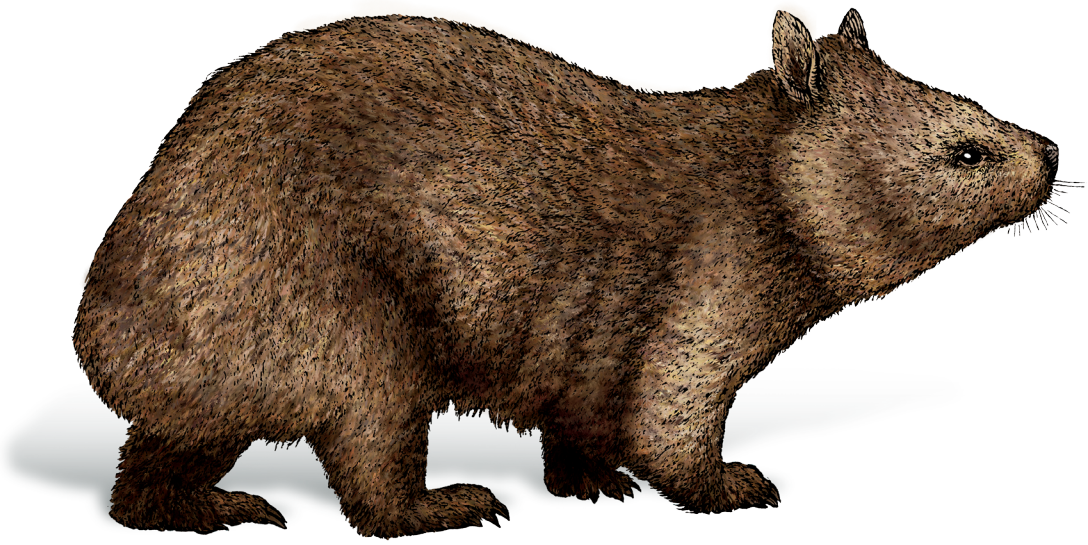
O'REILLY®

Deutsche Ausgabe

Praxisbuch Unsupervised Learning

Machine-Learning-Anwendungen für ungelabelte Daten mit Python programmieren



Ankur A. Patel

Übersetzung von Frank Langenau

Praxisbuch Unsupervised Learning

Machine-Learning-Anwendungen für ungelabelte Daten mit Python programmieren

Ankur A. Patel

Deutsche Übersetzung von Frank Langenau

O'REILLY®

Ankur A. Patel

Lektorat: Alexandra Follenius
Übersetzung: Frank Langenau
Korrektorat: Sibylle Feldmann, *www.richtiger-text.de*
Satz: III-Satz, *www.drei-satz.de*
Herstellung: Stefanie Weidner
Umschlaggestaltung: Karen Montgomery, Michael Oréal, *www.oreal.de*
Druck und Bindung: mediaprint solutions GmbH, 33100 Paderborn

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

ISBN:
Print 978-3-96009-127-1
PDF 978-3-96088-876-5
ePub 978-3-96088-877-2
mobi 978-3-96088-878-9

1. Auflage 2020

Wieblinger Weg 17
69123 Heidelberg

Hinweis:
Dieses Buch wurde auf PEFC-zertifiziertem Papier aus nachhaltiger Waldwirtschaft gedruckt. Der Umwelt zuliebe verzichten wir zusätzlich auf die Einschweißfolie.

Schreiben Sie uns:
Falls Sie Anregungen, Wünsche und Kommentare haben, lassen Sie es uns wissen: kommentar@oreilly.de.

5 4 3 2 1 0

Inhalt

Einleitung

Eine kurze Geschichte des maschinellen Lernens

Maschinelles Lernen ist ein Teilgebiet der *künstlichen Intelligenz* (KI, engl. *Artificial Intelligence*, AI), bei der Computer aus Daten lernen – üblicherweise, um ihre Performance für eine eng definierte Aufgabe zu verbessern –, ohne explizit dafür programmiert zu werden. Der Begriff *maschinelles Lernen* (engl. *Machine Learning*) wurde schon 1959 geprägt (von Arthur Samuel, einer Legende auf dem Gebiet der KI), doch im 21. Jahrhundert gab es nur wenige größere kommerzielle Erfolge im maschinellen Lernen zu verzeichnen. Stattdessen fristete das Gebiet ein Nischendasein im Rahmen wissenschaftlicher Forschungen an Universitäten.

Schon ziemlich früh (bereits in den 1960er-Jahren) waren viele Mitglieder der KI-Community viel zu optimistisch hinsichtlich der Zukunft der künstlichen Intelligenz. Forscher dieser Zeit, wie zum Beispiel Herbert Simon und Marvin Minsky, behaupteten, dass die KI innerhalb von Jahrzehnten das Niveau der menschlichen Intelligenz erreichen würde:[1]

> *Innerhalb von zwanzig Jahren werden Maschinen in der Lage sein, jede Arbeit zu verrichten, zu der ein Mensch fähig ist.*
>
> – Herbert Simon, 1965

> *In drei bis acht Jahren werden wir eine Maschine mit der allgemeinen Intelligenz eines durchschnittlichen Menschen haben.*
>
> – Marvin Minsky, 1970

Von ihrem Optimismus geblendet, konzentrierten sich Forscher auf Projekte der sogenannten *starken KI* oder *allgemeinen künstlichen Intelligenz* (engl. *Artificial General Intelligence*, AGI), um damit KI-Agenten zu schaffen, die Problemlösung, Wissensdarstellung, Lernen und Planen, Natural Language Processing, Wahrnehmung und Bewegungskontrolle realisieren können. Zwar half dieser Optimismus,

1 Inspiriert von solchen Ansichten kreierte Stanley Kubrick 1968 im Film *2001: Odyssee im Weltraum* den KI-Agenten HAL 9000.

beträchtliche Mittel von großen Akteuren wie z. B. dem Verteidigungsministerium zu beschaffen, doch nahmen diese Forscher zu anspruchsvolle Probleme in Angriff und waren letztlich zum Scheitern verurteilt.

Die KI-Forschung schaffte nur gelegentlich den Sprung vom akademischen Umfeld in die Industrie, und es folgte eine Reihe sogenannter KI-Winter. In diesen KI-Wintern (eine Analogie, die sich am nuklearen Winter in der Ära des Kalten Kriegs orientierte) gingen das Interesse an der KI und ihre Finanzierung zurück. Gelegentlich auftretende Hype-Zyklen um KI hielten kaum an. Anfang der 1990er-Jahre hatte das Interesse an der KI und ihrer Finanzierung einen Tiefpunkt erreicht.

KI ist zurück, aber warum gerade jetzt?

KI ist in den letzten zwei Jahrzehnten mit Vehemenz wieder aufgetaucht – zuerst als rein akademischer Interessenbereich und jetzt inzwischen als ausgewachsenes Gebiet, das die hellsten Köpfe von Universitäten wie auch von Unternehmen in ihren Bann zieht.

Drei entscheidende Entwicklungen stehen hinter diesem Wiederaufleben: Durchbrüche bei den Algorithmen für maschinelles Lernen, die Verfügbarkeit großer Datenbestände und superschnelle Computer.

Erstens haben Forscher ihre Aufmerksamkeit auf eng definierte Teilprobleme der starken KI gerichtet, auch als *schwache KI* bezeichnet, anstatt sich auf übermäßig ambitionierte starke KI-Projekte zu versteifen. Dieser Fokus auf die Verbesserung von Lösungen für eng definierte Aufgaben führte zu algorithmischen Durchbrüchen, die den Weg für erfolgreiche kommerzielle Anwendungen ebneten. Viele dieser Algorithmen – oftmals ursprünglich an Universitäten oder privaten Forschungseinrichtungen entwickelt – wurden schnell als Open Source zugänglich gemacht, was die Akzeptanz dieser Technologien durch die Industrie beschleunigte.

Zweitens wurde die Datenerfassung zu einem Schwerpunkt für die meisten Unternehmen, und die Kosten für das Speichern der Daten fielen aufgrund der Fortschritte in der digitalen Datenspeicherung drastisch. Dank des Internets wurden Unmengen von Daten auch in einem noch nie gekannten Umfang weithin und öffentlich zugänglich.

Drittens wurden die Computer immer leistungsfähiger und über die Cloud verfügbar, sodass KI-Forscher ihre IT-Infrastruktur bei Bedarf einfach und preiswert skalieren konnten, ohne zunächst riesige Mittel in Hardware zu investieren.

Das Entstehen der angewandten KI

Die oben genannten Kräfte haben die KI aus dem akademischen Umfeld in die Industrie befördert und dazu beigetragen, das Interesse und die Finanzierung von Jahr zu Jahr auf ein höheres Niveau zu heben. KI ist nicht mehr nur ein theoretischer Interessenbereich, sondern ein vollwertiges Anwendungsgebiet. Abbildung 1

zeigt ein Diagramm aus Google Trends, das das wachsende Interesse am maschinellen Lernen im Verlauf der letzten fünf Jahre darstellt.

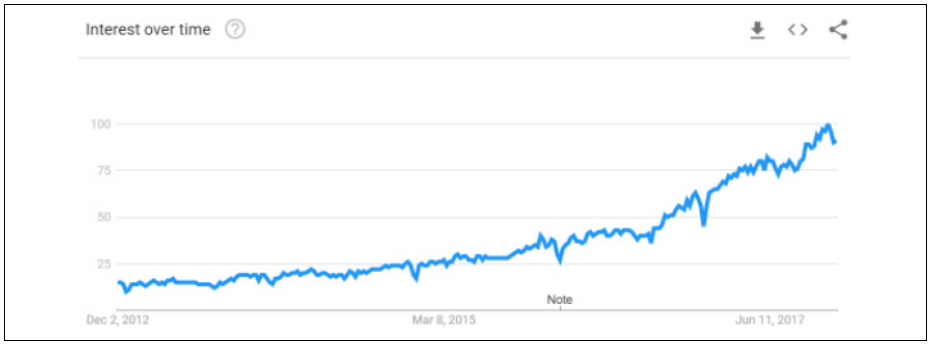


Abbildung 1: Interesse am maschinellen Lernen in den letzten Jahren

KI gilt heute als bahnbrechende horizontale Technologie – ähnlich dem Aufkommen von Computern und Smartphones –, die in den nächsten zehn Jahren erhebliche Auswirkungen auf jede einzelne Branche haben wird.[2]

Zu den erfolgreichen kommerziellen Anwendungen, die sich auf maschinelles Lernen stützen, gehören unter anderem optische Zeichenerkennung, Filtern von Spam-Mails, Bildklassifizierung, Computervision, Spracherkennung, maschinelle Übersetzung, Gruppensegmentierung und Clustering, Generieren von synthetischen Daten, Anomalieerkennung, Prävention von Cyberkriminalität, Erkennung von Kreditkartenbetrug, Erkennung von Betrug im Internet, Zeitreihenvorhersage, Natural Language Processing, Brett- und Videospiele, Dokumentklassifizierung, Empfehlungssysteme, Suchen, Robotik, Onlinewerbung, Sentimentanalyse, DNA-Sequenzierung, Finanzmarktanalyse, Informationsgewinnung, Beantwortung von Fragen und Entscheidungsfindung im Gesundheitswesen.

Meilensteine der angewandten KI in den letzten 20 Jahren

Die hier beschriebenen Meilensteine halfen, die KI von einem meist akademischen Gesprächsthema zu einem wichtigen Bestandteil der heutigen Technologie zu machen.

- 1997: Deep Blue, ein KI-Bot, der seit Mitte der 1980er-Jahre entwickelt wird, schlägt den Schachweltmeister Garry Kasparov in einem medienwirksamen Schachereignis.
- 2004: Die DARPA führt die DARPA Grand Challenge ein, einen in der Mojave-Wüste stattfindenden Wettbewerb für unbemannte Landfahrzeuge. Im Jahr 2005 erhält Stanford den Hauptpreis. Im Jahr 2007 veranstaltet die Carnegie

2 Laut McKinsey Global Institute könnte sich bis 2055 mehr als die Hälfte aller beruflichen Aktivitäten, für die Menschen bezahlt werden, automatisieren lassen.

Mellon University diesen Wettbewerb in einem städtischen Umfeld. Bis 2015 haben viele große Technologieunternehmen, darunter Tesla, Waymo von Alphabet und Uber, finanziell gut ausgestattete Programme aufgelegt, um eine allgemein verfügbare Selbstfahrtechnologie aufzubauen.

- 2006: Geoffrey Hinton von der University of Toronto stellt einen schnellen Lernalgorithmus vor, um neuronale Netze mit vielen Schichten zu trainieren, und leitet damit die Deep-Learning-Revolution ein.
- 2006: Netflix startet den mit einer Million Dollar dotierten Wettbewerb Netflix Prize, bei dem die Teams durch maschinelles Lernen die Genauigkeit ihres Empfehlungssystems um wenigstens 10 % verbessern sollen. Im Jahr 2009 hat zum ersten Mal ein Team diesen Preis gewonnen.
- 2007: KI erreicht übermenschliche Performance im Damespiel, was von einem Team der University of Alberta erreicht wurde.
- 2010: ImageNet startet einen jährlichen Wettbewerb – die *ImageNet Large Scale Visual Recognition Challenge* (ILSVRC) –, bei der Teams mithilfe von Algorithmen des maschinellen Lernens Objekte in einem großen, gut gepflegten Bild-Dataset korrekt erkennen und klassifizieren. Sowohl Akademiker als auch Technologieriesen sind stark daran interessiert. Der Klassifizierungsfehler fällt von 25 % im Jahr 2011 auf nur wenige Prozent bis 2015, was Fortschritten bei tiefen Faltungsnetzen zu verdanken ist. Dies führt zu kommerziellen Anwendungen von Computervision und Objekterkennung.
- 2010: Microsoft bringt die Steuerung Kinect für die Spielkonsole Xbox 360 auf den Markt. Die vom Computervision-Team bei Microsoft Research entwickelte Kinect kann Körperbewegungen des Menschen verfolgen und in Softwarebefehle zur Steuerung von Videospielen übersetzen.
- 2010: Siri, einer der ersten allgemein verfügbaren digitalen Sprachassistenten, wird von Apple übernommen und im Oktober 2011 als Teil des iPhone 4S veröffentlicht. Schließlich führt Apple Siri für alle seine Produkte ein. Auf der Basis von Convolutional Neural Networks (Faltungsnetzen) und rekurrenten neuronalen *Long-Short-Term-Memory*-Netzwerken beherrscht Siri sowohl die Spracherkennung als auch das Natural Language Processing. Schließlich greifen auch Amazon, Microsoft und Google mit Alexa (2014), Cortana (2014) sowie Google Assistant (2016) ins Rennen ein.
- 2011: IBM Watson, ein Fragen beantwortender KI-Agent, der von einem Team unter der Leitung von David Ferruci entwickelt wurde, schlägt die ehemaligen Jeopardy!-Gewinner Brad Rutter und Ken Jennings. IBM Watson wird heute in mehreren Branchen eingesetzt, darunter im Gesundheitswesen und im Einzelhandel.
- 2012: Das Google-Brain-Team unter Leitung von Andrew Ng und Jeff Dean trainiert ein neuronales Netz, um Katzen auf unbezeichneten Bildern aus YouTube-Videos zu erkennen.

- 2013: Google gewinnt die Robotics Challenge der DARPA, bei der teilautonome Bots komplexe Aufgaben in tückischen Umgebungen ausführen, beispielsweise ein Fahrzeug führen, durch Trümmer gehen, Schutt aus einem blockierten Eingang wegräumen, eine Tür öffnen und eine Leiter hochsteigen.
- 2014: Facebook veröffentlicht Arbeiten zu DeepFace, einem auf neuronalen Netzen basierenden System, das Gesichter mit einer Genauigkeit von 97% identifizieren kann. Dies entspricht nahezu der Leistung, die ein Mensch erreicht, und bedeutet eine Verbesserung von mehr als 27% gegenüber früheren Systemen.
- 2015: KI wird salonfähig und ist häufig Thema in Medienkanälen auf der ganzen Welt.
- 2015: AlphaGo von Google DeepMind schlägt den Weltklasseprofi Fan Hui im Spiel Go. Im Jahr 2016 besiegt AlphaGo Lee Sedol und 2017 Ke Jie. Die neue Version AlphaGo Zero besiegt im Jahr 2017 die vorherige AlphaGo-Version mit 100 zu 0. AlphaGo Zero bezieht unüberwachte Lerntechniken ein und meistert Go, indem es gegen sich selbst spielt.
- 2016: Google startet eine umfassende Überarbeitung seiner Sprachübersetzung Google Translate, bei der das vorhandene phrasengestützte Übersetzungssystem durch ein auf Deep Learning basierendes neuronales Maschinenübersetzungssystem ersetzt wird, was Übersetzungsfehler um bis zu 87% reduziert und sich einer Genauigkeit auf menschlichem Niveau nähert.
- 2017: Das von Carnegie Mellon entwickelte KI-Programm Libratus gewinnt beim Eins-gegen-eins-Poker in der Variante Texas Hold'em Heads-Up No Limit, d.h., es spielen zwei Spieler mit je zwei verdeckten Handkarten.
- 2017: Der von OpenAI trainierte Bot schlägt professionelle Spieler beim Dota-2-Turnier.

Von schwacher KI zu AGI

Natürlich sind diese Erfolge bei der Anwendung der KI auf eng definierte Probleme lediglich ein Ausgangspunkt. Die KI-Community glaubt zunehmend daran, dass wir – durch Kombination mehrerer schwacher KI-Systeme – starke KI entwickeln können. Dieser starke KI- oder AGI-Agent wird in der Lage sein, bei vielen breit definierten Aufgaben Leistungen auf Augenhöhe mit dem Menschen zu erbringen.

Bald nachdem die KI eine Performance auf menschlichem Niveau erreicht hat, wird diese starke KI die menschliche Intelligenz übertreffen und eine sogenannte *Superintelligenz* erreichen – so die Voraussagen einiger Forscher. Schätzungen für das Erreichen einer derartigen Superintelligenz reichen von mindestens 15 Jahren bis zu 100 Jahren, wobei aber die meisten Forscher davon überzeugt sind, dass sich KI schnell genug entwickelt, um es in wenigen Generationen zu schaffen. Ist dieser Hype erneut zu aufgebläht (wie in den vorherigen KI-Zyklen), oder ist das Ganze dieses Mal anders? Nur die Zeit wird es zeigen.

Ziel und Vorgehensweise

Die meisten der heute kommerziell erfolgreichen Anwendungen – in Bereichen wie Computervision, Spracherkennung, Maschinenübersetzung und Natural Language Processing – arbeiten mit Supervised Learning, das von *gelabelten Datasets* profitiert. Die meisten Daten in der Welt sind jedoch *ungelabelt*.

In diesem Buch geht es um den Bereich des *Unsupervised Learning* (einen Zweig des maschinellen Lernens, mit dem sich verdeckte Muster finden lassen) und des Lernens der zugrunde liegenden Struktur in ungelabelten Daten. Nach Ansicht vieler Branchenexperten, wie etwa Yann LeCun, Director of AI Research bei Facebook und Professor an der NYU, ist Unsupervised Learning die nächste große Herausforderung in der KI und kann den Schlüssel zur AGI beinhalten. Aus diesem und vielen anderen Gründen gehört Unsupervised Learning heute zu den gängigsten Themen in der KI.

Das Buch skizziert die Konzepte und Tools, die Sie brauchen, damit Sie die erforderliche Intuition entwickeln können, um diese Technik jeden Tag auf Probleme anzuwenden, an denen Sie arbeiten. Mit anderen Worten: Dies ist ein Praxisbuch, das Sie in die Lage versetzt, reale Systeme aufzubauen. Wir untersuchen auch, wie sich ungelabelte Datasets benennen lassen, um unüberwachte in semiüberwachte Lernprobleme zu überführen.

Der praxisorientierte Ansatz dieses Buchs bietet darüber hinaus einige theoretische Einführungen, die sich vorwiegend auf die Anwendung von unüberwachten Lerntechniken konzentrieren, um praktische Probleme zu lösen. Die Datasets und die Codebeispiele sind online als Jupyter Notebooks auf GitHub (*http://bit.ly/2Gd4v7e*) verfügbar.

Ausgerüstet mit dem konzeptionellen Verständnis und den praktischen Erfahrungen, die Sie sich mit diesem Buch aneignen, werden Sie unter anderem in der Lage sein, Unsupervised Learning auf große, ungelabelte Datasets anzuwenden, um versteckte Muster aufzudecken, tiefergehende Einblicke in Unternehmen zu erhalten, Anomalien zu erkennen, Gruppen nach Ähnlichkeiten zu clustern, automatische Merkmalskonstruktion (Feature Engineering) und Merkmalsauswahl (Feature Selection) durchzuführen sowie synthetische Datasets zu generieren.

Voraussetzungen

In diesem Buch gehen wir davon aus, dass Sie über Erfahrungen im Programmieren mit Python verfügen und insbesondere mit NumPy und Pandas vertraut sind.

Mehr zu Python finden Sie auf der offiziellen Python-Website (*https://www.python.org/*). Für Jupyter Notebooks sei auf die offizielle Jupyter-Website (*http://jupyter.org/index.html*) verwiesen. Um Ihr Wissen auf dem Niveau von Hochschulmathematik, linearer Algebra, Wahrscheinlichkeitsrechnung und Statistik aufzufrischen, lesen Sie am besten Teil 1 des *Deep Learning*-Lehrbuchs von Ian Goodfellow und

Yoshua Bengio (*http://www.deeplearningbook.org/*). Für eine Auffrischung in Bezug auf maschinelles Lernen sollten Sie *The Elements of Statistical Learning* (*https://stanford.io/2Tju4al*) lesen.

Roadmap

Das Buch ist in vier Teile gegliedert, die sich mit den folgenden Themen befassen:

Teil I, Grundlagen des Unsupervised Learning
: Unterschiede zwischen Supervised und Unsupervised Learning, ein Überblick über bekannte überwachte und unüberwachte Algorithmen sowie ein durchgängiges Projekt zum maschinellen Lernen.

Teil II, Unsupervised Learning mit Scikit-learn
: Dimensionalitätsreduktion, Anomalieerkennung sowie Clustering und Gruppensegmentierung.

Weitere Informationen zu den Konzepten, die in den Teilen A und B diskutiert werden, finden Sie in der Dokumentation zu Scikit-learn (*https://scikit-learn.org/stable/modules/classes.html*).

Teil III, Unsupervised Learning mit TensorFlow und Keras
: Repräsentationslernen und automatische Feature Extraction, Autoencoder und Semi-supervised Learning.

Teil IV, Deep Unsupervised Learning mit TensorFlow und Keras
: Eingeschränkte Boltzmann-Maschinen, Deep-Belief-Netze und Generative Adversarial Networks.

Konventionen, die in diesem Buch verwendet werden

In diesem Buch werden folgende typografische Konventionen verwendet:

Kursiv
: Kennzeichnet neue Begriffe, URLs, E-Mail-Adressen, Dateinamen und Dateierweiterungen.

`Schreibmaschinenschrift`
: Wird in Programmlistings verwendet und im Fließtext für Programmelemente wie zum Beispiel Variablen- oder Funktionsnamen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter.

`Schreibmaschinenschrift fett`
: Kennzeichnet Befehle oder andere Texte, die vom Benutzer buchstäblich eingegeben werden sollen.

Schreibmaschinenschrift kursiv

Zeigt Text, der ersetzt werden soll durch Werte, die der Benutzer bereitstellt, oder Werte, die sich aus dem Kontext ergeben.

Dieses Element kennzeichnet einen Tipp oder Vorschlag.

Dieses Element kennzeichnet einen allgemeinen Hinweis.

Codebeispiele verwenden

Ergänzungsmaterialien (Codebeispiele usw.) stehen auf GitHub zum Download bereit (*http://bit.ly/2Gd4v7e*).

Dieses Buch soll Ihnen bei Ihrer täglichen Arbeit helfen. Falls Beispielcode zum Buch angeboten wird, dürfen Sie ihn im Allgemeinen in Ihren Programmen und für Dokumentationen verwenden. Sie müssen uns nicht um Erlaubnis bitten, es sei denn, Sie kopieren einen erheblichen Teil des Codes. Wenn Sie zum Beispiel ein Programm schreiben, das einige Codeblöcke aus diesem Buch verwendet, benötigen Sie keine Erlaubnis. Sollten Sie aber eine CD-ROM mit den Beispielen von O'Reilly-Büchern verkaufen oder verteilen, ist eine Erlaubnis erforderlich. Wenn Sie eine Frage beantworten und dabei dieses Buch oder Beispielcode aus diesem Buch zitieren, brauchen Sie wiederum keine Erlaubnis. Aber wenn Sie erhebliche Teile des Beispielcodes aus diesem Buch in die Dokumentation Ihres Produkts einfließen lassen, ist eine Erlaubnis einzuholen.

Wir schätzen eine Quellenangabe, verlangen sie aber nicht. Eine Quellenangabe umfasst in der Regel Titel, Autor, Verlag und ISBN. Zum Beispiel: *Praxisbuch Unsupervised Learning* von Ankur A. Patel (O'Reilly). Copyright 2019, Human AI Collaboration, Inc., 978-3-96009-127-1.«

Wenn Sie der Meinung sind, dass Sie die Codebeispiele in einer Weise verwenden, die über die oben erteilte Erlaubnis hinausgeht, kontaktieren Sie uns bitte unter *kommentar@oreilly.de*.

Danksagungen

Das ganze Jahr 2018 war eine unglaubliche Reise, manchmal frustrierend, doch überwiegend voller Freude. Ich möchte meinen ehemaligen ThetaRay-Kollegen dafür danken, dass sie mir geholfen haben, Unsupervised Learning zu erkunden.

Insbesondere haben Mark Gazit, Amir Averbuch, David Segev, Gil Shabat, Ovad Harari und Udi Menkes zu diesem Prozess beigetragen.

Ich möchte auch meinen derzeitigen Kollegen bei 7Park Data – insbesondere Brian Lichtenberger, Alex Nephew und Rishit Shah – dafür danken, dass sie mir die Möglichkeit gegeben haben, meinen Hintergrund zu maschinellem Lernen zu nutzen und auf Dutzende sehr interessanter alternativer Datasets anzuwenden. Mein Dank geht auch an Vista Equity Partners, die mir eine umfangreiche Plattform für den KI-Bereich zur Verfügung gestellt haben.

Besonderer Dank gilt Sarah Nagy, Charles Givre, Matthew Harrison und Eric Perkins. Diese großzügigen Menschen haben unzählige Stunden damit verbracht, mein Buch und meinen Code im Detail zu überprüfen. Ohne sie wäre dieses Buch nicht annähernd so ausgefeilt geworden, wie es heute ist.

Die Zusammenarbeit mit dem O'Reilly-Team hat mir durchweg Spaß gemacht, und ich freue mich, auch in den kommenden Jahren mit ihm zusammenzuarbeiten. Die Mitarbeiter des Teams haben den gesamten Schreibprozess fast zum Kinderspiel werden lassen, und zwar von der ersten Konzeption des Werks bis zur endgültigen Produktion. Insbesondere meine Redakteure Michele Cronin und Nicole Tache waren während des gesamten Ablaufs sehr aufmerksam und geduldig und haben das Projekt bei jedem Schritt begleitet.

Großer Dank geht an Rachel Roumeliotis, die an dieses Projekt von Anfang an geglaubt und den Startschuss dafür gegeben hat. Melanie Yarbrough, Katherine Tozer, Jasmine Kwityn, Jonathan Hassell, Eszter Schoell, Daisy Wizda und Scott Murray haben alle herausragende Rollen gespielt, indem sie sowohl die Herstellung des Buchs als auch die Bereitstellung der zusätzlichen Onlinematerialien ermöglicht haben. Ich bin wirklich dankbar dafür, mit ihnen zusammenarbeiten zu dürfen.

Zu guter Letzt habe ich das Glück, dass mich wunderbare Menschen in meinem Leben bei jedem Schritt unterstützen. Danken möchte ich dabei vor allem meinen Eltern Amrat und Ila, meiner Schwester Bhavini und meinem Bruder Jigar. Und natürlich bin ich meiner Freundin Maria Koval auf ewig dankbar, die meine ständige Meisterin ist. Ich bin sehr glücklich, sie in meinem Leben zu haben.

TEIL I

Grundlagen des Unsupervised Learning

Zu Beginn untersuchen wir das aktuelle Ökosystem des maschinellen Lernens und sehen uns an, wo sich Unsupervised Learning einordnen lässt. Außerdem erstellen wir von Grund auf ein Projekt, an dem sich die Grundlagen maschinellen Lernens zeigen lassen – d.h. die Programmierumgebung einrichten, Daten erfassen und vorbereiten, Daten untersuchen, Algorithmen des maschinellen Lernens und Kostenfunktionen auswählen sowie die Ergebnisse auswerten.

KAPITEL 1

Unsupervised Learning im Ökosystem des maschinellen Lernens

Das Lernen bei Mensch und Tier geschieht überwiegend unüberwacht. Betrachtet man Intelligenz als Kuchen, wäre Unsupervised Learning der Kuchen selbst, Supervised Learning das Sahnehäubchen auf dem Kuchen und Reinforcement Learning die Praline on top. Wir wissen zwar, wie wir das Sahnehäubchen und die Praline hinbekommen, doch wir wissen nicht, wie man den Kuchen macht. Wir müssen das Problem des Unsupervised Learning lösen, bevor wir überhaupt daran denken können, zu echter KI zu gelangen.

– Yann LeCun

In diesem Kapitel untersuchen wir den Unterschied zwischen einem regelbasierten System und maschinellem Lernen, den Unterschied zwischen Supervised Learning (überwachtem Lernen) und Unsupervised Learning (unüberwachtem Lernen) sowie die relativen Stärken und Schwächen beider Methoden.

Außerdem befassen wir uns mit vielen bekannten Algorithmen des Supervised und Unsupervised Learning und untersuchen auch, wie Semi-supervised Learning (semiüberwachtes Lernen) und Reinforcement Learning (bestärkendes Lernen) in diese Mischung passen.

Grundbegriffe des maschinellen Lernens

Bevor wir uns eingehend mit den verschiedenen Typen des maschinellen Lernens befassen, werfen wir einen Blick auf ein einfaches und häufig verwendetes Beispiel für maschinelles Lernen, um die eingeführten Konzepte verständlicher zu machen: den E-Mail-Spam-Filter. Wir möchten ein einfaches Programm erstellen, das E-Mails korrekt entweder als »Spam« oder als »nicht Spam« klassifiziert. Dies ist ein simples Klassifizierungsproblem.

Zur Auffrischung folgt erst mal etwas Terminologie zum maschinellen Lernen: die *Eingabevariablen* für dieses Problem sind die Wörter im Text der E-Mails. Diese Eingabevariablen bezeichnet man auch als *Features*, *Prädiktoren* oder *unabhängige Variablen*. Die Ausgabevariable – die wir vorhersagen wollen – ist die Bezeichnung »Spam« oder »nicht Spam«. Man bezeichnet sie auch als *Zielvariable*, *abhängige*

Variable oder *Antwortvariable* (oder Klasse, da es sich um ein Klassifizierungsproblem handelt).

Die Menge der Beispiele, mit denen KI trainiert, ist das sogenannte *Trainingsset*. Die einzelnen Beispiele des Trainingssets heißen *Trainingsinstanzen* oder *Stichproben*. Während des Trainings versucht die KI, ihre *Kostenfunktion* oder *Fehlerrate* zu minimieren oder, positiver formuliert, ihre *Wertfunktion* zu maximieren – in diesem Fall die Quote der korrekt klassifizierten E-Mails. Die KI optimiert während des Trainings aktiv auf eine minimale Fehlerrate. Die Fehlerrate wird berechnet, indem die von der KI vorhergesagte Benennung mit der wahren Benennung verglichen wird.

Allerdings interessiert uns am meisten, wie die KI ihr Training auf noch nie zuvor gesehene E-Mails verallgemeinert. Dies wird der ultimative Test für die KI: Kann sie E-Mails, die sie zuvor nicht gesehen hat, durch das, was sie beim Training auf den Beispielen im Trainingsset gelernt hat, korrekt klassifizieren? Dieser *Generalisierungsfehler* oder *Out-of-sample-Fehler* ist das wichtigste Instrument, mit dem wir Lösungen des maschinellen Lernens bewerten.

Der Satz von noch nie zuvor gesehenen Beispielen ist das sogenannte *Testset* oder *Holdout-Set* (weil die Daten aus dem Training herausgehalten werden). Wenn wir uns für mehrere Holdout-Sets entscheiden (um etwa den Generalisierungsfehler beim Training abzuschätzen, was ratsam ist), können wir zwischengeschaltete Holdout-Sets einrichten, mit denen wir unseren Fortschritt bewerten, bevor wir zum endgültigen Testset kommen. Diese zwischengeschalteten Holdout-Sets sind die sogenannten *Validierungssets*.

Alles in allem trainiert KI auf den Trainingsdaten (*Erfahrung*), um beim Markieren von Spam (*Aufgabe*) die Fehlerrate zu minimieren (*Performance*). Das ultimative Erfolgskriterium besteht darin, wie gut sich die gewonnenen Erfahrungen auf neue, zuvor nicht gesehene Daten verallgemeinern lassen (*Generalisierungsfehler*).

Regelbasiertes vs. maschinelles Lernen

Nach einem regelbasierten Ansatz können wir einen Spam-Filter mit expliziten Regeln entwerfen, um Spam zu erfassen, wie zum Beispiel E-Mails, die »u« anstelle von »you«, »4« anstelle von »for« oder »BUY NOW« usw. verwenden. Allerdings wäre dieses System mit der Zeit schwer zu pflegen, da Betrüger ihr Spam-Verhalten ständig anpassen, um solche Regeln zu umgehen. Bei einem regelbasierten System müssten wir häufig die Regeln manuell anpassen, um auf dem Laufenden zu bleiben. Zudem wäre es ein recht teures Setup – denken Sie nur an all die Regeln, die wir erzeugen müssen, um ein brauchbares System zu bekommen.

Anstelle eines regelbasierten Ansatzes können wir auf maschinelles Lernen zurückgreifen, um mit E-Mail-Daten zu trainieren und automatisch Regeln erzeugen zu

lassen, die bösartige E-Mails korrekt als Spam markieren. Dieses System auf Basis maschinellen Lernens könnte auch im Laufe der Zeit automatisch angepasst werden. Der Aufwand für Training und Verwaltung wäre bei einem solchen System viel geringer.

Bei diesem einfachen E-Mail-Problem könnten wir die Regeln zwar auch manuell erstellen, doch bei vielen Problemen ist eine solche Vorgehensweise überhaupt nicht praktikabel. Nehmen Sie beispielsweise an, ein selbstfahrendes Auto zu entwerfen – stellen Sie sich vor, welche Regeln Sie formulieren müssten, um das Verhalten des Fahrzeugs in jeder nur denkbaren Situation zu beschreiben. Dies ist ein unlösbares Problem, es sei denn, das Fahrzeug kann lernen und sich selbst aufgrund seiner Erfahrungen anpassen.

Systeme des maschinellen Lernens ließen sich auch als Erkundungs- oder Datenerkennungstool einsetzen, um tiefere Einblicke in das zu lösende Problem zu gewinnen. So können wir im Beispiel des E-Mail-Filters lernen, welche Wörter oder Phrasen am ehesten auf Spam hinweisen, und neu entstehende bösartige Spam-Muster erkennen.

Supervised vs. Unsupervised

Das Gebiet des maschinellen Lernens gliedert sich in zwei Bereiche – Supervised und Unsupervised Learning (*überwachtes* und *unüberwachtes* Lernen) – sowie viele Unterbereiche, die die beiden überbrücken.

Beim überwachten Lernen hat der KI-Agent Zugriff auf *Labels* (Kennungen), mit deren Hilfe er seine Performance bei einer Aufgabe verbessern kann. Beim Problem des E-Mail-Spam-Filters haben wir ein Dataset von E-Mails mit dem gesamten Text jeder einzelnen E-Mail. Außerdem wissen wir, welche dieser E-Mails Spam sind und welche nicht (anhand der sogenannten Labels). Diese Labels sind sehr wertvoll, da sie der überwacht lernenden KI helfen, die Spam-E-Mails vom Rest zu trennen.

Beim unüberwachten Lernen sind keine Labels verfügbar. Demzufolge ist die Aufgabe des KI-Agenten nicht klar umrissen, und die Performance lässt sich nicht so deutlich messen. Nehmen wir das Problem des E-Mail-Spam-Filters – dieses Mal ohne Labels. Nun versucht der Agent, die zugrunde liegende Struktur von E-Mails zu verstehen, wobei er die Datenbank mit den E-Mails in verschiedene Gruppen einteilt, sodass E-Mails innerhalb einer Gruppe einander ähnlich sind, sich aber von E-Mails in anderen Gruppen unterscheiden.

Dieses unüberwachte Lernproblem ist weniger klar definiert als das überwachte Lernproblem und für den KI-Agenten schwieriger zu lösen. Doch wenn man es richtig angeht, ist die Lösung letztendlich leistungsfähiger.

Und hier ist der Grund dafür: Die unüberwacht lernende KI kann mehrere Gruppen finden, die sie später als »Spam« markiert – doch die KI kann auch Gruppen

finden, die sie später als »wichtig« kennzeichnet oder als »Familie«, »beruflich«, »Nachrichten«, »Shopping« usw. Mit anderen Worten: Da das Problem keine streng definierte Aufgabe hat, kann der KI-Agent weit mehr interessante Muster finden als die, nach denen man anfangs gesucht hat.

Darüber hinaus ist dieses unüberwachte System besser als das überwachte System, wenn es darum geht, neue Muster in zukünftigen Daten zu finden. Das macht die unüberwachte Lösung auf längere Sicht flexibler. Dies ist die Stärke des Unsupervised Learning.

Die Stärken und Schwächen des Supervised Learning

Supervised Learning zeichnet sich dadurch aus, dass es die Performance in genau definierten Aufgaben mit zahlreichen Labels optimiert. Nehmen Sie zum Beispiel ein sehr großes Dataset mit Bildern von Objekten, bei dem jedes Bild gelabelt ist. Wenn das Dataset genügend groß ist, wir mit dem richtigen Algorithmus für maschinelles Lernen (d.h. Convolutional Neural Networks) trainieren und zudem ausreichend leistungsfähige Computer einsetzen, können wir ein sehr gutes Bildklassifizierungssystem auf Basis von Supervised Learning erstellen.

Da die überwacht lernende KI auf den Daten trainiert, wird sie in der Lage sein, ihre Performance (über eine Kostenfunktion) zu messen, indem sie ihr vorhergesagtes Bild-Label mit dem wahren Bild-Label, das in der Datei verzeichnet ist, vergleicht. Die KI wird ausdrücklich versuchen, diese Kostenfunktion zu minimieren, sodass ihr Fehler bei niemals zuvor gesehenen Bildern (aus einem Holdout-Set) so gering wie möglich ist.

Aus diesem Grund sind Labels so leistungsfähig – sie helfen, den KI-Agenten zu führen, indem sie ihn mit einem Fehlermaß versorgen. Die KI verwendet das Fehlermaß, um ihre Performance im Laufe der Zeit zu verbessern. Ohne derartige Labels weiß die KI nicht, wie erfolgreich sie Bilder korrekt klassifizieren kann (oder nicht kann).

Allerdings sind die Kosten für das manuelle Labeln eines Bild-Datasets sehr hoch. Und selbst die am besten kuratierten Bild-Datasets enthalten nur Tausende von Labels. Dies ist ein Problem, weil überwachte Lernsysteme Bilder von Objekten, für die Labels existieren, sehr gut klassifizieren können, aber schlecht abschneiden bei Bildern von Objekten, für die es keine Labels gibt. So leistungsfähig überwachte Lernsysteme sind, so begrenzt sind sie auch, Wissen zu verallgemeinern, das über die gelabelten Elemente hinausgeht, mit denen sie trainiert wurden. Da die meisten Daten in der Welt nicht gelabelt sind, ist die Fähigkeit der KI, ihre Performance auf nie zuvor gesehene Instanzen auszudehnen, recht beschränkt.

Mit anderen Worten: Supervised Learning ist hervorragend geeignet, um eng begrenzte KI-Probleme zu lösen, aber nicht so gut für das Lösen anspruchsvollerer, weniger klar definierter Probleme der starken KI.

Die Stärken und Schwächen des Unsupervised Learning

Bei eng definierten Aufgaben, für die wir klar definierte Muster haben, die sich im Laufe der Zeit nicht wesentlich ändern, und für ausreichend große Datasets, die unmittelbar zugänglich und gelabelt sind, wird Supervised Learning das Unsupervised Learning schlagen.

Bei Problemen aber, bei denen die Muster unbekannt sind oder sich ständig ändern oder für die wir keine ausreichend großen gelabelten Datasets haben, wird Unsupervised Learning wirklich glänzen.

Anstatt sich von Labels leiten zu lassen, lernt Unsupervised Learning die zugrunde liegende Struktur der Daten, mit denen es trainiert hat. Dazu wird versucht, die Trainingsdaten mit einem Satz von Parametern darzustellen, der deutlich kleiner ist als die Anzahl der Beispiele, die im Dataset verfügbar sind. Dieses Lernen von Datenrepräsentationen ermöglicht dem unüberwachten Lernen, unterschiedliche Muster im Dataset zu identifizieren.

Im Beispiel des Bild-Datasets (dieses Mal ohne Label) kann die unüberwacht lernende KI in der Lage sein, Bilder zu identifizieren und zu gruppieren, und zwar basierend darauf, wie ähnlich sie sich einander sind und wie sehr sie sich vom Rest unterscheiden. Zum Beispiel werden alle Bilder, auf denen Stühle zu sehen sind, in einer Gruppe zusammengefasst, alle Bilder, die Hunde zeigen, in einer anderen Gruppe usw.

Natürlich kann die unüberwacht lernende KI diese Gruppen nicht selbst als »Stühle« oder »Hunde« kennzeichnen. Da aber ähnliche Bilder jetzt gruppiert sind, ist es für den Menschen viel einfacher, die Benennung durchzuführen. Anstatt Millionen von Bildern per Hand zu labeln, kann der Mensch jeweils die unterschiedlichen Gruppen manuell labeln, und die zugeordneten Labels werden allen Mitgliedern innerhalb jeder Gruppe zugewiesen.

Wenn die unüberwacht lernende KI nach dem anfänglichen Training Bilder findet, die zu keiner der gelabelten Gruppen gehören, erzeugt sie separate Gruppen für die nicht klassifizierten Bilder und veranlasst, dass der Mensch die neuen, noch zu benennenden Bilder mit Labels versieht.

Unsupervised Learning macht bis dahin unlösbare Probleme lösbar und ist wesentlich flinker beim Aufspüren von verborgenen Mustern – sowohl in den historischen Daten, die für das Training zur Verfügung stehen, als auch in zukünftig erfassten Daten. Darüber hinaus haben wir jetzt einen KI-Ansatz für die riesigen Mengen an nicht gelabelten Daten, die es weltweit gibt.

Selbst wenn Unsupervised Learning beim Lösen spezifischer, eng definierter Probleme weniger geschickt ist als Supervised Learning, schneidet es besser ab, wenn es darum geht, ergebnisoffene Probleme in der Art der starken KI anzugehen und dieses Wissen zu verallgemeinern.

Genauso wichtig ist es, dass Unsupervised Learning viele der allgemeinen Probleme angehen kann, mit denen Data Scientists zu tun haben, wenn Lösungen mit maschinellem Lernen zu erstellen sind.

Lösungen mit maschinellem Lernen durch Unsupervised Learning verbessern

Die jüngsten Erfolge im maschinellen Lernen sind der Verfügbarkeit sehr umfangreicher Daten, den Fortschritten in der Computerhardware und den cloudbasierten Ressourcen sowie den Durchbrüchen bei den Algorithmen zum maschinellen Lernen zu verdanken. Diese Erfolge konzentrieren sich aber vorwiegend auf enge KI-Probleme wie zum Beispiel Bildklassifizierung, Computervision, Spracherkennung, Natural Language Processing und Maschinenübersetzung.

Um anspruchsvollere KI-Probleme zu lösen, müssen wir das Potenzial des Unsupervised Learning freisetzen. Untersuchen wir einmal die am häufigsten vorkommenden Herausforderungen, denen sich Data Scientists gegenübersehen, wenn sie Lösungen erstellen, und wie Unsupervised Learning ihnen dabei helfen kann.

Unzureichend gelabelte Daten

> *Meiner Ansicht nach ist KI wie der Bau einer Rakete. Man braucht einen riesigen Motor und jede Menge Treibstoff. Mit einem kleinen Motor und einer winzigen Menge Treibstoff schafft man es nicht bis in den Orbit. Mit einem winzigen Motor und einer Tonne Treibstoff kann man nicht einmal abheben. Um eine Rakete zu bauen, braucht man einen riesigen Motor und sehr viel Treibstoff.*
>
> – Andrew Ng

Wäre maschinelles Lernen ein Raumschiff, würden die Daten den Treibstoff liefern – ohne Unmengen von Daten kann das Raumschiff nicht fliegen. Doch nicht alle Daten werden gleich erzeugt. Um überwachte Algorithmen zu verwenden, brauchen wir sehr viele gelabelte Daten, die aber nur schwer und kostenintensiv zu generieren sind.[1]

Beim Unsupervised Learning können wir ungelabelte Beispiele automatisch kennzeichnen. Das funktioniert folgendermaßen: Wir würden alle Beispiele clustern und dann die Labels von gelabelten Beispielen auf die ungelabelten innerhalb desselben Clusters anwenden. Ungelabelte Beispiele würden das Label der gelabelten erhalten, denen sie am ähnlichsten sind. Mit Clustering beschäftigt sich Kapitel 5.

1 Es gibt Start-ups wie zum Beispiel Figure Eight, die ausdrücklich diesen *Human-in-the-Loop*-Dienst anbieten.