

Annalyn Ng · Kenneth Soo

Data Science – was ist das eigentlich?!

Algorithmen
des maschinellen
Lernens verständlich
erklärt

EBOOK INSIDE



Springer

Data Science – was ist das eigentlich?!

Annalyn Ng · Kenneth Soo

Data Science – was ist das eigentlich?!

Algorithmen des maschinellen
Lernens verständlich erklärt

Aus dem Englischen übersetzt von
Matthias Delbrück

 Springer

Annalyn Ng
Singapur, Singapur

Kenneth Soo
Singapur, Singapur

ISBN 978-3-662-56775-3 ISBN 978-3-662-56776-0 (eBook)
<https://doi.org/10.1007/978-3-662-56776-0>

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

Übersetzung der englischsprachigen Ausgabe: Numsense! Data Science for the Layman: No Math Added von Annalyn Ng und Kenneth Soo. © 2017 by Annalyn Ng and Kenneth Soo. Alle Rechte vorbehalten.

© Springer-Verlag GmbH Deutschland, ein Teil von Springer Nature 2018

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung, die nicht ausdrücklich vom Urheberrechtsgesetz zugelassen ist, bedarf der vorherigen Zustimmung des Verlags. Das gilt insbesondere für Vervielfältigungen, Bearbeitungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

Der Verlag, die Autoren und die Herausgeber gehen davon aus, dass die Angaben und Informationen in diesem Werk zum Zeitpunkt der Veröffentlichung vollständig und korrekt sind. Weder der Verlag noch die Autoren oder die Herausgeber übernehmen, ausdrücklich oder implizit, Gewähr für den Inhalt des Werkes, etwaige Fehler oder Äußerungen. Der Verlag bleibt im Hinblick auf geografische Zuordnungen und Gebietsbezeichnungen in veröffentlichten Karten und Institutionsadressen neutral.

Einbandabbildung: © monsitj/stock.adobe.com

Springer ist ein Imprint der eingetragenen Gesellschaft Springer-Verlag GmbH, DE und ist ein Teil von Springer Nature
Die Anschrift der Gesellschaft ist: Heidelberger Platz 3, 14197 Berlin, Germany

Dieses Buch ist Ihnen, unseren Leserinnen und Lesern, gewidmet – von zwei Data-Science-Enthusiasten, Annalyn Ng (University of Cambridge) und Kenneth Soo (Stanford University).

Uns fiel auf, dass Data Science zwar immer häufiger in betrieblichen Entscheidungsprozessen eingesetzt wird, jedoch kaum jemand wirklich etwas darüber weiß. Darum haben wir aus unseren Tutorials ein Buch zusammengestellt – für wissbegierige Studierende, Professionals in Unternehmen oder schlicht alle Neugierigen.

Jedes Tutorial deckt die wichtigsten Funktionen und Annahmen einer Data-Science-Methode ab, und das ohne Mathematik und Fachausdrücke. Illustriert werden die Methoden mit „Real-Life-Daten“ und Beispielen.

Allein hätten wir dieses Buch jedoch nicht schreiben können. Wir danken Sonya Chan, unserer Lektorin und guten Freundin, die unsere unterschiedlichen Schreibstile gekonnt zusammengefügt und dafür gesorgt hat, dass der rote Faden nahtlos durchläuft.

Dora Tan danken wir für ihre Ratschläge zum Layout und zu den Grafiken. Unseren Freunden Michelle Poh, Dennis Chew and Mark Ho verdanken wir unschätzbare Tipps, wie sich der Inhalt noch verständlicher ausdrücken ließ.

Ebenso danken wir Prof. Long Nguyen (University of Michigan, Ann Arbor), Prof. Percy Liang (Stanford University) und Dr. Michal Kosinski (Stanford University) für die Geduld, mit der sie uns ausgebildet haben, und für ihre Expertise, an der sie uns bereitwillig teilhaben ließen.

Zum Schluss möchten wir uns selbst gegenseitig danken – wir zanken uns wie gute Freunde, bleiben aber immer dran, bis das fertig ist, was wir zusammen begonnen haben.

Vorwort

„Big Data“ bedeutet heute „Big Business“. In dem Maß, in dem immer größere Datensammlungen fast alle Aspekte unseres Lebens bestimmen, stürzen sich immer mehr Unternehmen darauf, solche Daten zu Geld zu machen. Techniken zur Mustererkennung und datenbasierte Vorhersagen eröffnen neue Dimensionen geschäftlicher Strategien. Intelligente Produktempfehlungen etwa sind eine Win-win-Situation für Verkäufer und Kunden, wenn sie Kunden auf Produkte aufmerksam machen, an denen sie mit hoher Wahrscheinlichkeit tatsächlich interessiert sind, und die Gewinne der Verkäufer gleichzeitig durch die Decke gehen.

Big Data ist jedoch nur ein Aspekt des Phänomens. Data Science¹ erlaubt es uns, Daten in nahezu beliebiger Menge zu analysieren und zu bearbeiten. Diese neue interdisziplinäre Wissenschaft umfasst unter anderem maschinelles Lernen, Statistik und verwandte Zweige der Mathematik. Das maschinelle Lernen steht hier übrigens nicht zufällig an erster Stelle. Es ist das primäre Werkzeug für das Erkennen von Mustern in Datensätzen und für daraus abgeleitete Prognosen. Mithilfe der Algorithmen des maschinellen Lernens ermöglicht Data Science unschätzbare Einsichten und ganz neue Wege, Informationen aus Daten zu destillieren.

Um zu erkennen, wie Data Science die derzeitige Datenrevolution antreibt, brauchen Uneingeweihte allerdings zumindest ein Grundverständnis dieses vielfältigen Wissensgebiets. Die dazu benötigten Kenntnisse stellen leider für manche eine große Hürde dar, der weitverbreitete Datenanalfabetismus führt in der Praxis zu einem massiven Bedarf an Fachkräften. An dieser Stelle kommt *Data Science – was ist das eigentlich?!* ins Spiel.

Wer die Arbeiten von Annalyn Ng und Kenneth Soo kennt, ist nicht verwundert, dass ihr Buch die Frage im Titel leicht verständlich beantwortet. Hier geht es um Data Science für Uneingeweihte, und die oft sehr komplexe Mathematik – welche das Buch durchaus auf hohem Niveau beschreibt – wird absichtlich nicht im Detail hergeleitet.

¹Man könnte „Data Science“ als Datenwissenschaft übersetzen, im Deutschen wird jedoch in der Regel der englische Begriff verwendet. Ähnliches gilt für viele andere Fachbegriffe wie Support vector oder Random Forests (Anm. d. Übers.).

Verstehen Sie dies nicht falsch: Der Inhalt wird in keiner Weise versimpelt, vielmehr enthält das Buch robuste Informationen, die knapp und präzise zusammengefasst sind.

Was nützt einem dieser Ansatz, könnten Sie sich jetzt fragen. Ziemlich viel – ich würde sagen, dass dies für Anfängerinnen und Anfänger auf jeden Fall der denkbar sinnvollste Einstieg ist. Denken Sie an einen Fahr Schüler in der ersten Fahrstunde: Funktion und Bedienung der wichtigsten Schalter sind erst einmal wichtiger als die thermodynamische Theorie des Verbrennungsmotors. Das Gleiche gilt für Data Science: Wenn Sie das Fach kennenlernen wollen, ist es besser, mit allgemeinen Konzepten zu beginnen, anstatt sich sofort in mathematischen Formalien zu verlieren.

Die Einführung des Buchs macht den Laien auf wenigen Seiten mit den grundlegenden Konzepten vertraut, sodass jede und jeder auf demselben Fundament zu bauen beginnt. Wichtige Themen wie das Auswählen von geeigneten Algorithmen – die oft in einführenden Texten ausgelassen werden – werden hier direkt angesprochen. Auf diese Weise versteht der Leser auch gleich, wie wichtig es ist, sich weiter mit dem Thema zu beschäftigen, und er erhält gleichzeitig einen umfassenden Bezugsrahmen dafür.

Es gibt natürlich sehr viele weitere Themen der Data Science, die Annalyn und Kenneth mit Fug und Recht auch noch in ihr Buch hätten aufnehmen können, und ebenso natürlich eine Vielzahl von möglichen Darstellungsweisen. Ihre Entscheidung, sich auf die wichtigsten Algorithmen des maschinellen Lernens zu konzentrieren, zusammen mit ein paar thematisch passenden Szenarios, war großartig. Gut eingeführte Algorithmen wie

k -Means-Clustering, Entscheidungsbäume und k -nächste Nachbarn werden gebührend gewürdigt. Neuere Klassifikations- und Ensemble-Algorithmen wie Support-Vektor-Maschinen – deren komplexe Mathematik einen schon ziemlich einschüchtern kann – oder Random Forests werden ebenso erklärt wie die neuronalen Netze, die hinter dem aktuellen Deep-Learning-Hype stecken.

Eine weitere Stärke von *Data Science – was ist das eigentlich?!* sind die intuitiven Anwendungsbeispiele. Seien es die Verbrechensvorhersage mit Random Forests oder das Clustering von Blockbuster-Fans, die gewählten Beispiele verbinden Klarheit mit Alltagsbezug. Gleichzeitig hält die Abwesenheit von jeglicher höheren Mathematik das Interesse und die Motivation lebendig, sich auf das spannende Studium von Data Science einzulassen.

Ich kann dieses Buch wirklich nur empfehlen: *Data Science für Laien* ist der ideale Einstieg in die Welt von Data Science und ihren Algorithmen. Ich fände es schwierig, eine vergleichbare Alternative zu benennen. Mit *Data Science – was ist das eigentlich?!* gibt es keinen Grund mehr, sich von der Mathematik den Spaß am Lernen ausreden zu lassen!

Matthew Mayo
Data-Science-Experte und
Mitherausgeber von KDnuggets
@mattmayo13

Inhaltsverzeichnis

1	Das Wichtigste in Kürze ...	1
1.1	Datenaufbereitung	2
1.2	Auswahl des Algorithmus	7
1.3	Parameter	11
1.4	Evaluation der Ergebnisse	14
1.5	Zusammenfassung	18
2	<i>k</i>-Means-Clustering	19
2.1	Wie man Kunden-Cluster findet	19
2.2	Beispiel: Persönlichkeitsprofile von Filmfans	20
2.3	Cluster definieren	22
2.4	Grenzen	27
2.5	Zusammenfassung	28

3	Hauptkomponentenanalyse	29
3.1	Der Nährwertgehalt von Lebensmitteln	29
3.2	Hauptkomponenten	31
3.3	Beispiel: Nahrungsmittelgruppen	34
3.4	Grenzen	40
3.5	Zusammenfassung	43
4	Assoziationsanalyse	45
4.1	Muster im Einkaufsverhalten	45
4.2	Support, Konfidenz und Lift	46
4.3	Beispiel: Daten aus einem Lebensmittelgeschäft	49
4.4	Das A-priori-Prinzip	52
4.5	Grenzen	55
4.6	Zusammenfassung	56
5	Soziale Netzwerkanalyse	57
5.1	Beziehungen abbilden	57
5.2	Beispiel: Waffenhandel und Geopolitik	59
5.3	Die Louvain-Methode	63
5.4	PageRank-Algorithmus	65
5.5	Grenzen	69
5.6	Zusammenfassung	71
6	Regressionsanalyse	73
6.1	Trendlinien	73
6.2	Beispiel: Vorhersage von Hauspreisen	74

6.3	Gradientenverfahren	78
6.4	Regressionskoeffizienten	80
6.5	Korrelationskoeffizienten	82
6.6	Grenzen	84
6.7	Zusammenfassung	85
7	<i>k</i>-nächste Nachbarn und Ausreißererkennung	87
7.1	Der Weindetektiv	87
7.2	Gleich und gleich gesellt sich gern	88
7.3	Beispiel: Der statistische Sommelier	90
7.4	Ausreißererkennung	92
7.5	Grenzen	94
7.6	Zusammenfassung	94
8	Support-Vektor-Maschine	97
8.1	„Nein“ oder „Oh Nein“?	97
8.2	Beispiel: Diagnose einer Herzerkrankung	98
8.3	Die optimale Grenzlinie	100
8.4	Grenzen	104
8.5	Zusammenfassung	105
9	Entscheidungsbaum	107
9.1	Wie man eine Katastrophe überlebt	107
9.2	Beispiel: Rettung von der Titanic	109

XIV Inhaltsverzeichnis

9.3	Einen Entscheidungsbaum erstellen	110
9.4	Grenzen	113
9.5	Zusammenfassung	115
10	Random Forests	117
10.1	Die Weisheit der Crowd	117
10.2	Beispiel: Verbrechensvorhersage	118
10.3	Ensembles	123
10.4	Bootstrap Aggregating	124
10.5	Grenzen	126
10.6	Zusammenfassung	127
11	Neuronale Netze	129
11.1	Bauen Sie sich ein Gehirn!	129
11.2	Beispiel: Handgeschriebene Zahlen erkennen	132
11.3	Wie ein neuronales Netz denkt	135
11.4	Aktivierungsregeln	139
11.5	Grenzen	140
11.6	Zusammenfassung	144
12	A/B-Tests und vielarmige Banditen	147
12.1	Grundlagen des A/B-Tests	147
12.2	Grenzen von A/B-Tests	148
12.3	Abnehmendes-Epsilon-Strategie	149
12.4	Beispiel: Vielarmige Banditen	150

12.5	Nett zu wissen: van Gaals Elfmeterschützen	153
12.6	Grenzen einer Abnehmendes- Epsilon-Strategie	154
12.7	Zusammenfassung	156
	Anhang	157
	Glossar	165
	Literatur	175

Über die Autoren

Annalyn Ng schloss ihr Studium an der University of Michigan in Ann Arbor ab, wo sie auch als Statistik-Tutorin arbeitete. Anschließend erwarb sie einen Master of Philosophy am Psychometrie-Zentrum der University of Cambridge, wo sie mithilfe von Data Mining an Daten aus sozialen Medien gezielte Werbekampagnen erarbeitete und kognitive Einstellungstests programmierte. Disney Research gewann sie später für ihr Behavioral-Sciences-Team, wo sie psychologische Kundenprofile erforschte.

Kenneth Soo erwarb seinen Master of Science in Statistik an der Stanford University. Vorher erzielte er während seiner drei Studienjahre an der University of Warwick durchgängig Top-Ergebnisse in „Mathematics, Operational Research, Statistics and Economics“ (MORSE). Dort arbeitete er auch als Research Assistant in der Gruppe

XVIII Über die Autoren

„Operational Research & Management Sciences“, und zwar über bikriterielle robuste Optimierung mit Anwendungen in zufallsfehleranfälligen Netzwerken.

Die Autoren freuen sich über Feedback zum Buch und sind hierfür unter contact@algobeans.com erreichbar.

Warum Data Science?

Stellen Sie sich vor, Sie sind eine junge Ärztin oder ein junger Arzt.

Ein Patient kommt in Ihre Klinik und klagt über Kurzatmigkeit, Brustschmerzen und gelegentliches Sodbrennen. Sie prüfen Blutdruck und Puls, beides normal, es gibt keine relevanten Vorerkrankungen.

Der Patient macht allerdings einen etwas rundlichen Eindruck. Und da seine Symptome ziemlich typisch für übergewichtige Menschen sind, versichern Sie ihm, dass alles unter Kontrolle ist, und raten für alle Fälle zu etwas sportlicher Betätigung.

Allzu oft führt eine solche Situation zu einer nicht diagnostizierten Herzerkrankung – denn Patienten mit Herzproblemen zeigen häufig ganz ähnliche Symptome wie sonst gesunde Übergewichtige. So werden weitere Tests

unterlassen, welche die ernstere Erkrankung aufgezeigt hätten (siehe hierzu auch Abschn. 8.2).

Unsere menschlichen Urteile werden von begrenzten, subjektiven Erfahrungen und unvollständigem Wissen geleitet. Dies beeinträchtigt unsere Entscheidungsfindung und kann, wie im Fall eines unerfahrenen Arztes, weitere Untersuchungen verhindern, welche zu akkurateren Schlüssen geführt hätten.

Dies ist die Stelle, an der Data Science helfen kann.

Anstatt sich auf das Urteil eines Einzelnen zu verlassen, erlauben uns Data-Science-Techniken, Informationen aus vielen Quellen zu verwerten, um so zu besseren Entscheidungen zu kommen. Wir könnten zum Beispiel archivierte Patientendaten auswerten, in denen ähnliche Symptome vorkommen, und so auf Diagnosen stoßen, auf die wir von allein nicht gekommen wären.

Mit modernen Rechnern und fortschrittlichen Algorithmen können wir

- versteckte Trends in großen Datensätzen aufspüren,
- mithilfe von Trends Vorhersagen treffen,
- die Wahrscheinlichkeit jedes möglichen Ergebnisses berechnen und
- schnell exakte Resultate erhalten.

Dieses Buch wurde, wie bereits im Vorwort angekündigt, in Laiensprache geschrieben und führt ganz sanft in die Konzepte und Algorithmen von Data Science ein. Dazu greifen wir auf intuitive Erklärungen und jede Menge Grafiken zurück.