charakterisierende 223
    differenzierende 223
Beschreibungsmodell 74
Betrugserkennung 128, 396
Beurteilung 57, 403
    Projektvorschlag 406
    von Modellen 137, 159, 275
    von Projektvorschlägen 379, 401
Beurteilungskriterien 228, 245
Beurteilungsphase 61, 225, 393
Bewegungsprofil 390
Bewerbungsgespräch 14
Bias 359, 361
Big Data 31
Big Data 2.0 32
Bigramm 311
Binäre Klasse 112
Binäre Zielvariable 279
Binäres Label 78
Blatt 92
Blattknoten 92
    Anzahl der Instanzen 167
Borders 368
Börsenkurs 314

C

Capital One 35
Cloud-Arbeiten 398
Cluster 204
    Distanzfunktion 206
Cluster-Beschreibungen 220
Clustering 46, 203
    Ergebnis 218
    hierarchisches 204
Clustering-Algorithmus 209
Clustering-Güte 214
Coase, Ronald 145
Compaq 368
Cookies 278
CRISP-Datamining-Prozess 53
CRISP-DM (Cross Industry Standard Process for Data Mining) 39
Crowd-Sourcing 398

D

Data Engineering 31
Data Mining 409
    Anwendung 51
    Begriff 24
    Ergebnis 51
    überwachtes 72
Data Science
    Arbeitsmarkt 375
    Ausgereiftheit 382
    Begriff 24
    Handwerk 372
    Konzepte 14, 39
    Praxis 378
    Ziele 27
Data Scientist 14, 40, 371
Data Scientists, LLC 377
Data Warehouse 66
Data-Mining-Algorithmen 13, 38, 298
Data-Mining-Aufgaben 43
Data-Mining-Aufgabenstellung 53
Data-Mining-Prozess 43, 52, 70
Data-Mining-Techniken 23
Data-Mining-Toolkits 289
Data-Mining-Tools 295
Data-Mining-Verfahren 145
Data-Science-Community 373
Data-Science-Fähigkeiten 33
Data-Science-Management 373
Data-Science-Team 375
Daten
    Kosten 54
    Objektivität 393
    persönliche 396
    zurückgehaltene 146
Daten als Gut 335, 396
Daten von Einzelpersonen 396
Datenabgleich 55
Datenanalyse 24, 26, 36, 402
Datenanalysetechnologien 62
Datenanalytische Denkweise 34, 36, 365, 386, 401
Datenaufbereitung 56, 215, 295, 395, 402
Datenbankabfrage 65
Datenbanktabelle 144
Datenbereinigung 409
Datenbeschaffung 55, 335
Datenerfassungsmöglichkeiten 23
Datenfeld 410
Datengestützte Aufgabe 44
Datengestützte Entscheidungsfindung 27, 30, 42, 365, 392
Datengestützte Modelle 35

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Datengestützte Vorhersagen 25
Dateninterpretation 395
Datenmenge 47, 75, 119, 410
Datenpunkte 123
Datenqualität 409
Datenquelle 182, 248
Datenreduzierung 47, 354
Datensammlung 23
Datensatz 410
Datenschutz 397
Datenverarbeitung 31
Datenverarbeitungstechnologien 31
Datenverständnis 54, 56, 325, 401
Datenvorverarbeitung 319
DDD (Data-Driven Decision-Making, datengestützte Entscheidungsfindung) 27
Decision Stump 248
Deduktion 76
Dell 215, 368
Dendrogramm 204, 206
Deployment 59, 410
Determinismus 360
Dictionary of Distances 197
Dimension 410
    latente 355
Diskriminanzfunktion
    lineare 115, 119
Diskriminatives Modell 294
Display Advertising 278
Distanz 178
    euklidische 180
Distanzmaß 179, 197
Distanz-Metriken 198
Dokument 298
Dokumenthäufigkeit
    inverse 302

E

Editierdistanz 200
Eigennamenerkennung 311
Eingabedaten 59
Einkaufsgewohnheiten 29
Einseitiges Auswahlverfahren 329
Eintrag 410
Einzigartigigkeit 369
Elder Research 377
Ellenbogenkriterium 214
Elternmenge 82
Empfehlung 357
Empfehlungsmodell 235, 358
Engagement als Berater 376
Engineering
    analytisches 325
Ensemble 273
Ensemblemethode 355, 359, 361
Ensemblemodell 358
Entropie 80, 88, 309
    Änderung 82
Entscheidungsbaum 92, 135, 248
    Lernkurve 165
    Überanpassung vermeiden 167
Entscheidungsbaumverfahren 72, 94, 149
Entscheidungsfindung 29
Entscheidungsgerade 100
Entscheidungsgrenze 100, 113, 134, 153
Entscheidungsknoten 92
Ergebnisse des Data Minings 51, 57, 59, 227
Erkennung von Überanpassung 175
Erklärende Variable 75
Ersatzdaten 378
Erwartungswert 235, 255, 310, 327, 330, 363
    Berechnung 250
    Beurteilung von Klassifizierern 238
Erwünschte Klassifizierung 280
Ethik 396
Euklidische Distanz 180, 197
Euklidisches Distanzmaß 196
Evidenz 277, 290
Experte 359
Explorative Datenanalyse 203, 389, 402
Extraktion nützlichen Wissens 39

F

Facebook 36, 37, 278, 291, 344
Facebook-Likes 292, 343
Fall 410
Fallbasiertes Schließen 189
Fallstudien 377
Falsch Negative 230, 241, 258
Falsch Positive 230, 241, 258
Fehlalarmquote 259
Fehlender Wert 410
Fehler
    absoluter 126
Fehlerquadrate 127
Fehlerquote 239, 410, 411
Fehlerrate 360
Fehlersuche 250
Fehlertypen 234
Fehlklassifizierungsrate 229

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Filmbewertung 358
Filmempfehlung 235, 354, 357
Fitfunktion 146, 147, 175, 269
Fläche unter der ROC-Kurve 263
F-Maß 245
Förderungsprogramm 376
Formulierung der Aufgabe 334, 391
Formulierung der Aufgabenstellung 53
Freiheitsgrad 155
Fundamentale Konzepte 77, 339, 365
Fundamentale Konzepte der Data Science 329, 385
Funktion
    zusammenfassende 200
Funktionsanpassung 141

G

Gängige Tools 340
Gaussian Mixture Model (GMM) 351
Gauß-Verteilung 348, 350
    zweidimensionale 351
Geistiges Eigentum 370
Genauigkeit 245
General Electric Capital 225
Generatives Modell 294
Geometrische Interpretation 189
Geschäftliche Aufgaben 54, 68, 389
Geschäftsanalyse 62
Geschäftsgeheimnis 371
Gewichtetes Abstimmungsverfahren 188
Gewichtetes Scoring 188
Gewichtungen 358
Gewichtungsfunktion 201
Gezieltes Marketing 396
Gini-Koeffizient 263
Gleichzeitiges Auftreten 340
Glockenkurve 349
Google 366
Google Finance 316
GPS-Daten 388
Grundgesamtheit 145
Gruppierung nach gleichzeitigem Auftreten 46, 340

H

Hadoop 41
Harrah´s Casino 36
Häufigkeitsbasierte Abschätzung 103
Helicobacter pylori 207
Heterogene Merkmale 196
Hewlett-Packard 177, 215
Hierarchisches Clustering 204
Hinge-Verlustfunktion 124, 173
Holdout-Daten 146
Homograf 297
Hurrikan Frances 25
Hyperebene 100
Hypothesenerzeugung 64
Hypothesentest 64, 69, 167, 175

I

IBM 46, 177, 219, 377
Ideen übernehmen 378
IDF 302
IEEE Conference on Data Mining 398
Immaterieller Wert 371
Implementierung 59
Implementierungsdetail 371
Implizite Klassifizierungsbereiche 189
Individuum 45
Induktion 76, 410
Induktionsalgorithmus 76
Informationen
    latente 313
Informationsgewinn 80, 81, 82, 88, 105
Informationstechnologie 39
Informationsverlust 354
Informative Merkmale 72
Informative Variablen 71, 78
Inkrementelles Lernen 288
Instanz 74, 75, 149, 278, 410
Instanzenbasiertes Lernen 189
Instanzenraum 98, 190, 209
Intervall 197
Inverse Dokumenthäufigkeit 302
Investition in Daten 33, 336, 398
IP-Adresse 388
Iris-Datensammlung 119, 152
Iteration 53, 213

J

Jaccard-Distanz 198
Jaccard-Koeffizient 198
Jargon 400
Jobs, Steve 395

K

Kategoriale Merkmale 412
Kategoriale Ziele 50
Kaufquote 361
Kausalanalyse 363
Kausalmodell 48, 362

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

KDD 411
KDD (Knowledge Discovery and Data Mining) 68
KDD Cup 2009 266
Kernel-Funktion 139
Kindmenge 82
Klasse 411
Klassenbezeichnung 184, 191
Klassenverteilung 250
Klassenvorhersage 45
Klassenzugehörigkeit 45, 121
Klassifizierer 411
    Beurteilung 228
    diskreter 258
    linearer 114, 134
Klassifizierung 45, 50, 284
Klassifizierungsaufgabe 78, 87
Klassifizierungsbaum 92, 134
    Leistung 163
Klassifizierungsbaum als Regelsatz 101
Klassifizierungsbereich 189
k-Means-Algorithmus 209
k-Means-Clustering 213
Knoten 92
Kombination mehrerer Modelle 358
Kommunikation 387
Komplexität 146
    Steuerung 167, 171, 341
Komplexität mathematischer Funktionen 155
Komplexitätsparameter 170, 191
Konfidenz (Assoziationsregel) 341
Konfidenzintervall 64
Konfusionsmatrix 411
Konfusionsmatrix Siehe Wahrheitsmatrix
Kontext 297
Kontrafaktische Analyse 48
Konventionen 16
Konzeptuelle Frameworks 339
Korrektklassifizierungsrate 108, 137, 147, 163, 229, 268, 411
Korrelation 45, 65, 157, 218
Kosinus-Ähnlichkeit 199, 307
Kosinus-Distanz 198
Kosten und Nutzen 250, 257
Kosten/Nutzen/Verlust 411
Kosten-Nutzen-Analyse 320
Kosten-Nutzen-Matrix 242
Kosten-Nutzen-Verhältnis 233, 274
Kosten-Nutzen-Werte 240
Kreditausfall 33, 34, 102
Kreditkarte 33
Kreuzvalidierung 159, 161, 176, 268, 411
    verschachtelte 169
Kumulative Reaktionskurve 264
Kundenabwanderung 30, 329
Kundenbindung 27, 35, 340, 406
Kundenfluktuation 26
Kundengewinnung 35
Kundensegment 49
Kundenverhalten 351
Kundenwert 331

L

L1-Norm 173, 197
L1-Regularisierung 173
L2-Norm 173, 197
Label 50, 55, 76, 411
Laplace-Korrektur 103
LASSO-Verfahren 173
Latent Dirichlet Allocation 313
Latente Dimensionen 355, 357
Latente Informationen 313
Lebensbaum 207
Leck 57, 394
Leistungsanalyse 266
Leistungsbeurteilung 246
Lernen
    inkrementelles 288
    überwachtes 220
Lernkurve 165, 176, 360
Levenshtein-Distanz 200
Leverage 342
Lift 15, 264, 290, 342, 387
Lift-Kurve 264, 266, 272, 321
Lifts der Evidenz 294
Lineare Diskriminanzfunktion 115, 122
Lineare Modellbildungsverfahren 141
Lineare Regression 112, 126
    Lernkurve 165
Linearer Klassifizierer 114, 134
Lineares Modell 151
Linkage-Funktion 206
Logarithmische Skalierung 350
Logistische Regression 128
    Leistung 163
Logit 129
Logit-Wert 130
Log-Normalverteilung 350
Lösen geschäftlicher Aufgaben 385
Lösung
    Formulierung 401

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

M

Machine Learning 67, 412
Magnum Opus 344
Manhattan-Distanz 197
Margin 123
Marketing
virales 362
Marktverhalten 47
Mathematische Funktionen 151
Mathematische Notation 17
Matrix-Faktorisierung 313, 358
Maximale Reinheit 96
Maximum-Likelihood-Methode 133, 172, 348
McKinsey 38
Mechanical Turk 398
Medianwert 63
Medicare 55
Mehrfache Vergleiche 175
Mehrheitsentscheidung 184, 188, 200, 247
Menge 299
Menschliches Eingreifen 393
Merkmal 39, 71, 74, 278, 412
informatives 77, 78
Intervalle 85
irrelevantes 194
nominales 412
numerisches 85
ordinales 412
Merkmale 88
Definition 402
Merkmalsauswahl 86, 108, 152, 195
Merkmalsgewichtung 116
Merkmalsraum 139
Merkmalsvektor 74, 144, 181, 324, 412
Methode der kleinsten Quadrate 127
Methoden des Data Minings 385
Metrik 197
Micro-Outsourcing 398
Mittel
arithmetisches 63
Mittelwert 349
Mittelwertbildung 361
Mobilgeräte 388
Modell 144, 412
Beurteilung 159
Definition 73
diskriminatives 294
generatives 294
Komplexitätsgrad 147
Leistung 150
nichtlineares 138
parametrisiertes 116
Vergleich 247
Modellbildung 57, 146, 263, 334, 403
parametrisierte 111
Modellbildungsverfahren 148
Modellinduktion 76
Modellkomplexität 145
Modellparameter 111
Monogramm 311
Mozilla Thunderbird 289
Multimenge 299
Multivariate Segmentierung 92
Muster 39, 57, 143, 145
Muster in den Daten 413

N

Nächste Nachbarn 181, 184
Anzahl 187
Dimensionalität 194
Fachwissen 194
Klassifizierung 184
Regression 186
Verständlichkeit 193
Wahrscheinlichkeitsabschätzung 186
Nächste-Nachbarn-Klassifizierer 191
Naiver Bayes-Klassifizierer 271, 287, 288
Negatives Ergebnis 228
Netflix 178, 193
Netflix-Wettbewerb 354
Neuronales Netz 140
N-Gramm 311
Nichtlineares Modell 138
Nichttriviale Klassifizierung 149
Nominales Merkmal 412
Notation
mathematische 17
Numerische Funktionen 138
Numerische Merkmale 182, 194
Numerisches Klassifizierungsmodell 116
Numerisches Merkmal 85
Numerisches Modell 148
Nutzen 331

O

Odds 129
OLAP 412
Online Analytical Processing (OLAP) 66
Onlinewerbung 266, 277
Optimalpunkt 150, 361
Orange 267
Ordinales Merkmal 412

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

P

Parameter 111
Parameterlernen 111
Parameterwerte 111
Parametrisierte Modellbildung 126, 141
Parametrisiertes Modell 116
Pasteur, Louis 367
Patent 371
Personenpaar 354
Persönliche Daten 396
Pilzarten 86
Pilz-Datensammlung 86
Placeboeffekt 48
Plausibilität der Evidenz 279
Positiver Vorhersagewert 246
Positives Ergebnis 228
Prävalenz 283
Prior 242, 257
Privatsphäre 396, 397
Probabilistische Modelle 313
Probabilistisches Clustering 352
Probengenauigkeit 146
Produktakzeptant 36
Produktempfehlungen 46
Profiling 47, 347
Profit 240
Profitabilität 34
Profitabilitätsmodell 34
Profitkurve 254, 274, 276
Profitmaximierung 256
Projektvorschlag 405
Projektvorschläge
    beurteilen 379, 401
Publikation 376
Punkte im ROC-Raum 260
p-Wert 168

Q

Quadratischer Abweichungsverlust 125
Qualitative Bewertung 58

R

Randomisierte Experimente 48, 59
Rangfolge 252, 253, 288
Rastersuche 174
Reaktionskurve
    kumulative 264
Regelsatz 101
Regression 45, 50
    lineare 126
    logistische 128
Regressionsanalyse 67
Regressionsbaum 93
Regularisierung 171, 172, 176
Reinheit 79
Reinheitsmaß 80, 109
rekursiv 94
Rekursive Aufteilung 148
Relative Häufigkeit 341
Reuters 214
Richtig Negative 230, 241, 258
Richtig Positive 230, 241, 258
Ridge-Regression 173
ROC-AUC 263
ROC-Diagramm 258
ROC-Kurve 261, 276, 320
Rückwärtseliminierung
    sequenzielle 170

S

Satz von Bayes 282, 294
SBS-Algorithmus 170
Schema 413
Schutz der Privatsphäre 396
Schwellenwert 98, 167, 252, 274
Score 121, 186, 252
Scoring 45
    gewichtetes 188
Segmentierung 114
    überwachte 71, 77
    unüberwachte 178, 203
    Visualisierung 98
Selbstbeurteilung 382
Semantische Indizierung 313
Sensitivität 246, 413
Sequential Backward Elemination Siehe SBS-Algorithmus
Sequential Forward Selection Siehe SFS-Algorithmus
SFS-Algorithmus 170
Shannon, Claude 80
Sigmoid 132
Signet-Bank 33, 34
Skalierung 196
Softwareentwicklungskenntnisse 62
Softwareentwicklungszyklus 61
Spam-Klassifizierer 288
Speicherbasiertes Lernen 189
Spendenaufruf 326
Spezifität 246, 413
SQL (Structured Query Language) 65
SQL-Abfrage 66

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Standardabweichung 269, 349
Standardklassifizierer 289
Standard-Prozess-Modell für Data Mining 52
Standardzielfunktion 133
Standortdaten 388
Stapelmodelle 140
Stärke 343
Stärke (Assoziationsregel) 341
Statistik 64
    zusammenfassende 63
Stellvertreterbezeichnung 335
Stemming 301, 305
Stetige Merkmale 412
Steuerung der Komplexität 167, 171, 341
Stichprobe 333
Stoppwörter 301
Strafterm 124, 125, 172
Streudiagramm 98, 388
Suchmaschinenwerbung 278
Summit Technologies 317
Support 341, 343
Support Vector Machine (SVM) 118, 122, 173
    nichtlineare 139
Synonym 297

T

Tabelle 75
Tabellenkalkulation 293
Tabellenmodell 144
Target 29
Taxi-Distanz 198
Teilaufgaben 44
Teilmenge 72, 161
Testdaten 161, 163, 168
Testdatenmenge 146
Testlabor 160
Text
    Auswertung 297
    Bedeutung 296
    Dokument 298
    Fehlerhaftigkeit 297
    Kontext 297
    Repräsentierung 298, 304
    Suchmaschine 307
    Token 298, 302
    Zahlen 301
Text Mining 296, 323
Textdaten 295
Textklassifizierung 199, 399
Textkorpus 215, 298
Textorientierte Data Science 296
TFIDF-Gewichtung 306
TFIDF-Maß 215, 303, 390
TFIDF-Repräsentierung 307
Token 298, 302, 308, 322
Tools 65
Topic Layer 312
Träges Lernen 189
Trainierter Klassifizierer 221
Trainingsdaten 76, 145, 146, 161, 163, 278
Trainingsdatenmenge 268
Trainingsmenge 108
Trainingsuntermenge 169
Trefferquote 245, 259
Trennfunktion 115
Trennlinie 153
Trigramm 311
Tumblr 278
Tupel 413
Twitter 37

U

Überanpassung 40, 141, 143, 145, 268
    Entscheidungsbaumverfahren 149
    lineare Funktionen 152
    mathematische Funktionen 151
    Nachteile 157
    Vermeidung 167
Überwachte Modellbildung 203
Überwachte Segmentierung 77, 202
Überwachtes Data Mining 344
Überwachtes Lernen 74, 220, 413
Überwachtes Verfahren 49
Unabhängige Variable 75
Unabhängigkeit
    bedingte 286
Unabhängigkeit der Merkmale 290
Unausgewogene Klassenverteilung 231
Unausgewogenheit 231
Unerwartetheit 341
Untermenge 149
Unüberwachte Modellbildung 203
Unüberwachte Segmentierung 178, 203
Unüberwachtes Data Mining 344
Unüberwachtes Lernen 413
Unüberwachtes Verfahren 49
Ursächliche Zusammenhänge 362

V

Validierungsmenge 169
Variable 248, 413
    informative 71

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546

Varianz 85, 160, 268, 360, 361
Verallgemeinerungsfähig 156
Verallgemeinerungsfähigkeit 143, 145, 160, 228, 336
Verborgene Strukturen 218
Verbraucherbeeinflussung 362
Verbraucherkredit 30
Verbraucherverhalten 35
Verbundkaufanalyse 46
Verbundwahrscheinlichkeit 281
Vereinfachende Annahmen 112
Verfahren
    überwachte und unüberwachte 49
Verfügbare Daten 54
Vergleiche
    mehrfache 175
Verhaltensmuster 26
Verknüpfungsvorhersage 47, 353, 355
Verlust 413
Verlustfunktion 125
Verteilung (Statistik) 63
Verteilzentrum 340
Verzerrung 213
Virales Marketing 362
Visualisierung 252, 275
Vokabular 215
Vorhersagbarkeit 320
Vorhersage 25, 44, 50, 73
Vorhersagegenauigkeit 263
Vorhersagekraft 35
Vorhersagekraft eines Modells 142
Vorhersagemodell 29, 33, 69, 184, 353
    implementieren 59
Vorhersagemodellbildung 71, 104, 108
Vorkommenshäufigkeit 215, 299, 301
Vorwärtsauswahl
    sequenzielle 170

W

Wahrheitsmatrix 230, 239, 245, 253, 413
Wahrscheinlichkeit 51, 280
    bedingte 242
Wahrscheinlichkeit der Klassenzugehörigkeit 130
Wahrscheinlichkeitsabschätzung 44, 45, 51, 101, 121, 252, 327, 333
    der Klassenzugehörigkeit 285
    häufigkeitsbasierte 103
Wahrscheinlichkeitsabschätzungsbaum 93
Walmart 24, 25
Warenkorbanalyse 46, 343
Wartezeit 349
Web 2.0 32, 296
Website zum Buch 16
Wert 331
Wert eines Guts 368
Wertebereich 196
Wertekombinationen 182
Wettbewerbsnachteil 369
Wettbewerbsvorteil 29, 368
Wettervorhersage 247
Whisky-Analyse 181, 203
Wilcoxon-Mann-Whitney-Test 263
Wirtschaftsnachrichten 214
Wisconsin-Brustkrebs-Datensammlung 135
Wissensentdeckung 413
Wissensextraktion 40, 386
Worthäufigkeit 301
Wortkombination 311
Wortreihenfolge 310
Wurzelknoten 92

Y

Yahoo! 278
Yahoo! Finance 316

Z

Zeitstempel 319
Zentroid 209
Zerlegung des Erwartungswerts 333
Zerlegung in Teilaufgaben 44
Ziel 49, 73, 118
Ziel des Data Minings 223
Zielfunktion 118, 127, 358
Zielgruppe 330
Zielkriterium 393
Zielvariable 50, 71, 74, 75, 88, 109, 278, 329, 402
Zielwert 50
Zufälligkeit 360
Zurückgehaltene Daten 146
Zusammenarbeit von Mensch und Maschine 393
Zusammenfassende Funktion 200
Zusammenfassende Statistik 63

© des Titels »Data Science für Unternehmen« (ISBN 9783958455467) 2017 by
mitp Verlags GmbH & Co. KG, Frechen. Nähere Informationen unter: http://www.mitp.de/546