O'REILLY®

2. Auflage
Aktuell zu
TensorFlow 2

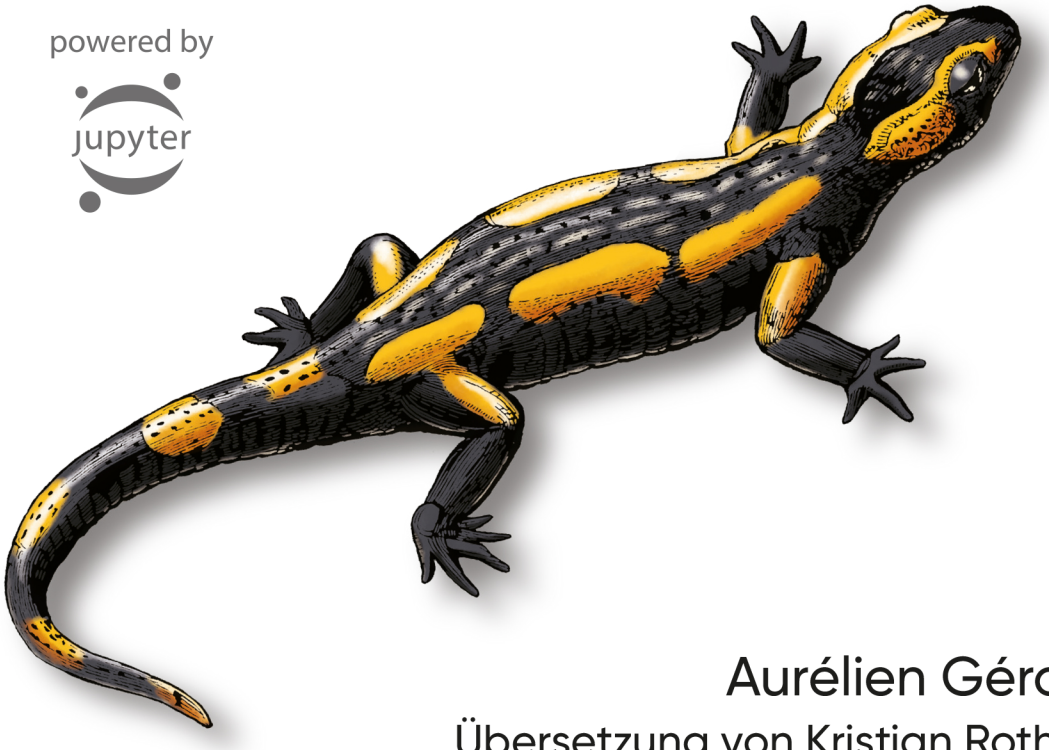
Praxiseinstieg Machine Learning

mit Scikit-Learn, Keras und TensorFlow

Konzepte, Tools und Techniken
für intelligente Systeme

powered by

jupyter



Aurélien Géron

Übersetzung von Kristian Rother
und Thomas Demmig

2. AUFLAGE

Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow

Konzepte, Tools und Techniken für intelligente Systeme

Aurélien Géron

Deutsche Übersetzung von Kristian Rother & Thomas Demmig

O'REILLY®

Aurélien Géron

Lektorat: Alexandra Follenius
Übersetzung: Kristian Rother, Thomas Demmig
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 978-3-96009-124-0
PDF 978-3-96010-339-4
ePub 978-3-96010-340-0
mobi 978-3-96010-341-7

2. Auflage

Wieblinger Weg 17
69123 Heidelberg

Authorized German translation of the English edition of *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition*, ISBN 9781492032649 © 2019 Kiwisoft S.A.S. This translation is published and sold by permission of O'Reilly Media, Inc., which owns or controls all rights to publish and sell the same.

Dieses Buch erscheint in Kooperation mit O'Reilly Media, Inc. unter dem Imprint »O'REILLY«.
O'REILLY ist ein Markenzeichen und eine eingetragene Marke von O'Reilly Media, Inc. und wird mit Einwilligung des Eigentümers verwendet.

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

Vorwort

Der Machine-Learning-Tsunami

Im Jahr 2006 erschien ein Artikel (*https://homl.info/136*) von Geoffrey Hinton et al.,[1] in dem vorgestellt wurde, wie sich ein neuronales Netz zum Erkennen handgeschriebener Ziffern mit ausgezeichneter Genauigkeit (> 98 %) trainieren lässt. Ein Deep Neural Network ist ein (sehr) vereinfachtes Modell unseres zerebralen Kortex, und es besteht aus einer Folge von Schichten mit künstlichen Neuronen. Die Autoren nannten diese Technik »Deep Learning«. Zu dieser Zeit wurde das Trainieren eines Deep-Learning-Netzes im Allgemeinen als unmöglich angesehen,[2] und die meisten Forscher hatten die Idee in den 1990ern aufgegeben. Dieser Artikel ließ das Interesse der wissenschaftlichen Gemeinde wieder aufleben, und schon nach kurzer Zeit zeigten weitere Artikel, dass Deep Learning nicht nur möglich war, sondern umwerfende Dinge vollbringen konnte, zu denen kein anderes Machine-Learning-(ML-)Verfahren auch nur annähernd in der Lage war (mithilfe enormer Rechenleistung und riesiger Datenmengen). Dieser Enthusiasmus breitete sich schnell auf weitere Teilgebiete des Machine Learning aus.

Zehn Jahre später hat Machine Learning ganze Industriezweige erobert: Es ist zu einem Herzstück heutiger Spitzentechnologien geworden und dient dem Ranking von Suchergebnissen im Web, kümmert sich um die Spracherkennung Ihres Smartphones, gibt Empfehlungen für Videos und schlägt den Weltmeister im Brettspiel Go. Über kurz oder lang wird ML vermutlich auch Ihr Auto steuern.

1 Geoffrey Hinton et al., »A Fast Learning Algorithm for Deep Belief Nets«, *Neural Computation* 18 (2006): 1527–1554.

2 Obwohl die Konvolutionsnetze von Yann Lecun bei der Bilderkennung seit den 1990ern gut funktioniert hatten, auch wenn sie nicht allgemein anwendbar waren.

Machine Learning in Ihren Projekten

Deshalb interessieren Sie sich natürlich auch für Machine Learning und möchten an der Party teilnehmen!

Womöglich möchten Sie Ihrem selbst gebauten Roboter einen eigenen Denkapparat geben? Ihn Gesichter erkennen lassen? Oder lernen lassen, herumzulaufen?

Oder vielleicht besitzt Ihr Unternehmen Unmengen an Daten (Logdateien, Finanzdaten, Produktionsdaten, Sensordaten, Hotline-Statistiken, Personalstatistiken und so weiter), und Sie könnten vermutlich einige verborgene Schätze heben, wenn Sie nur wüssten, wo Sie danach suchen müssten, beispielsweise:

- Kundensegmente finden und für jede Gruppe die beste Marketingstrategie entwickeln.
- Jedem Kunden anhand des Kaufverhaltens ähnlicher Kunden Produktempfehlungen geben.
- Betrügerische Transaktionen mit hoher Wahrscheinlichkeit erkennen.
- Den Unternehmensgewinn im nächsten Jahr vorhersagen.

Was immer der Grund ist, Sie haben beschlossen, Machine Learning zu erlernen und in Ihren Projekten umzusetzen. Eine ausgezeichnete Idee!

Ziel und Ansatz

Dieses Buch geht davon aus, dass Sie noch so gut wie nichts über Machine Learning wissen. Unser Ziel ist es, Ihnen die Grundbegriffe, ein Grundverständnis und die Werkzeuge an die Hand zu geben, mit denen Sie Programme zum *Lernen aus Daten* entwickeln können.

Wir werden eine Vielzahl von Techniken besprechen, von den einfachsten und am häufigsten eingesetzten (wie der linearen Regression) bis zu einigen Deep-Learning-Verfahren, die regelmäßig Wettbewerbe gewinnen.

Anstatt eigene Übungsversionen jedes Algorithmus zu entwickeln, werden wir dazu für den Produktionsbetrieb geschaffene Python-Frameworks verwenden:

- Scikit-Learn (*http://scikit-learn.org/*) ist sehr einfach zu verwenden, enthält aber effiziente Implementierungen vieler Machine-Learning-Algorithmen. Damit ist es ein großartiger Ausgangspunkt, um Machine Learning zu erlernen.
- TensorFlow (*http://tensorflow.org/*) ist eine komplexere Bibliothek für verteiltes Rechnen. Mit ihr können Sie sehr große neuronale Netze effizient trainieren und ausführen, indem Sie die Berechnungen auf bis zu Hunderte von Servern mit mehreren GPUs (*Graphics Processing Units*) verlagern. TensorFlow (TF) wurde von Google entwickelt und läuft in vielen großflächigen Machine-Learning-Anwendungen. Die Bibliothek wurde im November 2015 als Open Source veröffentlicht.

- Keras (*https://keras.io/*) ist eine High-Level-Deep-Learning-API, die das Trainieren und Ausführen neuronaler Netze sehr einfach macht. Sie kann auf TensorFlow, Theano oder Microsoft Cognitive Toolkit (früher bekannt als CNTK) aufsetzen. TensorFlow bringt seine eigene Implementierung dieser API namens *tf.keras* mit, die einige der fortgeschritteneren TensorFlow-Features unterstützt (zum Beispiel die Möglichkeit, Daten effizient zu laden).

Dieses Buch verfolgt einen praxisorientierten Ansatz, bei dem Sie ein intuitives Verständnis von Machine Learning entwickeln, indem Sie sich mit konkreten Beispielen und ein klein wenig Theorie beschäftigen. Auch wenn Sie dieses Buch lesen können, ohne Ihren Laptop in die Hand zu nehmen, empfehlen wir Ihnen, mit den als Jupyter-Notebooks unter *https://github.com/ageron/handson-ml2* verfügbaren Codebeispielen herumzuexperimentieren.

Voraussetzungen

Dieses Buch geht davon aus, dass Sie ein wenig Programmiererfahrung mit Python haben und dass Sie mit den wichtigsten wissenschaftlichen Bibliotheken in Python vertraut sind, insbesondere mit NumPy (*http://numpy.org/*), pandas (*http://pandas.pydata.org/*) und Matplotlib (*http://matplotlib.org/*).

Wenn Sie sich dafür interessieren, was hinter den Kulissen passiert, sollten Sie ein Grundverständnis von Oberstufenmathematik haben (Analysis, lineare Algebra, Wahrscheinlichkeiten und Statistik).

Sollten Sie Python noch nicht kennen, ist *http://learnpython.org/* ein ausgezeichneter Ausgangspunkt. Das offizielle Tutorial auf python.org (*https://docs.python.org/3/tutorial/*) ist ebenfalls recht gut.

Falls Sie Jupyter noch nie verwendet haben, führt Sie Kapitel 2 durch die Installation und die Grundlagen: Es ist ein leistungsfähiges Werkzeug in Ihrem Werkzeugkasten.

Und für den Fall, dass Sie mit den wissenschaftlichen Bibliotheken für Python nicht vertraut sind, beinhalten die mitgelieferten Jupyter-Notebooks einige Tutorials. Es gibt dort auch ein kurzes Mathematiktutorial über lineare Algebra.

Wegweiser durch dieses Buch

Dieses Buch ist in zwei Teile aufgeteilt. Teil I behandelt folgende Themen:

- Was ist Machine Learning? Welche Aufgaben lassen sich damit lösen? Welches sind die wichtigsten Kategorien und Grundbegriffe von Machine-Learning-Systemen?
- Die Schritte in einem typischen Machine-Learning-Projekt.
- Lernen durch Anpassen eines Modells an Daten.

- Optimieren einer Kostenfunktion.
- Bearbeiten, Säubern und Vorbereiten von Daten.
- Merkmale auswählen und entwickeln.
- Ein Modell auswählen und dessen Hyperparameter über Kreuzvalidierung optimieren.
- Die Herausforderungen beim Machine Learning, insbesondere Underfitting und Overfitting (das Gleichgewicht zwischen Bias und Varianz).
- Die verbreitetsten Lernalgorithmen: lineare und polynomielle Regression, logistische Regression, k-nächste Nachbarn, Support Vector Machines, Entscheidungsbäume, Random Forests und Ensemble-Methoden.
- Dimensionsreduktion der Trainingsdaten, um dem »Fluch der Dimensionalität« etwas entgegenzusetzen.
- Andere Techniken des unüberwachten Lernens, unter anderem Clustering, Dichteabschätzung und Anomalieerkennung.

Teil II widmet sich diesen Themen:

- Was sind neuronale Netze? Wofür sind sie geeignet?
- Erstellen und Trainieren neuronaler Netze mit TensorFlow und Keras.
- Die wichtigsten Architekturen neuronaler Netze: Feed-Forward-Netze für Tabellendaten, Convolutional Neural Networks zur Bilderkennung, rekurrente Netze und Long-Short-Term-Memory-(LSTM-)Netze zur Sequenzverarbeitung, Encoder/Decoder und Transformer für die Sprachverarbeitung, Autoencoder und Generative Adversarial Networks (GANs) zum generativen Lernen.
- Techniken zum Trainieren von Deep-Learning-Netzen.
- Wie man einen Agenten erstellt (zum Beispiel einen Bot in einem Spiel), der durch Versuch und Irrtum gute Strategien erlernt und dabei Reinforcement Learning einsetzt.
- Effizientes Laden und Vorverarbeiten großer Datenmengen.
- Trainieren und Deployen von TensorFlow-Modellen im großen Maßstab.

Der erste Teil baut vor allem auf Scikit-Learn auf, der zweite Teil verwendet TensorFlow.

Springen Sie nicht zu schnell ins tiefe Wasser: Auch wenn Deep Learning zweifelsohne eines der aufregendsten Teilgebiete des Machine Learning ist, sollten Sie zuerst Erfahrungen mit den Grundlagen sammeln. Außerdem lassen sich die meisten Aufgabenstellungen recht gut mit einfacheren Techniken wie Random Forests und Ensemble-Methoden lösen (die in Teil I besprochen werden). Deep Learning ist am besten für komplexe Aufgaben wie Bilderkennung, Spracherkennung und Sprachverarbeitung geeignet, vorausgesetzt, Sie haben genug Daten und Geduld.

Änderungen in der zweiten Auflage

Diese zweite Auflage hat sechs zentrale Ziele:

1. Die Behandlung zusätzlicher ML-Themen: weitere Techniken zum unüberwachten Lernen (unter anderem Clustering, Anomalieerkennung, Dichteabschätzung und Mischmodelle), zusätzliche Techniken zum Trainieren von Deep Networks (einschließlich sich selbst normalisierender Netze), weitere Techniken der Bilderkennung (unter anderem Xception, SENet, Objekterkennung mit YOLO und semantische Segmentierung mit R-CNN), das Verarbeiten von Sequenzen mit Convolutional Neural Networks (CNNs, einschließlich WaveNet), Verarbeitung natürlicher Sprache mit rekurrenten neuronalen Netzen (RNNs), CNNs und Transformers, GANs.
2. Zusätzliche Bibliotheken und APIs (Keras, die Data-API, TF-Agents für das Reinforcement Learning) sowie das Trainieren und Deployen von TF-Modellen im großen Maßstab mithilfe der Distribution Strategies API, TF Serving und Google Cloud AI Platform; zudem eine kurze Vorstellung von TF Transform, TFLite, TF Addons/Seq2Seq und TensorFlow.js.
3. Vorstellen einiger der neuesten Forschungsergebnisse aus dem Deep Learning.
4. Migrieren aller TensorFlow-Kapitel nach TensorFlow 2 und Verwenden der TensorFlow-Implementierung der Keras-API (tf.keras), wann immer das möglich ist.
5. Aktualisieren der Codebeispiele auf die neuesten Versionen von Scikit-Learn, NumPy, pandas, Matplotlib und anderer Bibliotheken.
6. Anpassen einiger Abschnitte zum besseren Verständnis und Beheben von Fehlern dank sehr vieler Rückmeldungen von Lesern.

Es wurden ein paar Kapitel hinzugefügt, andere wurden umgeschrieben, und ein paar wurden neu angeordnet. Unter *https://homl.info/changes2* finden Sie detailliertere Angaben darüber, was sich in der zweiten Auflage geändert hat.

Ressourcen im Netz

Es gibt viele ausgezeichnete Ressourcen, mit deren Hilfe sich Machine Learning erlernen lässt. Der ML-Kurs auf Coursera (*https://homl.info/ngcourse*) von Andrew Ng ist faszinierend, auch wenn er einen beträchtlichen Zeitaufwand bedeutet (in Monaten).

Darüber hinaus finden Sie viele interessante Webseiten über Machine Learning, darunter natürlich den ausgezeichneten User Guide (*https://homl.info/skdoc*) von Scikit-Learn. Auch Dataquest (*https://www.dataquest.io/*), das sehr ansprechende Tutorials und ML-Blogs bietet, sowie die auf Quora (*https://homl.info/1*) aufgeführ-

ten ML-Blogs könnten Ihnen gefallen. Schließlich sind auf der Deep-Learning-Website (*http://deeplearning.net/*) Ressourcen aufgezählt, mit denen Sie mehr lernen können.

Natürlich bieten auch viele andere Bücher eine Einführung in Machine Learning, insbesondere:

- Joel Grus, *Einführung in Data Science: Grundprinzipien der Datenanalyse mit Python (https://www.oreilly.de/buecher/13335/9783960091233-einf%C3%BChrung-in-data-science.html)* (O'Reilly). Dieses Buch stellt die Grundlagen von Machine Learning vor und implementiert die wichtigsten Algorithmen in reinem Python (von null auf).
- Stephen Marsland, *Machine Learning: An Algorithmic Perspective* (Chapman & Hall). Dieses Buch ist eine großartige Einführung in Machine Learning, die viele Themen ausführlich behandelt. Es enthält Codebeispiele in Python (ebenfalls von null auf, aber mit NumPy).
- Sebastian Raschka, *Machine Learning mit Python und Scikit-Learn und TensorFlow: Das umfassende Praxis-Handbuch für Data Science, Predictive Analytics und Deep Learning* (mitp Professional). Eine weitere ausgezeichnete Einführung in Machine Learning. Dieses Buch konzentriert sich auf Open-Source-Bibliotheken in Python (Pylearn 2 und Theano).
- François Chollet, *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek* (mitp Professional). Ein sehr praxisnahes Buch, das klar und präzise viele Themen behandelt – wie Sie es vom Autor der ausgezeichneten Keras-Bibliothek erwarten können. Es zieht Codebeispiele der mathematischen Theorie vor.
- Andriy Burkov, *Machine Learning kompakt: Alles, was Sie wissen müssen* (mitp Professional). Dieses sehr kurze Buch behandelt ein beeindruckendes Themenspektrum gut verständlich, scheut dabei aber nicht vor mathematischen Gleichungen zurück.
- Yaser S. Abu-Mostafa, Malik Magdon-Ismail und Hsuan-Tien Lin, *Learning from Data* (AMLBook). Als eher theoretische Abhandlung von ML enthält dieses Buch sehr tiefgehende Erkenntnisse, insbesondere zum Gleichgewicht zwischen Bias und Varianz (siehe Kapitel 4).
- Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach, 3rd Edition* (Pearson). Dieses ausgezeichnete (und umfangreiche) Buch deckt eine unglaubliche Stoffmenge ab, darunter Machine Learning. Es hilft dabei, ML in einem breiteren Kontext zu betrachten.

Eine gute Möglichkeit zum Lernen sind schließlich Webseiten mit ML-Wettbewerben wie Kaggle.com (*https://www.kaggle.com/*). Dort können Sie Ihre Fähigkeiten an echten Aufgaben üben und Hilfe und Tipps von einigen der besten ML-Profis erhalten.

In diesem Buch verwendete Konventionen

Die folgenden typografischen Konventionen werden in diesem Buch verwendet:

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

`Konstante Zeichenbreite`
: Wird für Programmlistings und für Programmelemente in Textabschnitten wie Namen von Variablen und Funktionen, Datenbanken, Datentypen, Umgebungsvariablen, Anweisungen und Schlüsselwörter verwendet.

`Konstante Zeichenbreite, fett`
: Kennzeichnet Befehle oder anderen Text, den der Nutzer wörtlich eingeben sollte.

`Konstante Zeichenbreite, kursiv`
: Kennzeichnet Text, den der Nutzer je nach Kontext durch entsprechende Werte ersetzen sollte.

Dieses Symbol steht für einen Tipp oder eine Empfehlung.

Dieses Symbol steht für einen allgemeinen Hinweis.

Dieses Symbol steht für eine Warnung oder erhöhte Aufmerksamkeit.

Codebeispiele

Es gibt eine Reihe von Jupyter-Notebooks voll mit zusätzlichem Material, wie Codebeispielen und Übungen, die zum Herunterladen unter *https://github.com/ageron/handson-ml2* bereitstehen.

Einige der Codebeispiele im Buch lassen sich wiederholende Abschnitte oder Details weg, die offensichtlich sind oder nichts mit Machine Learning zu tun haben. Das sorgt dafür, dass Sie sich auf die wichtigen Teile des Codes konzentrieren können, und spart Platz, um mehr Themen unterzubringen. Wollen Sie sich die vollständigen Codebeispiele betrachten, finden Sie diese in den Jupyter-Notebooks.

Geben die Codebeispiele etwas aus, wird dies mit Python-Prompts (`>>>` und `...`) wie in einer Python-Shell dargestellt, um den Code deutlich von den Ausgaben trennen zu können. So definiert beispielsweise folgender Code die Funktion `square()`, rechnet dann damit und gibt das Quadrat von 3 aus:

```
>>> def square(x):
...     return x ** 2
...
>>> result = square(3)
>>> result
9
```

Gibt Code nichts aus, werden keine Prompts verwendet. Aber das Ergebnis wird manchmal als Kommentar angegeben, wie zum Beispiel hier:

```
def square(x):
    return x ** 2

result = square(3)  # Ergebnis ist 9
```

Verwenden von Codebeispielen

Dieses Buch ist dazu da, Ihnen beim Erledigen Ihrer Arbeit zu helfen. Im Allgemeinen dürfen Sie die Codebeispiele aus diesem Buch in Ihren eigenen Programmen und der dazugehörigen Dokumentation verwenden. Sie müssen uns dazu nicht um Erlaubnis fragen, solange Sie nicht einen beträchtlichen Teil des Codes reproduzieren. Beispielsweise benötigen Sie keine Erlaubnis, um ein Programm zu schreiben, in dem mehrere Codefragmente aus diesem Buch vorkommen. Wollen Sie dagegen eine CD-ROM mit Beispielen aus Büchern von O'Reilly verkaufen oder verteilen, benötigen Sie eine Erlaubnis. Eine Frage zu beantworten, indem Sie aus diesem Buch zitieren und ein Codebeispiel wiedergeben, benötigt keine Erlaubnis. Eine beträchtliche Menge Beispielcode aus diesem Buch in die Dokumentation Ihres Produkts aufzunehmen, bedarf hingegen einer Erlaubnis.

Wir freuen uns über Zitate, verlangen diese aber nicht. Ein Zitat enthält Titel, Autor, Verlag und ISBN. Beispiel: »*Praxiseinstieg Machine Learning mit Scikit-Learn, Keras und TensorFlow* von Aurélien Géron (O'Reilly). Copyright 2020, ISBN 978-3-96009-124-0.«

Wenn Sie glauben, dass Ihre Verwendung von Codebeispielen über die übliche Nutzung hinausgeht oder außerhalb der oben vorgestellten Nutzungsbedingungen liegt, kontaktieren Sie uns bitte unter *kommentar@oreilly.de*.

Danksagungen

In meinen wildesten Träumen hätte ich mir niemals vorgestellt, dass die erste Auflage dieses Buchs solch eine Verbreitung finden würde. Ich habe so viele Nachrich-

ten von Lesern erhalten – oft mit Fragen, manche mit freundlichen Hinweisen auf Fehler und die meisten mit ermutigenden Worten. Ich kann gar nicht sagen, wie dankbar ich all diesen Lesern für ihre Unterstützung bin. Vielen, vielen Dank! Scheuen Sie sich nicht, sich auf GitHub (*https://homl.info/issues2*) zu melden, wenn Sie Fehler in den Codebeispielen finden (oder einfach etwas fragen wollen) oder um auf Fehler im Text aufmerksam (*https://homl.info/errata2*) zu machen. Manche Leser haben mir auch geschrieben, wie dieses Buch ihnen dabei geholfen hat, ihren ersten Job zu bekommen oder ein konkretes Problem zu lösen, an dem sie gearbeitet haben. Ich finde ein solches Feedback unglaublich motivierend. Hat Ihnen dieses Buch geholfen, würde ich mich freuen, wenn Sie mir Ihre Geschichte erzählen würden – entweder privat (zum Beispiel über LinkedIn (*https://www.linkedin.com/in/aurelien-geron/*)) oder öffentlich (beispielsweise in einem Tweet oder über ein Amazon-Review (*https://homl.info/amazon2*)).

Ich bin all den fantastischen Menschen unglaublich dankbar, die in ihrem geschäftigen Leben die Zeit gefunden haben, mein Buch im Detail gegenzulesen. Insbesondere möchte ich François Chollet dafür danken, dass er alle Kapitel zu Keras und TensorFlow kontrolliert und mir großartiges und detailliertes Feedback gegeben hat. Da Keras eine meiner wichtigsten Ergänzungen dieser zweiten Auflage ist, war die Review durch den Autor unbezahlbar. Ich empfehle Ihnen François' Buch *Deep Learning mit Python und Keras: Das Praxis-Handbuch vom Entwickler der Keras-Bibliothek* (mitp Professional): Es bietet die Präzision, Klarheit und Tiefe, die auch die Keras-Bibliothek selbst besitzt. Besonderer Dank geht ebenfalls an Ankur Patel, der jedes Kapitel dieser zweiten Auflage begutachtet und mir ausgezeichnetes Feedback gegeben hat, insbesondere zu Kapitel 9, das sich um unüberwachtes Lernen dreht. Er könnte glatt ein ganzes Buch zu dem Thema schreiben ... Moment mal, das hat er ja! Schauen Sie sich mal *Praxisbuch Unsupervised Learning: Machine-Learning-Anwendungen für ungelabelte Daten mit Python programmieren* (*https://www.oreilly.de/buecher/13534/9783960091271-praxisbuch-unsupervised-learning.html*) (O'Reilly) an. Ein großes Dankeschön auch an Olzhas Akpambetov, der alle Kapitel im zweiten Teil des Buchs begutachtet, sehr viel Code getestet und viele großartige Verbesserungsvorschläge gemacht hat. Ich bin dankbar, dass Mark Daoust, Jon Krohn, Dominic Monn und Josh Patterson den zweiten Teil des Buchs so genau begutachtet und ihre Erfahrungen eingebracht haben. Sie ließen keinen Stein auf dem anderen und lieferten ausgezeichnete Hinweise.

Beim Schreiben dieser zweiten Auflage hatte ich das Glück, sehr viel Hilfe von Mitgliedern des TensorFlow-Teams zu bekommen, insbesondere von Martin Wicke, der unermüdlich Dutzende meiner Fragen beantwortet und den Rest an die richtigen Leute weitergeleitet hat, unter anderen an Karmel Allison, Paige Bailey, Eugene Brevdo, William Chargin, Daniel »Wolff« Dobson, Nick Felt, Bruce Fontaine, Goldie Gadde, Sandeep Gupta, Priya Gupta, Kevin Haas, Konstantinos Katsiapis, Viacheslav Kovalevskyi, Allen Lavoie, Clemens Mewald, Dan Moldovan, Sean Morgan, Tom O'Malley, Alexandre Passos, André Susano Pinto, Anthony Platanios, Oscar Ramirez, Anna Revinskaya, Saurabh Saxena, Ryan Sepassi, Jiri Simsa,

Xiaodan Song, Christina Sorokin, Dustin Tran, Todd Wang, Pete Warden (der auch die erste Auflage begutachtet hat), Edd Wilder-James und Yuefeng Zhou, die alle außerordentlich hilfreich waren. Ein großer Dank an euch alle und auch an alle anderen Mitglieder des TensorFlow-Teams – nicht nur für eure Hilfe, sondern auch dafür, dass ihr solch eine tolle Bibliothek geschaffen habt. Ein besonderer Dank geht an Irene Giannoumis und Robert Crowe vom TFCX-Team, die die Kapitel 13 und 19 im Detail durchgearbeitet haben.

Ich danke auch den fantastischen Menschen bei O'Reilly, insbesondere Nicole Taché für ihr aufschlussreiches, immer freundliches, ermutigendes und hilfreiches Feedback. Eine bessere Lektorin hätte ich mir nicht vorstellen können. Ein großer Dank geht an Michele Cronin, die zu Beginn dieser zweiten Auflage sehr hilfreich (und geduldig) war, und an Kristen Brown, Production Editor für die zweite Auflage, die sie auf allen Schritten begleitet hat (sie hat auch Korrekturen und Aktualisierungen jedes Nachdrucks der ersten Auflage koordiniert). Ich danke Rachel Monaghan und Amanda Kersey für ihr umfassendes Copyediting (der ersten bzw. zweiten Auflage) und Johnny O'Toole, der die Beziehung zu Amazon gemanagt und viele meiner Fragen beantwortet hat. Dank geht an Marie Beaugureau, Ben Lorica, Mike Loukides und Laurel Ruma dafür, dass sie an dieses Projekt geglaubt und mir geholfen haben, den Rahmen abzustecken. Ich danke Matt Hacker und dem gesamten Atlas-Team für das Beantworten aller meiner technischen Fragen zu Formatierung, AsciiDoc und LaTeX sowie Nick Adams, Rebecca Demarest, Rachel Head, Judith McConville, Helen Monroe, Karen Montgomery, Rachel Roumeliotis und allen bei O'Reilly, die zu diesem Buch beigetragen haben.

Ich möchte auch meinen früheren Kollegen bei Google danken, insbesondere dem Team zur Klassifikation von YouTube-Videos, von denen ich sehr viel über Machine Learning gelernt habe. Ohne sie hätte ich die erste Auflage niemals starten können. Besonderer Dank gebührt meinen persönlichen ML-Gurus: Clément Courbet, Julien Dubois, Mathias Kende, Daniel Kitachewsky, James Pack, Alexander Pak, Anosh Raj, Vitor Sessak, Wiktor Tomczak, Ingrid von Glehn und Rich Washington. Und danke an alle anderen, mit denen ich bei YouTube und in den großartigen Google-Forschungsteams in Mountain View zusammengearbeitet habe. Vielen Dank auch an Martin Andrews, Sam Witteveen und Jason Zaman, dass sie mich in ihre Google-Developer-Experts-Gruppe in Singapur aufgenommen haben – mit freundlicher Unterstützung durch Soonson Kwon – und für all die tollen Diskussionen über Deep Learning und TensorFlow. Jeder, der in Singapur an Deep Learning interessiert ist, sollte auf jeden Fall das Deep Learning Singapore Meetup (*https://homl.info/meetupsg*) besuchen. Jason verdient besonderen Dank für sein TFLite-Know-how, das in Kapitel 19 eingeflossen ist!

Nie werde ich all die netten Leute vergessen, die die erste Auflage dieses Buchs Korrektur gelesen haben, unter anderem David Andrzejewski, Lukas Biewald, Justin Francis, Vincent Guilbeau, Eddy Hung, Karim Matrah, Grégoire Mesnil, Salim Sémaoune, Iain Smears, Michel Tessier, Ingrid von Glehn, Pete Warden und natür-

lich mein lieber Bruder Sylvain. Ein besonderer Dank geht an Haesun Park, der mir sehr viel ausgezeichnetes Feedback gab und viele Fehler fand, während er die koreanische Übersetzung der ersten Auflage schrieb. Er hat auch die Jupyter-Notebooks ins Koreanische übersetzt, nicht zu vergessen die Dokumentation von TensorFlow. Ich spreche kein Koreanisch, aber in Anbetracht der Qualität seines Feedbacks müssen alle seine Übersetzungen wirklich ausgezeichnet sein. Darüber hinaus hat Haesun freundlicherweise ein paar der Lösungen zu den Übungen in dieser zweiten Auflage beigetragen.

Schließlich bin ich meiner geliebten Frau Emmanuelle und unseren drei wunderbaren Kindern Alexandre, Rémi und Gabrielle unendlich dafür dankbar, dass sie mich zur Arbeit an diesem Buch ermutigt haben. Auch danke ich ihnen für ihre unersättliche Neugier: Indem ich einige der schwierigsten Konzepte in diesem Buch meiner Frau und meinen Kindern erklärt habe, konnte ich meine Gedanken ordnen und viele Teile direkt verbessern. Und sie haben mir sogar Kekse und Kaffee vorbeigebracht. Was kann man sich noch mehr wünschen?

TEIL I

Die Grundlagen des Machine Learning

KAPITEL 1

Die Machine-Learning-Umgebung

Die meisten Menschen denken beim Begriff »Machine Learning« an einen Roboter: einen zuverlässigen Butler oder einen tödlichen Terminator, je nachdem, wen Sie fragen. Aber Machine Learning ist keine futuristische Fantasie, es ist bereits Gegenwart. Tatsächlich gibt es Machine Learning in bestimmten, spezialisierten Anwendungsbereichen wie der optischen Zeichenerkennung (OCR) schon seit Jahrzehnten. Aber die erste weitverbreitete Anwendung von ML, die das Leben von Hunderten Millionen Menschen verbesserte, hat die Welt in den 1990er-Jahren erobert: Es war der *Spamfilter*. Es ist nicht gerade ein Skynet mit eigenem Bewusstsein, aber technisch gesehen ist es Machine Learning (es hat inzwischen so gut gelernt, dass Sie nur noch selten eine E-Mail als Spam kennzeichnen müssen). Dem Spamfilter folgten etliche weitere Anwendungen von ML, die still und heimlich Hunderte Produkte und Funktionen aus dem Alltag steuern, darunter Einkaufsempfehlungen und Stimmsuche.

Wo aber beginnt Machine Learning, und wo hört es auf? Worum genau geht es, wenn eine Maschine etwas *lernt*? Wenn ich mir eine Kopie von Wikipedia herunterlade, hat mein Computer dann schon etwas gelernt? Ist er auf einmal schlauer geworden? In diesem Kapitel werden wir erst einmal klarstellen, was Machine Learning ist und wofür Sie es einsetzen könnten.

Bevor wir aber beginnen, den Kontinent des Machine Learning zu erforschen, werfen wir einen Blick auf die Landkarte und lernen die wichtigsten Regionen und Orientierungspunkte kennen: überwachtes und unüberwachtes Lernen, Online- und Batch-Learning, instanzbasiertes und modellbasiertes Lernen. Anschließend betrachten wir die Arbeitsabläufe in einem typischen ML-Projekt, diskutieren die dabei wichtigsten Herausforderungen und besprechen, wie Sie ein Machine-Learning-System auswerten und optimieren können.

In diesem Kapitel werden diverse Grundbegriffe (und Fachjargon) eingeführt, die jeder Data Scientist auswendig kennen sollte. Es wird ein abstrakter und recht einfacher Überblick bleiben (das einzige Kapitel mit wenig Code), aber Ihnen sollte alles glasklar sein, bevor Sie mit dem Buch fortfahren. Schnappen Sie sich also einen Kaffee, und los geht's!

Wenn Sie bereits sämtliche Grundlagen von Machine Learning kennen, können Sie direkt mit Kapitel 2 fortfahren. Falls Sie sich nicht sicher sind, versuchen Sie, die Fragen am Ende des Kapitels zu beantworten, bevor Sie fortfahren.

Was ist Machine Learning?

Machine Learning ist die Wissenschaft (und Kunst), Computer so zu programmieren, dass sie *anhand von Daten lernen*.

Hier ist eine etwas allgemeinere Definition:

> [Maschinelles Lernen ist das] Fachgebiet, das Computern die Fähigkeit zu lernen verleiht, ohne explizit programmiert zu werden.
>
> *– Arthur Samuel 1959*

Und eine eher technisch orientierte:

> Man sagt, dass ein Computerprogramm dann aus Erfahrungen *E* in Bezug auf eine Aufgabe *T* und ein Maß für die Leistung *P* lernt, wenn seine durch *P* gemessene Leistung bei *T* mit der Erfahrung *E* anwächst.
>
> *– Tom Mitchell 1997*

Ihr Spamfilter ist ein maschinelles Lernprogramm, das aus Beispielen für Spam-E-Mails (z. B. vom Nutzer markierten) und gewöhnlichen E-Mails (Nicht-Spam, auch »Ham« genannt) lernt, Spam zu erkennen. Diese vom System verwendeten Lernbeispiele nennt man den *Trainingsdatensatz*. Jedes Trainingsbeispiel nennt man einen *Trainingsdatenpunkt* (oder *Instanz*). In diesem Fall besteht die Aufgabe *T* darin, neue E-Mails als Spam zu kennzeichnen, die Erfahrung *E* entspricht den *Trainingsdaten*. Nur das Leistungsmaß *P* ist noch zu definieren; Sie könnten z. B. den Anteil korrekt klassifizierter E-Mails verwenden. Dieses Leistungsmaß nennt man *Genauigkeit*. Es wird bei Klassifikationsaufgaben häufig verwendet.

Falls Sie gerade eine Kopie von Wikipedia heruntergeladen haben, verfügt Ihr Computer über eine Menge zusätzlicher Daten, verbessert sich dadurch aber bei keiner Aufgabe. Deshalb ist dies kein Machine Learning.

Warum wird Machine Learning verwendet?

Überlegen Sie einmal, wie Sie mit herkömmlichen Programmiertechniken einen Spamfilter schreiben würden (siehe Abbildung 1-1):

1. Zuerst würden Sie sich ansehen, wie Spam typischerweise aussieht. Sie würden feststellen, dass einige Wörter oder Phrasen (wie »Für Sie«, »Kreditkarte«, »kostenlos«, und »erstaunlich«) in der Betreffzeile gehäuft auftreten. Möglicherweise würden Ihnen auch weitere Muster im Namen des Absenders, dem Text und in anderen Teilen der E-Mail auffallen.

2. Sie würden für jedes der von Ihnen erkannten Muster einen Algorithmus schreiben, der dieses erkennt, und Ihr Programm würde E-Mails als Spam markieren, sobald eine bestimmte Anzahl dieser Muster erkannt wird.
3. Sie würden Ihr Programm testen und die Schritte 1 und 2 wiederholen, bis es gut genug ist.

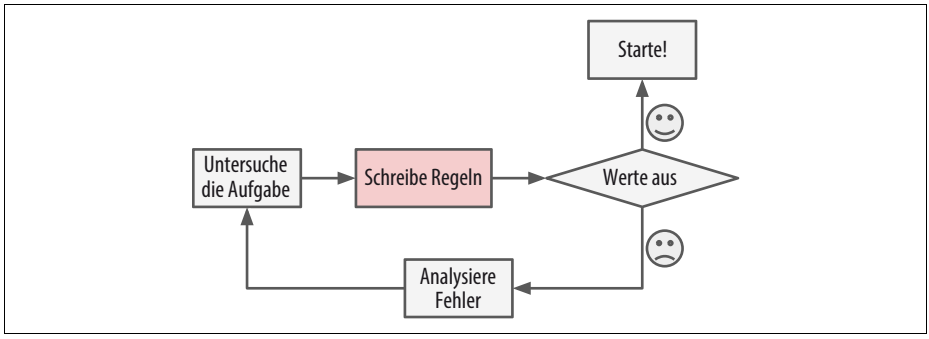


Abbildung 1-1: Die herkömmliche Herangehensweise

Da diese Aufgabe schwierig ist, wird Ihr Programm vermutlich eine lange Liste komplexer Regeln beinhalten – und ganz schön schwer zu warten sein.

Dagegen lernt ein mit Machine-Learning-Techniken entwickelter Spamfilter automatisch, welche Wörter und Phrasen Spam gut vorhersagen, indem er im Vergleich zu den Ham-Beispielen ungewöhnlich häufige Wortmuster in den Spambeispielen erkennt (siehe Abbildung 1-2). Das Programm wird viel kürzer, leichter zu warten und wahrscheinlich auch treffsicherer.

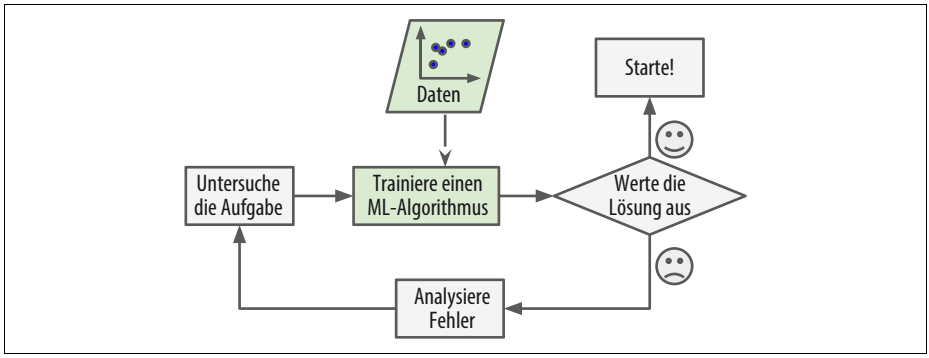


Abbildung 1-2: Der Machine-Learning-Ansatz

Wenn außerdem die Spammer bemerken, dass alle ihre E-Mails mit »Für Sie« geblockt werden, könnten sie stattdessen auf »4U« umsatteln. Ein mit traditionellen Programmiertechniken entwickelter Spamfilter müsste aktualisiert werden, um die E-Mails mit »4U« zu markieren. Wenn die Spammer sich ständig um Ihren Spamfilter herumarbeiten, werden Sie ewig neue Regeln schreiben müssen.

Ein auf Machine Learning basierender Spamfilter bemerkt dagegen automatisch, dass »4U« auffällig häufig in von Nutzern als Spam markierten Nachrichten vorkommt, und beginnt, diese ohne weitere Intervention auszusortieren (siehe Abbildung 1-3).

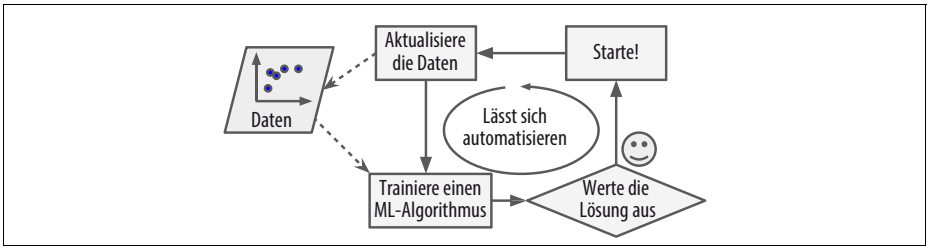


Abbildung 1-3: Automatisches Einstellen auf Änderungen

Machine Learning brilliert außerdem bei Aufgaben, die entweder zu komplex für herkömmliche Verfahren sind oder für die kein bekannter Algorithmus existiert. Betrachten Sie z. B. Spracherkennung: Sagen wir, Sie beginnen mit einer einfachen Aufgabe und schreiben ein Programm, das die Wörter »one« und »two« unterscheiden kann. Sie bemerken, dass das Wort »two« mit einem hochfrequenten Ton (»T«) beginnt, und könnten demnach einen Algorithmus hartcodieren, der die Intensität hochfrequenter Töne misst und dadurch »one« und »two« unterscheiden kann. Natürlich skaliert diese Technik nicht auf Tausende Wörter, die von Millionen von Menschen mit Hintergrundgeräuschen und in Dutzenden von Sprachen gesprochen werden. Die (zumindest heutzutage) beste Möglichkeit ist, einen Algorithmus zu schreiben, der eigenständig aus vielen aufgenommenen Beispielen für jedes Wort lernt.

Schließlich kann Machine Learning auch Menschen beim Lernen unterstützen (siehe Abbildung 1-4): ML-Algorithmen lassen sich untersuchen, um zu erkennen, was sie gelernt haben (auch wenn dies bei manchen Algorithmen kompliziert sein kann). Wenn z. B. der Spamfilter erst einmal mit genug Spam trainiert wurde, lässt sich die Liste von Wörtern und Wortkombinationen inspizieren, die als gute Merkmale von Spam erkannt wurden. Manchmal kommen dabei überraschende Korrelationen oder neue Trends ans Tageslicht und führen dadurch zu einem besseren Verständnis der Aufgabe. Das Anwenden von ML-Techniken zum Durchwühlen großer Datenmengen hilft dabei, nicht unmittelbar ersichtliche Muster zu finden. Dies nennt man auch *Data Mining*.

Zusammengefasst, ist Machine Learning hervorragend geeignet für:

- Aufgaben, bei denen die existierenden Lösungen eine Menge Feinarbeit oder lange Listen von Regeln erfordern: Ein maschineller Lernalgorithmus vereinfacht oft den Code und schneidet besser ab als der klassische Ansatz.
- Komplexe Aufgaben, für die es mit herkömmlichen Methoden überhaupt keine gute Lösung gibt: Die besten Machine-Learning-Techniken können vielleicht eine Lösung finden.

- Fluktuierende Umgebungen: Ein Machine-Learning-System kann sich neuen Daten anpassen.
- Erkenntnisse über komplexe Aufgabenstellungen und große Datenmengen zu gewinnen.

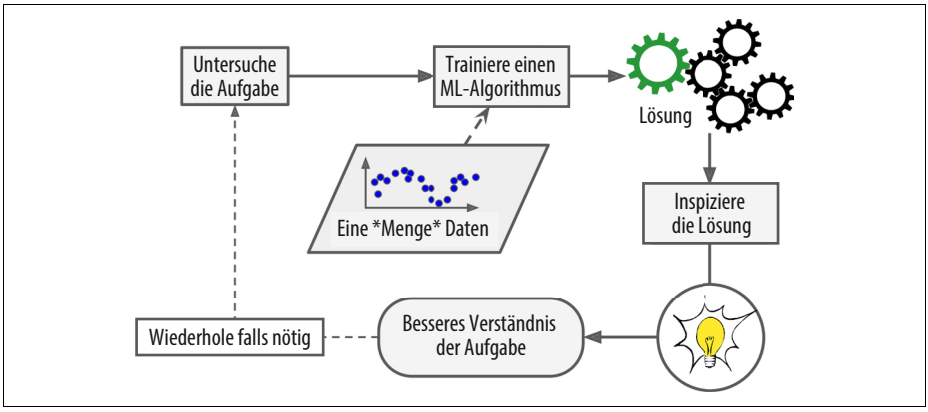


Abbildung 1-4: Machine Learning hilft Menschen beim Lernen.

Anwendungsbeispiel

Schauen wir uns ein paar konkrete Beispiele für Aufgaben des Machine Learning und die dabei eingesetzten Techniken an:

Produktbilder in der Herstellung analysieren, um sie automatisch zu klassifizieren
: Dies ist Bildklassifikation, meist ausgeführt durch Convolutional Neural Networks (CNNs, siehe Kapitel 14).

Tumoren in Gehirnscans erkennen
: Das ist eine semantische Segmentierung, bei der jedes Pixel im Bild klassifiziert wird (da wir die genaue Position und Form des Tumors bestimmen wollen), meist ebenfalls mithilfe von CNNs.

Nachrichtenartikel automatisch klassifizieren
: Dies ist linguistische Datenverarbeitung (NLP, Natural Language Processing) und, spezifischer, Textklassifikation, die sich über rekurrente neuronale Netze (RNNs), CNNs oder Transformer angehen lässt (siehe Kapitel 16).

Beleidigende Kommentare in Diskussionsforen automatisch markieren
: Dabei handelt es sich ebenfalls um Textklassifikation mit den gleichen NLP-Tools.

Automatisch lange Dokumente zusammenfassen
: Dies ist ein Zweig der NLP namens Textextrahierung, ebenfalls mit den gleichen Tools.

Einen Chatbot oder persönlichen Assistenten erstellen
: Dazu gehören viele NLP-Komponenten, unter anderem das Verstehen natürlicher Sprache (Natural Language Understanding, NLU) und Module zum Beantworten von Fragen.

Den Umsatz Ihrer Firma für das nächste Jahr basierend auf vielen Performancemetriken vorhersagen
: Dies ist eine Regressionsaufgabe (also das Vorhersagen von Werten), die über ein Regressionsmodell wie ein lineares oder polynomisches Regressionsmodell (siehe Kapitel 4), eine Regressions-SVM (siehe Kapitel 5), einen Regressions-Random-Forest (siehe Kapitel 7) oder ein künstliches neuronales Netzwerk (siehe Kapitel 10) angegangen werden kann. Wollen Sie Zeitreihen vergangener Metriken mit einbeziehen, können Sie RNNs, CNNs oder Transformer nutzen (siehe die Kapitel 15 und 16).

Kreditkartenmissbrauch erkennen
: Dies ist Anomalieerkennung (siehe Kapitel 9).

Kunden anhand ihrer Einkäufe segmentieren, sodass Sie für jeden Bereich unterschiedliche Marketingstrategien entwerfen können
: Das ist Clustering (siehe Kapitel 9).

Einen komplexen, hochdimensionalen Datensatz in einem klaren und verständlichen Diagramm darstellen
: Hier geht es um Datenvisualisierung, meist unter Verwendung von Techniken zur Datenreduktion (siehe Kapitel 8).

Einem Kunden basierend auf dessen bisherigen Käufen ein Produkt empfehlen, das ihn interessieren könnte
: Dies ist ein Empfehlungssystem. Ein Ansatz ist, Käufe aus der Vergangenheit (und andere Informationen über den Kunden) in ein künstliches neuronales Netzwerk einzuspeisen (siehe Kapitel 10) und es dazu zu bringen, den wahrscheinlichsten nächsten Kauf auszugeben. Dieses neuronale Netzwerk würde typischerweise mit bisherigen Abfolgen von Käufen aller Kunden trainiert werden.

Einen intelligenten Bot für ein Spiel bauen
: Dies wird oft durch Reinforcement Learning (RL, siehe Kapitel 18) angegangen. Dabei handelt es sich um einen Zweig des Machine Learning, der Agenten trainiert (zum Beispiel Bots), um die Aktionen auszuwählen, die mit der Zeit in einer gegebenen Umgebung (wie dem Spiel) ihre Belohnungen maximieren (beispielsweise kann ein Bot immer dann eine Belohnung erhalten, wenn der Spieler Lebenspunkte verliert). Das berühmte AlphaGo-Programm, das den Weltmeister im Go geschlagen hat, wurde mithilfe von RL gebaut.

Diese Liste könnte immer weiter fortgeführt werden, aber hoffentlich haben Sie auch so schon einen Eindruck von der unglaublichen Breite und Komplexität der Aufgaben erhalten, die Machine Learning angehen kann, und die Art von Techniken, die Sie dafür verwenden würden.

Unterschiedliche Machine-Learning-Systeme

Es gibt so viele verschiedene Arten von Machine-Learning-Systemen, dass es hilfreich ist, die Verfahren nach folgenden Kriterien in grobe Kategorien einzuteilen:

- Ob sie mit menschlicher Überwachung trainiert werden oder nicht (überwachtes, unüberwachtes und halbüberwachtes Lernen sowie Reinforcement Learning).
- Ob sie inkrementell dazulernen können oder nicht (Online-Learning gegenüber Batch-Learning).
- Ob sie einfach neue Datenpunkte mit den bereits bekannten Datenpunkten vergleichen oder stattdessen Muster in den Trainingsdaten erkennen, um ein Vorhersagemodell aufzubauen, wie es auch Wissenschaftler tun (instanzbasiertes gegenüber modellbasiertem Lernen).

Diese Kriterien schließen sich nicht gegenseitig aus; sie lassen sich beliebig miteinander kombinieren. Zum Beispiel kann ein moderner Spamfilter ständig mit einem neuronalen Netzwerkmodell mit Beispielen für Spam und Ham dazulernen; damit ist er ein modellbasiertes, überwachtes Onlinelernsystem.

Betrachten wir jedes dieser Kriterien etwas genauer.

Überwachtes/unüberwachtes Lernen

Machine-Learning-Systeme lassen sich entsprechend der Menge und Art der Überwachung beim Trainieren einordnen. Es gibt dabei vier größere Kategorien: überwachtes Lernen, unüberwachtes Lernen, halbüberwachtes Lernen und verstärkendes Lernen (Reinforcement Learning).

Überwachtes Lernen

Beim *überwachten Lernen* enthalten die dem Algorithmus gelieferten Trainingsdaten die gewünschten Lösungen, genannt *Labels* (siehe Abbildung 1-5).

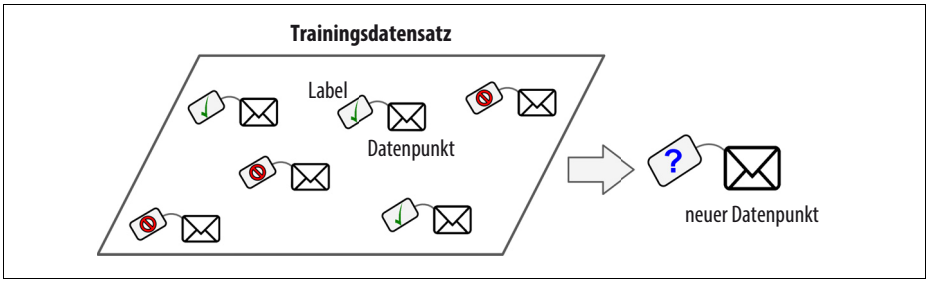


Abbildung 1-5: Ein Trainingsdatensatz mit Labels zur Spamerkennung (ein Beispiel für überwachtes Lernen)

Klassifikation ist eine typische überwachte Lernaufgabe. Spamfilter sind hierfür ein gutes Beispiel: Sie werden mit vielen Beispiel-E-Mails und deren *Kategorie* (Spam oder Ham) trainiert und müssen lernen, neue E-Mails zu klassifizieren.

Eine weitere typische Aufgabe ist, eine numerische *Zielgröße* vorherzusagen, wie etwa den Preis eines Autos auf Grundlage gegebener *Merkmale* (gefahrene Kilometer, Alter, Marke und so weiter), den sogenannten *Prädiktoren*. Diese Art Aufgabe bezeichnet man als *Regression* (siehe Abbildung 1-6).[1] Um das System zu trainieren, benötigt es viele Beispielfahrzeuge mitsamt ihren Prädiktoren und Labels (also den Preisen).

Beim Machine Learning ist ein *Attribut* ein Datentyp (z. B. »Kilometerzahl«), wohingegen ein *Merkmal* je nach Kontext mehrere Bedeutungen haben kann. Meist ist damit ein Attribut und dessen Wert gemeint (z. B. »Kilometerzahl = 15000«). Allerdings verwenden viele Anwender *Attribut* und *Merkmal* synonym.

Viele Regressionsalgorithmen lassen sich auch zur Klassifikation einsetzen und umgekehrt. Zum Beispiel ist die *logistische Regression* eine verbreitete Methode für Klassifikationsaufgaben, da sie die Wahrscheinlichkeit der Zugehörigkeit zu einer bestimmten Kategorie als Ergebnis liefert (z. B. 20%ige Chance für Spam).

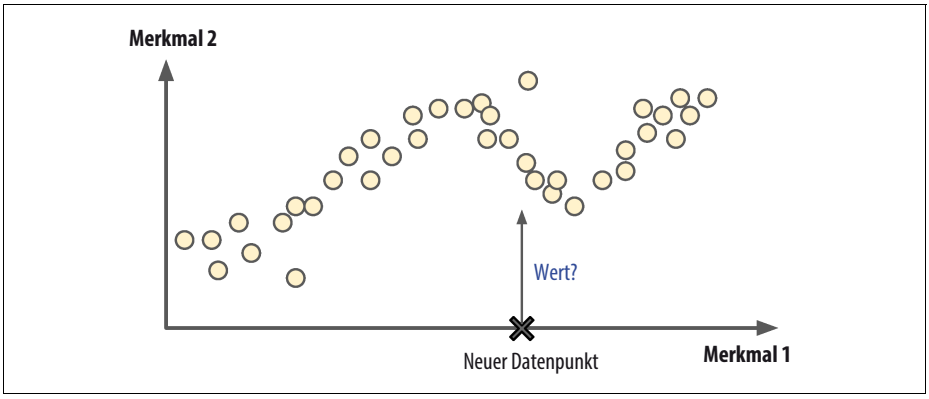


Abbildung 1-6: Ein Regressionsproblem: Sage mithilfe eines Eingangsmerkmals einen Wert voraus (es gibt meist mehrere Eingangsmerkmale und manchmal auch mehrere Ausgangsmerkmale).

Hier sind ein paar der wichtigsten (in diesem Buch besprochenen) überwachten Lernalgorithmen:

1 Wissenswert: Dieser seltsam anmutende Begriff wurde von Francis Galton in die Statistik eingeführt, der beobachtete, dass die Kinder hochgewachsener Eltern zu einer geringeren Körpergröße als ihre Eltern neigen. Da die Kinder kleiner waren, nannte er dies *Regression zum Mittelwert*. Dieser Name wurde anschließend auch für seine Methode zur Analyse von Korrelationen zwischen Variablen verwendet.

- k-nächste-Nachbarn
- lineare Regression
- logistische Regression
- Support Vector Machines (SVMs)
- Entscheidungsbäume und Random Forests
- neuronale Netzwerke[2]

Unüberwachtes Lernen

Beim *unüberwachten Lernen*sind die Trainingsdaten, wie der Name vermuten lässt, nicht gelabelt (siehe Abbildung 1-7). Das System versucht, ohne Anleitung zu lernen.

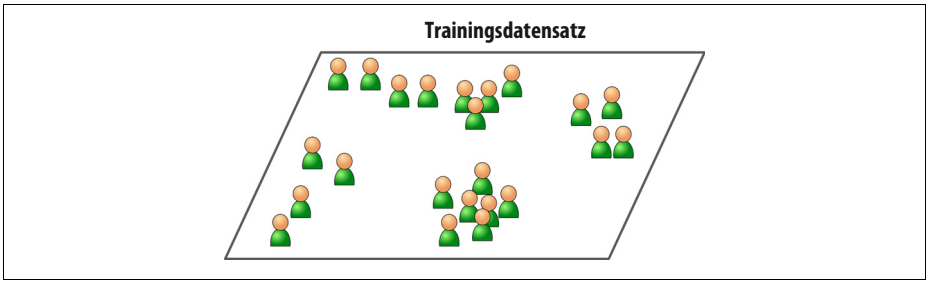


Abbildung 1-7: Ein Trainingsdatensatz ohne Labels für unüberwachtes Lernen

Hier sind ein paar der wichtigsten unüberwachten Lernalgorithmen (wir werden die Dimensionsreduktion in Kapitel 8 und Kapitel 9 behandeln):

- Clustering
  - k-Means
  - DBSCAN
  - hierarchische Cluster-Analyse (HCA)
- Anomalieerkennung und Novelty Detection
  - One-Class-SVM
  - Isolation Forest
- Visualisierung und Dimensionsreduktion
  - Hauptkomponentenzerlegung (PCA)
  - Kernel-PCA
  - Locally-Linear Embedding (LLE)
  - t-verteiltes Stochastic Neighbor Embedding (t-SNE)

2 Einige neuronale Netzwerkarchitekturen können unüberwacht sein, wie beispielsweise Autoencoder und Restricted Boltzmann Machines. Sie können auch halbüberwacht sein, wie bei Deep Belief Networks und unüberwachtem Vortrainieren.

- Lernen mit Assoziationsregeln
  - Apriori
  - Eclat

Nehmen wir an, Sie hätten eine Menge Daten über Besucher Ihres Blogs. Sie möchten einen *Clustering-Algorithmus* verwenden, um Gruppen ähnlicher Besucher zu entdecken (siehe Abbildung 1-8). Sie verraten dem Algorithmus nichts darüber, welcher Gruppe ein Besucher angehört, er findet die Verbindungen ohne Ihr Zutun heraus. Beispielsweise könnte der Algorithmus bemerken, dass 40 % Ihrer Besucher Männer mit einer Vorliebe für Comics sind, die Ihr Blog abends lesen, 20 % dagegen sind junge Science-Fiction-Fans, die am Wochenende vorbeischauen. Wenn Sie ein *hierarchisches Cluster-Verfahren* verwenden, können Sie sogar jede Gruppe in weitere Untergruppen zerlegen, was hilfreich sein kann, wenn Sie Ihre Blogartikel auf diese Zielgruppen zuschneiden möchten.

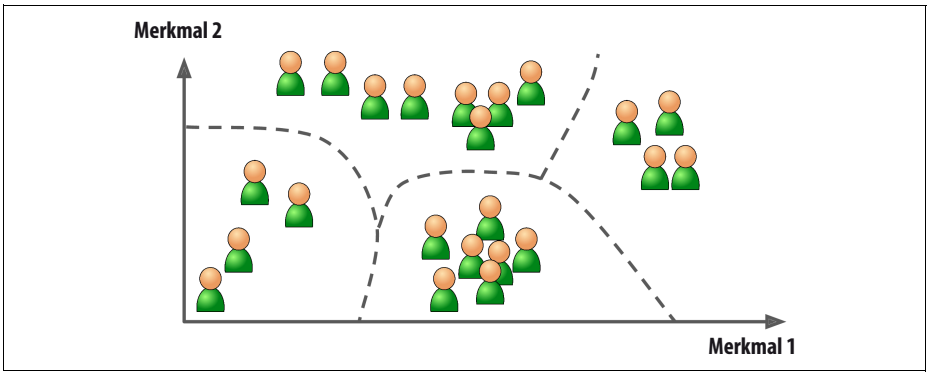


Abbildung 1-8: Clustering

Algorithmen zur *Visualisierung* sind ebenfalls ein gutes Beispiel für unüberwachtes Lernen: Sie übergeben diesen eine Menge komplexer Daten ohne Labels und erhalten eine 2-D- oder 3-D-Repräsentation der Daten, die Sie leicht grafisch darstellen können (siehe Abbildung 1-9). Solche Algorithmen versuchen, die Struktur der Daten so gut wie möglich zu erhalten (z. B. Cluster in den Eingabedaten am Überlappen in der Visualisierung zu hindern), sodass Sie leichter verstehen können, wie die Daten aufgebaut sind, und womöglich auf unvermutete Muster stoßen.

Eine verwandte Aufgabe ist die *Dimensionsreduktion*, bei der das Ziel die Vereinfachung der Daten ist, ohne dabei allzu viele Informationen zu verlieren. Dazu lassen sich mehrere korrelierende Merkmale zu einem vereinigen. Beispielsweise korreliert der Kilometerstand eines Autos stark mit seinem Alter, daher kann ein Algorithmus zur Dimensionsreduktion beide zu einem Merkmal verbinden, das die Abnutzung des Fahrzeugs repräsentiert. Dies nennt man auch *Extraktion von Merkmalen*.

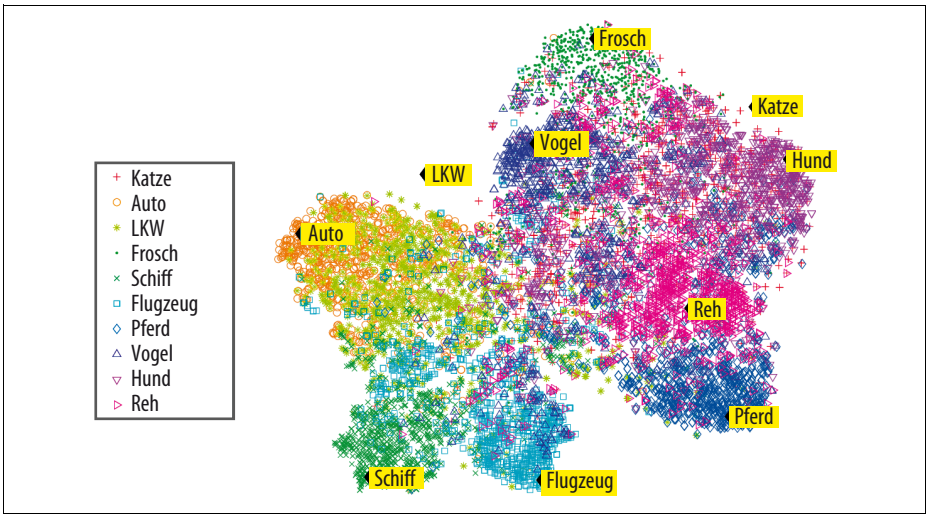


Abbildung 1-9: Beispiel für eine t-SNE-Visualisierung semantischer Cluster[3]

Meist ist es eine gute Idee, die Dimensionen Ihrer Trainingsdaten zu reduzieren, bevor Sie sie in einen anderen Machine-Learning-Algorithmus einspeisen (wie etwa einen überwachten Lernalgorithmus). Er wird viel schneller arbeiten, und die Daten beanspruchen weniger Platz auf der Festplatte und im Speicher. In manchen Fällen ist auch das Ergebnis besser.

Eine weitere wichtige unüberwachte Aufgabe ist das *Erkennen von Anomalien* – beispielsweise ungewöhnliche Transaktionen auf Kreditkarten, die auf Betrug hindeuten, das Abfangen von Produktionsfehlern oder das automatische Entfernen von Ausreißern aus einem Datensatz, bevor dieser in einen weiteren Lernalgorithmus eingespeist wird. Dem System werden beim Training vor allem gewöhnliche Datenpunkte präsentiert, sodass es lernt, sie zu erkennen. Sieht es dann einen neuen Datenpunkt, kann es entscheiden, ob dieser wie ein normaler Punkt oder eine Anomalie aussieht (siehe Abbildung 1-10). Eine sehr ähnliche Aufgabe ist die *Novelty Detection*: Ihr Ziel ist es, neue Instanzen zu erkennen, die anders als alle anderen Instanzen im Trainingsdatensatz aussehen. Dafür brauchen Sie einen sehr »sauberen« Trainingsdatensatz, der von allen Instanzen befreit ist, die der Algorithmus erkennen soll. Haben Sie beispielsweise Tausende von Bildern mit Hunden und sind nur auf 1 % dieser Bilder Chihuahuas zu sehen, sollte ein Algorithmus zur Novelty Detection neue Bilder von Chihuahuas nicht als Besonderheiten behandeln. Ein Algorithmus zur Anomalieerkennung mag diese Hunde allerdings als so

3 Beachten Sie, dass die Tiere recht gut von Fahrzeugen unterschieden werden und wie nahe sich Pferde und Rehe sind, aber wie weit entfernt von Vögeln. Abbildung mit Erlaubnis reproduziert, Quelle: Richard Socher et al., »Zero-Shot Learning Through Cross-Modal Transfer«, *Proceedings of the 26th International Conference on Neural Information Processing Systems* 1 (2013): 935–943.

selten ansehen – und als so anders als andere Hunde –, dass er sie als Anomalien klassifizieren würde (nichts gegen Chihuahuas).

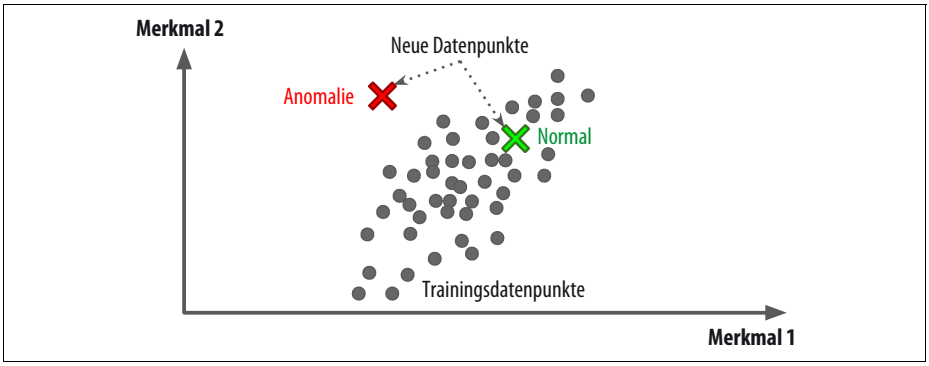


Abbildung 1-10: Erkennen von Anomalien

Schließlich ist auch das *Lernen von Assoziationsregeln* eine verbreitete unüberwachte Lernaufgabe, bei der das Ziel ist, in große Datenmengen einzutauchen und interessante Beziehungen zwischen Merkmalen zu entdecken. Wenn Sie beispielsweise einen Supermarkt führen, könnten Assoziationsregeln auf Ihren Verkaufsdaten ergeben, dass Kunden, die Grillsoße und Kartoffelchips einkaufen, tendenziell auch Steaks kaufen. Daher sollten Sie diese Artikel in unmittelbarer Nähe zueinander platzieren.

Halbüberwachtes Lernen

Da das Labeling normalerweise zeitaufwendig und teuer ist, werden Sie oftmals sehr viele ungelabelte und wenige gelabelte Instanzen haben. Einige Algorithmen können mit nur teilweise gelabelten Trainingsdaten arbeiten. Dies bezeichnet man als *halbüberwachtes Lernen* (siehe Abbildung 1-11).

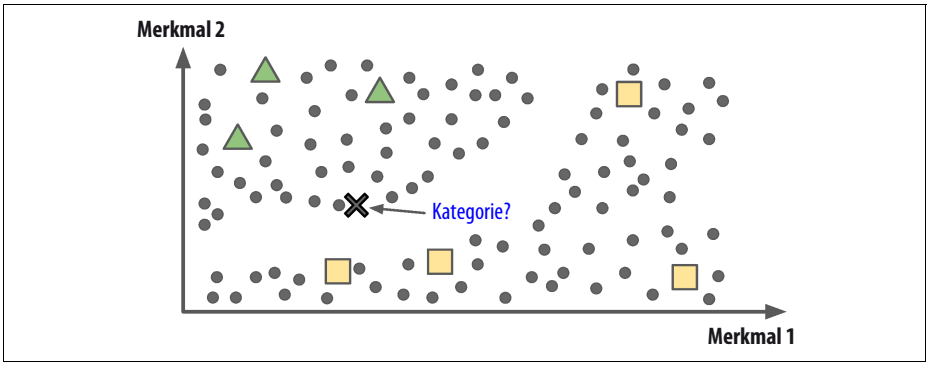


Abbildung 1-11: Halbüberwachtes Lernen mit zwei Klassen (Dreiecke und Quadrate): Die ungelabelten Beispiele (Kreise) helfen dabei, eine neue Instanz (das Kreuz) in die Dreiecksklasse statt in die Quadratklasse einzuordnen, auch wenn sie näher an den gelabelten Quadraten ist.

Einige Fotodienste wie Google Photos bieten hierfür ein gutes Beispiel. Sobald Sie all Ihre Familienfotos in den Dienst hochgeladen haben, erkennt dieser automatisch, dass die gleiche Person A auf den Fotos 1, 5 und 11 vorkommt, während Person B auf den Fotos 2, 5 und 7 zu sehen ist. Dies ist der unüberwachte Teil des Algorithmus (Clustering). Nun muss das System nur noch wissen, wer diese Personen sind. Ein Label pro Person[4] genügt, um jede Person in jedem Foto zuzuordnen, was bei der Suche nach Fotos äußerst nützlich ist.

Die meisten Algorithmen für halbüberwachtes Lernen sind Kombinationen aus unüberwachten und überwachten Verfahren. Beispielsweise beruhen *Deep Belief Networks* (DBNs) auf in Reihe geschalteten unüberwachten Komponenten namens *restricted Boltzmann Machines* (RBMs). Die RBMs werden nacheinander unüberwacht trainiert. Die Feinabstimmung des Gesamtsystems findet anschließend mit überwachten Lerntechniken statt.

Reinforcement Learning

Reinforcement Learning ist etwas völlig anderes. Das Lernsystem, in diesem Zusammenhang als *Agent* bezeichnet, beobachtet eine Umgebung, wählt Aktionen und führt diese aus. Dafür erhält es *Belohnungen* (oder *Strafen* in Form negativer Belohnungen wie in Abbildung 1-12). Das System muss selbst herausfinden, was die beste Strategie oder *Policy* ist, um mit der Zeit die meisten Belohnungen zu erhalten. Eine Policy definiert, welche Aktion der Agent in einer gegebenen Situation auswählt.

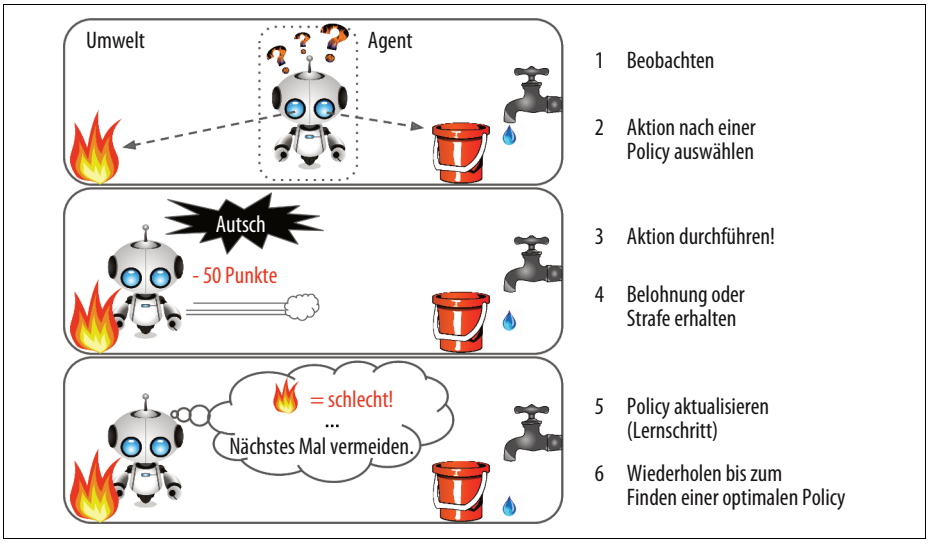


Abbildung 1-12: Reinforcement Learning

4 Dies ist der Fall, wenn das System perfekt funktioniert. In der Praxis werden meist einige Cluster pro Person erstellt. Manchmal werden zwei ähnliche Personen verwechselt, sodass Sie einige Labels pro Person angeben und außerdem ein paar Cluster von Hand aufräumen müssen.

Beispielsweise verwenden viele Roboter Reinforcement-Learning-Algorithmen, um laufen zu lernen. Auch das Programm AlphaGo von DeepMind ist ein gutes Beispiel für Reinforcement Learning: Es geriet im Mai 2017 in die Schlagzeilen, als es den Weltmeister Ke Jie im Brettspiel *Go* schlug. AlphaGo erlernte die zum Sieg führende Policy, indem es Millionen von Partien analysierte und anschließend viele Spiele gegen sich selbst spielte. Beachten Sie, dass das Lernen während der Partien gegen den Weltmeister abgeschaltet war; AlphaGo wandte nur die bereits erlernte Policy an.

Batch- und Online-Learning

Ein weiteres Kriterium zum Einteilen von Machine-Learning-Systemen ist, ob das System aus einem kontinuierlichen Datenstrom inkrementell lernen kann.

Batch-Learning

Beim *Batch-Learning* kann das System nicht inkrementell lernen, es muss mit sämtlichen verfügbaren Daten trainiert werden. Dies dauert meist lange und beansprucht Rechenkapazitäten. Es wird daher in der Regel offline durchgeführt. Zuerst wird das System trainiert und anschließend in einer Produktivumgebung eingesetzt, wo es ohne weiteres Lernen läuft; es wendet lediglich das bereits Erlernte an. Dies nennt man*Offline-Learning*.

Wenn Sie möchten, dass ein Batch-Learning-System etwas über neue Daten erfährt (beispielsweise neuartigen Spam), müssen Sie eine neue Version des Systems ein weiteres Mal mit dem gesamten Datensatz trainieren (nicht einfach nur den neuen Datensatz, sondern auch den alten). Anschließend müssen Sie das alte System anhalten und durch das neue ersetzen.

Glücklicherweise lässt sich der gesamte Prozess aus Training, Evaluation und Inbetriebnahme eines Machine-Learning-Systems recht leicht automatisieren (wie in Abbildung 1-3 gezeigt). So kann sich selbst ein Batch-Learning-System anpassen. Aktualisieren Sie einfach die Daten und trainieren Sie eine neue Version des Systems so oft wie nötig.

Dies ist eine einfache Lösung und funktioniert meist gut, aber das Trainieren mit dem gesamten Datensatz kann viele Stunden beanspruchen. Daher würde man das neue System nur alle 24 Stunden oder wöchentlich trainieren. Wenn Ihr System sich an schnell ändernde Daten anpassen muss (z. B. um Aktienkurse vorherzusagen), benötigen Sie eine anpassungsfähigere Lösung.

Außerdem beansprucht das Trainieren auf dem gesamten Datensatz eine Menge Rechenkapazität (CPU, Hauptspeicher, Plattenplatz, I/O-Kapazität, Netzwerkbandbreite und so weiter). Wenn Sie eine Menge Daten haben und Ihr System automatisch jeden Tag trainieren lassen, kann Sie das am Ende eine Stange Geld kosten.

Falls die Datenmenge sehr groß ist, kann der Einsatz von Batch-Learning sogar unmöglich sein.

Wenn Ihr System autonom lernen muss und die Ressourcen dazu begrenzt sind (z. B. eine Applikation auf einem Smartphone oder ein Fahrzeug auf dem Mars), ist das Herumschleppen großer Mengen an Trainingsdaten oder das Belegen einer Menge Ressourcen für mehrere Stunden am Tag kein gangbarer Weg.

In all diesen Fällen sind Algorithmen, die inkrementell lernen können, eine bessere Alternative.

Online-Learning

Beim *Online-Learning*, wird das System nach und nach trainiert, indem einzelne Datensätze nacheinander oder in kleinen Paketen, sogenannten *Mini-Batches*, hinzugefügt werden. Jeder Lernschritt ist schnell und billig, sodass das System aus neuen Daten lernen kann, sobald diese verfügbar sind (siehe Abbildung 1-13).

Online-Learning eignet sich großartig für ein System mit kontinuierlich eintreffenden Daten (z. B. Aktienkursen), das sich entweder schnell oder autonom an Veränderungen anpassen muss. Auch wenn Ihnen nur begrenzte Rechenkapazitäten zur Verfügung stehen, ist es eine sinnvolle Option: Sobald ein Online-Learning-System die neuen Datenpunkte erlernt hat, werden diese nicht mehr benötigt und können verworfen werden (es sei denn, Sie möchten in der Lage sein, zu einem früheren Zustand zurückzukehren und den Datenstrom erneut »abzuspielen«). Dies kann enorme Mengen an Speicherplatz einsparen.

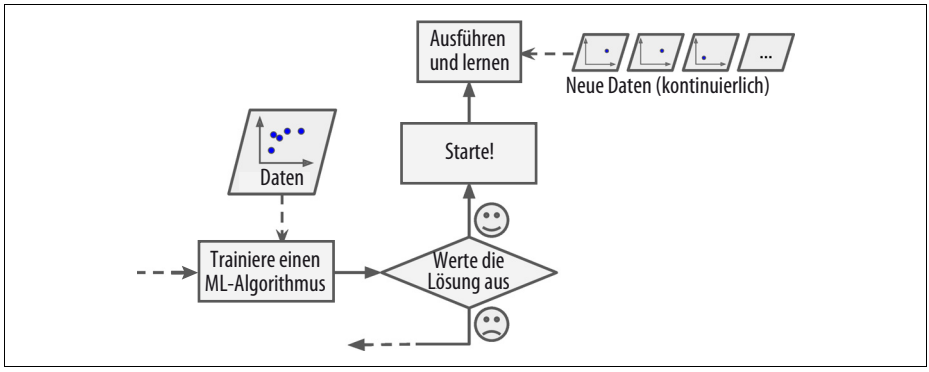


Abbildung 1-13: Beim Online-Learning wird ein Modell trainiert und in den Produktivbetrieb übernommen, wo es mit neu eintreffenden Daten weiterlernt.

Algorithmen zum Online-Learning lassen sich auch zum Trainieren von Systemen mit riesigen Datensätzen einsetzen, die nicht in den Hauptspeicher eines Rechners passen (dies nennt man auch *Out-of-Core*-Lernen). Der Algorithmus lädt einen Teil der Daten, führt einen Trainingsschritt auf den Daten aus und wiederholt den Prozess, bis er sämtliche Daten verarbeitet hat (siehe Abbildung 1-14).

Out-of-Core-Lernen wird für gewöhnlich offline durchgeführt (also nicht auf einem Produktivsystem), daher ist der Begriff *Online-Learning* etwas irreführend. Stellen Sie sich darunter eher *inkrementelles Lernen* vor.

Ein wichtiger Parameter bei Online-Learning-Systemen ist, wie schnell sie sich an sich verändernde Daten anpassen. Man spricht hier von der *Lernrate*. Wenn Sie die Lernrate hoch ansetzen, wird sich Ihr System schnell auf neue Daten einstellen, aber die alten Daten auch leicht wieder vergessen (Sie möchten sicher nicht, dass ein Spamfilter nur die zuletzt gesehenen Arten von Spam erkennt). Wenn Sie die Lernrate dagegen niedrig ansetzen, entwickelt das System eine höhere Trägheit; das bedeutet, es lernt langsamer, ist aber auch weniger anfällig für Rauschen in den neuen Daten oder für Folgen nicht repräsentativer Datenpunkte (Outlier).

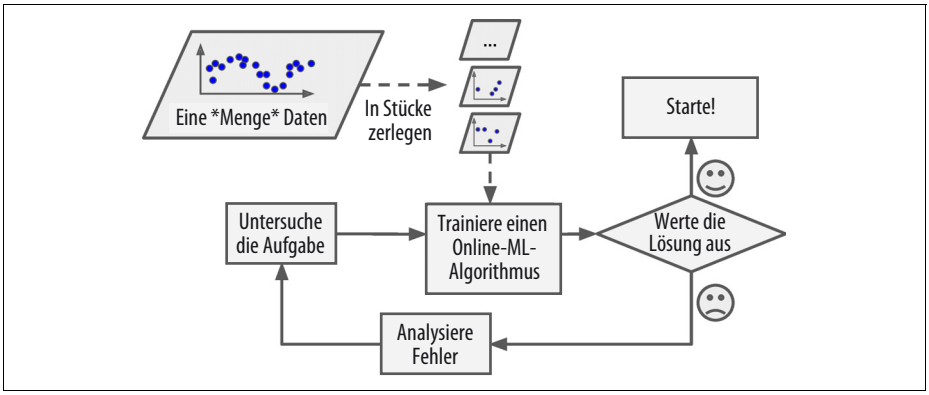


Abbildung 1-14: Verwenden von Online-Learning zum Bewältigen riesiger Datensätze

Eine große Herausforderung beim Online-Learning besteht darin, dass in das System eingespeiste minderwertige Daten zu einer allmählichen Verschlechterung seiner Leistung führen. Wenn es sich dabei um ein Produktivsystem handelt, werden Ihre Kunden dies bemerken. Beispielsweise könnten minderwertige Daten von einem fehlerhaften Sensor an einem Roboter oder auch von jemandem stammen, der sein Ranking in einer Suchmaschine durch massenhafte Anfragen zu verbessern versucht. Um dieses Risiko zu reduzieren, müssen Sie Ihr System aufmerksam beobachten und den Lernprozess beherzt abschalten (und eventuell auf einen früheren Zustand zurücksetzen), sobald Sie einen Leistungsabfall bemerken. Sie können auch die Eingabedaten verfolgen und auf ungewöhnliche Daten reagieren (z. B. über einen Algorithmus zur Erkennung von Anomalien).

Instanzbasiertes versus modellbasiertes Lernen

Eine weitere Möglichkeit, maschinelle Lernverfahren zu kategorisieren, ist die Art, wie diese *verallgemeinern*. Bei den meisten Aufgaben im Machine Learning geht es um das Treffen von Vorhersagen. Dabei muss ein System in der Lage sein, aus einer

Anzahl von Trainingsbeispielen auf nie zuvor gesehene Beispiele zu verallgemeinern. Es ist hilfreich, aber nicht ausreichend, eine gute Leistung auf den Trainingsdaten zu erzielen; das wirkliche Ziel ist, eine gute Leistung auf neuen Datenpunkten zu erreichen.

Es gibt beim Verallgemeinern zwei Ansätze: instanzbasiertes Lernen und modellbasiertes Lernen.

Instanzbasiertes Lernen

Die vermutlich trivialste Art zu lernen, ist das einfache Auswendiglernen. Wenn Sie auf diese Weise einen Spamfilter erstellten, würde dieser einfach alle E-Mails aussortieren, die mit bereits von Nutzern markierten E-Mails identisch sind – nicht die schlechteste Lösung, aber sicher nicht die beste.

Anstatt einfach mit bekannten Spam-E-Mails identische Nachrichten zu markieren, könnte Ihr Spamfilter auch so programmiert sein, dass darüber hinaus bekannten Spam-E-Mails sehr ähnliche Nachrichten markiert werden. Dazu ist ein *Ähnlichkeitsmaß* zwischen zwei E-Mails nötig. Ein (sehr einfaches) Maß für die Ähnlichkeit zweier E-Mails könnte die Anzahl gemeinsamer Wörter sein. Das System könnte eine E-Mail als Spam markieren, wenn diese viele gemeinsame Wörter mit einer bekannten Spamnachricht aufweist.

Dies nennt man *instanzbasiertes Lernen*: Das System lernt die Beispiele auswendig und verallgemeinert dann mithilfe eines Ähnlichkeitsmaßes auf neue Fälle, wobei es sie mit den gelernten Beispielen (oder einer Untermenge davon) vergleicht. So würden beispielsweise in Abbildung 1-15 die neuen Instanzen als Dreieck klassifiziert werden, weil die Mehrheit der ähnlichsten Instanzen zu dieser Klasse gehört.

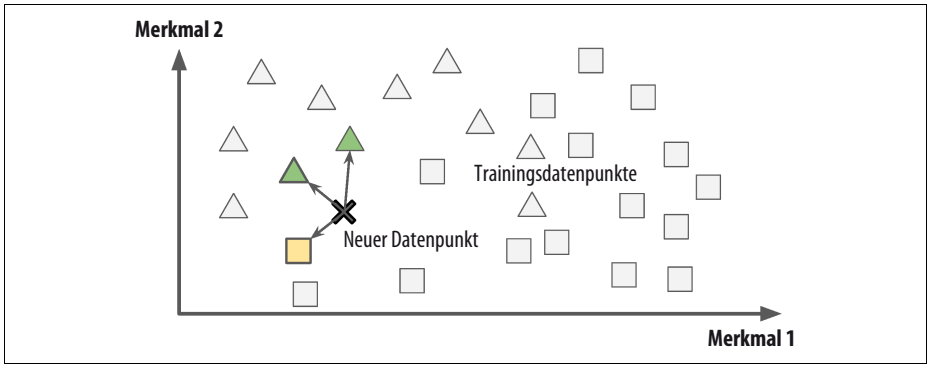


Abbildung 1-15: Instanzbasiertes Lernen

Modellbasiertes Lernen

Eine andere Möglichkeit, von einem Beispieldatensatz zu verallgemeinern, ist, ein Modell aus den Beispielen zu entwickeln und dieses Modell dann für *Vorhersagen* zu verwenden. Das wird als *modellbasiertes Lernen* bezeichnet (siehe Abbildung 1-16).

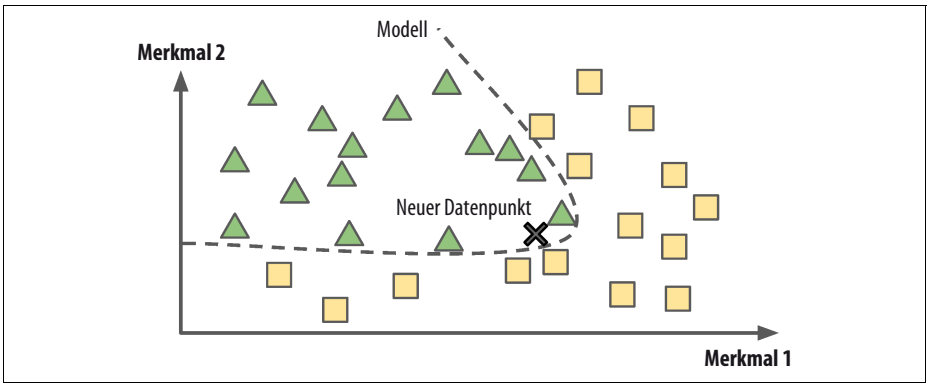


Abbildung 1-16: Modellbasiertes Lernen

Nehmen wir an, Sie möchten herausfinden, ob Geld glücklich macht. Sie laden dazu die Daten des *Better Life Index* von der Webseite des OECD (*https://homl.info/4*) herunter und Statistiken zum Pro-Kopf-Bruttoinlandsprodukt (BIP) von der Webseite des IMF (*https://homl.info/5*). Anschließend führen Sie beide Tabellen zusammen und sortieren nach dem BIP pro Kopf. Tabelle 1-1 zeigt einen Ausschnitt des Ergebnisses.

Tabelle 1-1: Macht Geld Menschen glücklicher?

Land	BIP pro Kopf (USD)	Zufriedenheit
Ungarn	12240	4,9
Korea	27195	5,8
Frankreich	37675	6,5
Australien	50962	7,3
Vereinigte Staaten	55805	7,2

Stellen wir die Daten für einige dieser Länder dar (siehe Abbildung 1-17).

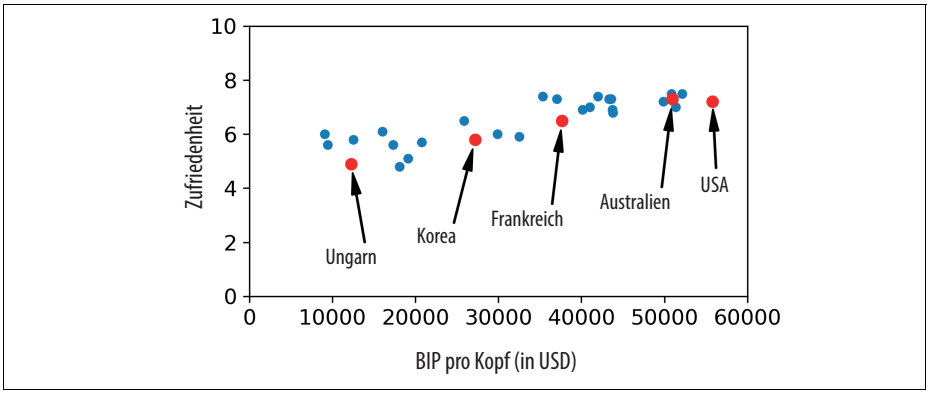


Abbildung 1-17: Sehen Sie hier eine Tendenz?

Es scheint so etwas wie einen Trend zu geben! Auch wenn die Daten *verrauscht* sind (also teilweise zufällig), sieht es so aus, als stiege die Zufriedenheit mehr oder weniger mit dem Pro-Kopf-BIP des Lands linear an. Sie beschließen also, die Zufriedenheit als lineare Funktion des Pro-Kopf-Bruttoinlandsprodukts zu modellieren. Diesen Schritt bezeichnet man als *Modellauswahl*: Sie wählen ein *lineares Modell* der Zufriedenheit mit genau einem Merkmal aus, nämlich dem Pro-Kopf-BIP (siehe Formel 1-1).

Formel 1-1: Ein einfaches lineares Modell

$$\text{Zufriedenheit} = \theta_0 + \theta_1 \times \text{BIP\_pro\_Kopf}$$

Diesen Modell enthält zwei *Modellparameter*, θ_0 und θ_1.[5] Indem Sie diese Parameter verändern, kann das Modell jede lineare Funktion annehmen, wie Sie in Abbildung 1-18 sehen können.

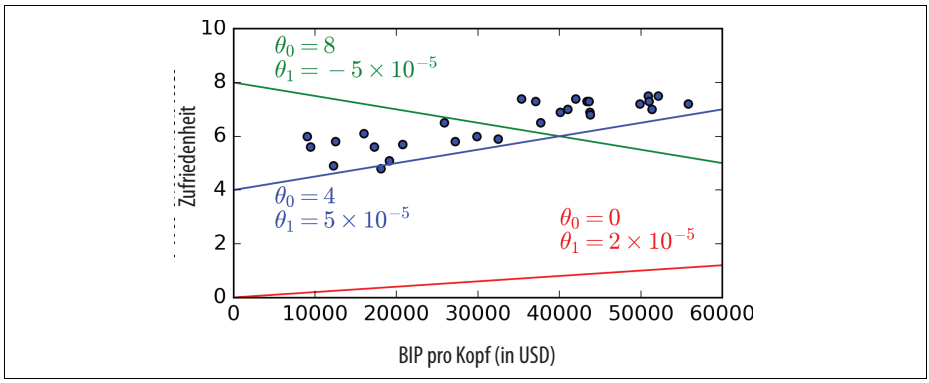


Abbildung 1-18: Einige mögliche lineare Modelle

Bevor Sie Ihr Modell verwenden können, müssen Sie die Werte der Parameter θ_0 und θ_1 festlegen. Woher sollen Sie wissen, welche Werte zur bestmöglichen Leistung führen? Um diese Frage zu beantworten, müssen Sie ein Maß für die Leistung festlegen. Sie können dafür entweder eine *Nutzenfunktion* (oder *Fitnessfunktion*) verwenden, die die *Güte* Ihres Modells bestimmt. Alternativ können Sie eine *Kostenfunktion* definieren, die misst, wie *schlecht* das Modell ist. Bei linearen Modellen verwendet man typischerweise eine Kostenfunktion, die die Entfernung zwischen den Vorhersagen des linearen Modells und den Trainingsbeispielen bestimmt; das Ziel ist, diese Entfernung zu minimieren.

An dieser Stelle kommt der Algorithmus zur linearen Regression ins Spiel: Sie speisen Ihre Trainingsdaten ein, und der Algorithmus ermittelt die für Ihre Daten bestmöglichen Parameter des linearen Modells. Dies bezeichnet man als *Trainieren* des Modells. In unserem Fall ermittelt der Algorithmus $\theta_0 = 4{,}85$ und $\theta_1 = 4{,}91 \times 10^{-5}$ als optimale Werte für die beiden Parameter.

5 Der griechische Buchstabe θ (Theta) wird konventionsgemäß häufig zum Darstellen von Modellparametern verwendet.

Verwirrenderweise kann sich das gleiche Wort »Modell« auf eine Art von Modell (zum Beispiel lineare Regression), auf eine vollständig spezifizierte Modellarchitektur (zum Beispiel lineare Regression mit einer Eingabe und einer Ausgabe) oder auf das abschließende trainierte Modell, das für Vorhersagen genutzt werden kann (zum Beispiel lineare Regression mit einer Eingabe und einer Ausgabe und $\theta_0 = 4{,}85$ und $\theta_1 = 4{,}91 \times 10^{-5}$), verwendet werden. Die Modellauswahl besteht darin, die Art des Modells auszuwählen und seine Architektur vollständig zu definieren. Beim Trainieren eines Modells geht es darum, einen Algorithmus laufen zu lassen, der die Modellparameter findet, mit denen es am besten zu den Trainingsdaten passt (und hoffentlich gute Vorhersagen für neue Daten trifft).

Nun passt das Modell (für ein lineares Modell) bestmöglich zu den Trainingsdaten, wie Abbildung 1-19 zeigt.

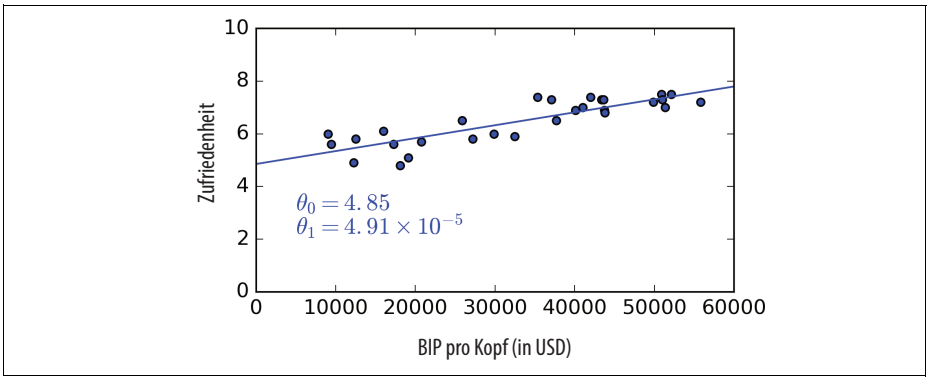


Abbildung 1-19: Das an die Trainingsdaten optimal angepasste lineare Modell

Sie sind nun endlich so weit, das Modell für Vorhersagen einzusetzen. Nehmen wir an, Sie möchten wissen, wie glücklich Zyprioten sind. Die Daten des OECD liefern darauf keine Antwort. Glücklicherweise können Sie unser Modell verwenden, um eine gute Vorhersage zu treffen: Sie schlagen das Pro-Kopf-BIP für Zypern nach, finden 22587 USD und wenden Ihr Modell an. Dabei finden Sie heraus, dass die Zufriedenheit irgendwo um $4{,}85 + 22{,}587 \times 4{,}91 \times 10^{-5} = 5{,}96$ liegt.

Um Ihren Appetit auf die folgenden Kapitel anzuregen, zeigt Beispiel 1-1 den Python-Code, der die Daten lädt, vorbereitet,[6] einen Scatterplot zeichnet, ein lineares Modell trainiert und eine Vorhersage trifft.[7]

6 Die Definition der Funktion `prepare_country_stats()` ist hier nicht gezeigt (Sie finden sie im Jupyter-Notebook zu diesem Kapitel, wenn Sie sie in all ihrer Schönheit betrachten wollen). Es handelt es nur um langweiligen pandas-Code, der die Daten zur Zufriedenheit der OECD mit den BIP-Daten des IMF verbindet.

7 Es ist in Ordnung, wenn Sie nicht gleich den gesamten Code nachvollziehen können; wir werden Scikit-Learn in den folgenden Kapiteln vorstellen.

Beispiel 1-1: Trainieren und Ausführen eines linearen Modells mit Scikit-Learn

```
import matplotlib.pyplot as plt
import numpy as np
import pandas as pd
import sklearn.linear_model

# Laden der Daten
oecd_bli = pd.read_csv("oecd_bli_2015.csv", thousands=',')
gdp_per_capita = pd.read_csv("gdp_per_capita.csv",thousands=',',delimiter='\t',
                             encoding='latin1', na_values="n/a")

# Vorbereiten der Daten
country_stats = prepare_country_stats(oecd_bli, gdp_per_capita)
X = np.c_[country_stats["GDP per capita"]]
y = np.c_[country_stats["Life satisfaction"]]

# Visualisieren der Daten
country_stats.plot(kind='scatter', x="Pro-Kopf-BIP", y='Zufriedenheit')
plt.show()

# Auswahl eines linearen Modells
model = sklearn.linear_model.LinearRegression()

# Trainieren des Modells
model.fit(X, y)

# Treffen einer Vorhersage für Zypern
X_new = [[22587]]  # Pro-Kopf-BIP für Zypern
print(model.predict(X_new)) # Ausgabe [[ 5.96242338]]
```

Hätten Sie einen instanzbasierten Lernalgorithmus verwendet, würden Sie herausbekommen, dass Slowenien das zu Zypern ähnlichste Pro-Kopf-BIP hat (20732 USD). Da uns die OECD-Daten die Zufriedenheit mit 5,7 angeben, hätten Sie für Zypern eine Zufriedenheit von 5,7 vorhergesagt. Wenn Sie ein wenig herauszoomen und sich die nächstgelegenen Länder ansehen, finden Sie Portugal und Spanien mit Zufriedenheitswerten von jeweils 5,1 und 6,5. Der Mittelwert dieser drei Werte ist 5,77, was sehr nah an Ihrer modellbasierten Vorhersage liegt. Dieses einfache Verfahren nennt man *k-nächste-Nachbarn*-Regression (in diesem Beispiel mit $k = 3$).

Das Ersetzen des linearen Regressionsmodells durch k-nächste-Nachbarn-Regression im obigen Code erfordert lediglich das Ersetzen dieser beiden Zeilen:

```
import sklearn.linear_model
model = sklearn.linear_model.LinearRegression()
```

durch diese zwei:

```
import sklearn.neighbors
model = sklearn.neighbors.KNeighborsRegressor(
n_neighbors=3)
```

Wenn alles gut gegangen ist, wird Ihr Modell gute Vorhersagen treffen. Wenn nicht, müssen Sie weitere Merkmale heranziehen (Beschäftigungsquote, Gesundheit, Luftverschmutzung und Ähnliches), sich mehr oder hochwertigere Trainingsdaten beschaffen oder ein mächtigeres Modell auswählen (z. B. ein polynomielles Regressionsmodell).

Zusammengefasst:

- Sie haben die Daten untersucht.
- Sie haben ein Modell ausgewählt.
- Sie haben es auf Trainingsdaten trainiert (d. h., der Trainingsalgorithmus hat nach den Modellparametern gesucht, die eine Kostenfunktion minimieren).
- Schließlich haben Sie das Modell verwendet, um für neue Fälle Vorhersagen zu treffen (dies nennt man *Inferenz*), und hoffen, dass das Modell gut verallgemeinert.

So sieht ein typisches Machine-Learning-Projekt aus. In Kapitel 2 können Sie dies selbst erfahren, indem Sie ein Projekt vom Anfang bis zum Ende durcharbeiten. Wir haben bisher ein weites Feld beschritten: Sie wissen bereits, worum es beim Machine Learning wirklich geht, warum es nützlich ist, welche Arten von ML-Systemen verbreitet sind und wie ein typischer Arbeitsablauf aussieht. Nun werden wir uns anschauen, was beim Lernen schiefgehen kann und präzise Vorhersagen verhindert.

Die wichtigsten Herausforderungen beim Machine Learning

Kurz gesagt, da Ihre Hauptaufgabe darin besteht, einen Lernalgorithmus auszuwählen und mit Daten zu trainieren, sind die zwei möglichen Fehlerquellen dabei ein »schlechter Algorithmus« sowie »schlechte Daten«. Beginnen wir mit Beispielen für schlechte Daten.

Unzureichende Menge an Trainingsdaten

Damit ein Kleinkind lernt, was ein Apfel ist, genügt es, dass Sie auf einen Apfel zeigen und »Apfel« sagen (und die Prozedur einige Male wiederholen). Damit ist das Kind in der Lage, Äpfel in allen möglichen Farben und Formen zu erkennen. Genial.

Machine Learning ist noch nicht ganz so weit; bei den meisten maschinellen Lernverfahren ist eine Vielzahl an Daten erforderlich, damit sie funktionieren. Selbst bei sehr einfachen Aufgaben benötigen Sie üblicherweise Tausende von Beispielen, und bei komplexen Aufgaben wie Bild- oder Spracherkennung können es auch Millionen sein (es sei denn, Sie können Teile eines existierenden Modells wiederverwenden).

Die unverschämte Effektivität von Daten

In einem berühmten Artikel (*https://homl.info/6*) aus dem Jahr 2001 zeigten die Forscher Michele Banko und Eric Brill bei Microsoft, dass sehr unterschiedliche maschinelle Lernalgorithmen, darunter sehr primitive, bei einem sehr komplexen Problem wie der Unterscheidung von Sprache etwa gleich gut abschnitten,[8] wenn man ihnen nur genug Daten zur Verfügung stellt (wie Sie in Abbildung 1-20 sehen können).

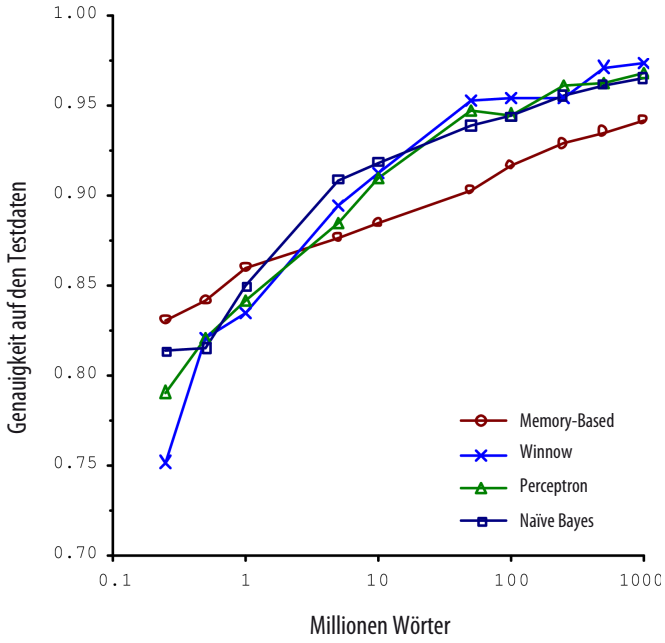


Abbildung 1-20: Die Wichtigkeit der Daten im Vergleich zum Algorithmus[9]

Die Autoren drücken dies folgendermaßen aus: »Diese Ergebnisse legen nahe, dass wir unsere Entscheidung über das Investieren von Zeit und Geld in die Entwicklung von Algorithmen gegenüber der Entwicklung eines Datenkorpus neu bewerten sollten.«

Dass Daten bei komplexen Problemen wichtiger als Algorithmen sind, wurde von Peter Norvig et al. in einem Artikel mit dem Titel »The Unreasonable Effectiveness of Data« (*https://homl.info/7*), veröffentlicht im Jahr 2009, weiter thematisiert.[10] Es

8 Beispielsweise aus dem Kontext zu ermitteln, ob man »to«, »two« oder »too« schreiben muss.

9 Die Abbildung wurde mit Erlaubnis von Michele Banko und Eric Brill reproduziert aus »Scaling to Very Very Large Corpora for Natural Language Disambiguation« Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (2001), 26–33.

10 Peter Norvig et al., »The Unreasonable Effectiveness of Data«, *IEEE Intelligent Systems* 24, no. 2 (2009): 8–12.

sollte jedoch betont werden, dass kleine und mittelgroße Datensätze nach wie vor sehr häufig sind und dass es nicht immer einfach oder billig ist, an zusätzliche Trainingsdaten heranzukommen. Daher schreiben Sie die Algorithmik besser nicht gleich ab.

Nicht repräsentative Trainingsdaten

Um gut zu verallgemeinern, ist es entscheidend, dass Ihre Trainingsdaten die zu verallgemeinernden neuen Situationen repräsentieren. Dies ist sowohl beim instanzbasierten als auch beim modellbasierten Lernen der Fall.

Beispielsweise waren die zuvor zum Trainieren eines linearen Modells eingesetzten Länder nicht perfekt repräsentativ; einige Länder fehlten. Abbildung 1-21 zeigt, wie die Daten mit den fehlenden Ländern aussehen.

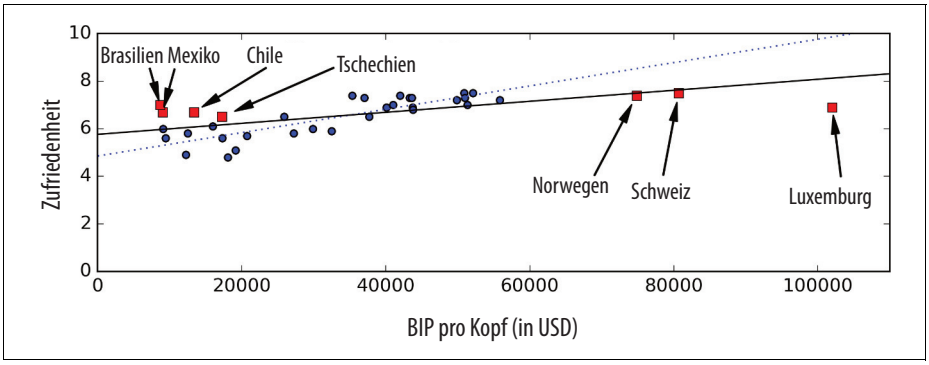


Abbildung 1-21: Ein repräsentativerer Trainingsdatensatz

Wenn Sie mit diesen Daten ein lineares Modell trainieren, erhalten Sie die durchgezogene Linie. Das alte Modell ist durch die gepunktete Linie gekennzeichnet. Wie Sie sehen, verändert sich nicht nur das Modell durch die Daten. Es wird auch deutlich, dass ein einfaches lineares Modell vermutlich nie gut funktionieren wird. Es sieht ganz danach aus, dass reiche Länder nicht glücklicher als Länder mit mittlerem Wohlstand sind (sie wirken sogar weniger glücklich). Auch einige arme Länder scheinen glücklicher als viele reiche Länder zu sein.

Das mit einem nicht repräsentativen Datensatz trainierte Modell trifft also ungenaue Vorhersagen, besonders bei sehr armen und sehr reichen Ländern.

Es ist daher wichtig, einen Trainingsdatensatz zu verwenden, in dem die zu verallgemeinernden Fälle abgebildet sind. Oft ist dies schwieriger, als es sich anhört: Wenn die Stichprobe zu klein ist, erhalten Sie *Stichprobenrauschen* (also durch Zufall nicht repräsentative Daten). Selbst sehr große Stichproben können nicht repräsentativ sein, wenn die Methode zur Erhebung fehlerhaft ist. Dies nennt man auch *Stichprobenverzerrung*.

Beispiele für Stichprobenverzerrungen

Das vermutlich berühmteste Beispiel für Stichprobenverzerrung stammt aus der US-Präsidentschaftswahl von 1936, bei der Landon gegen Roosevelt antrat: Der *Literary Digest* führte damals eine sehr große Umfrage durch, bei der Briefe an etwa 10 Millionen Menschen verschickt wurden. Nach dem Sammeln von 2,4 Millionen Antworten wurde daraus mit hoher Konfidenz vorhergesagt, dass Landon 57 % der Stimmen erhalten würde. Tatsächlich gewann aber Roosevelt mit 62 % der Stimmen. Der Fehler lag in der Methode, die *Literary Digest* beim Erheben der Stichprobe einsetzte:

- Zum einen verwendete *Literary Digest* Telefonbücher, Abonnentenlisten, Mitgliederlisten von Klubs und so weiter, um an die Adressen zum Verschicken der Umfrage zu kommen. In allen diesen Listen waren wohlhabendere Menschen stärker vertreten, die wahrscheinlich eher für die Republikaner (und damit Landon) stimmen würden.
- Zum anderen antworteten weniger als 25 % der Menschen auf die Umfrage. Auch dies führte zu einer Stichprobenverzerrung, da Menschen, die sich nicht für Politik interessieren, Menschen, die den *Literary Digest* nicht mögen, und andere wichtige Gruppen potenziell aussortiert wurden. Diese Art von Stichprobenverzerrung nennt man auch *Schweigeverzerrung*.

Ein weiteres Beispiel: Sagen wir, Sie möchten ein System zum Erkennen von Funk-Musikvideos konstruieren. Eine Möglichkeit zum Zusammenstellen der Trainingsdaten wäre, »funk music« bei YouTube einzugeben und die erhaltenen Videos zu verwenden. Allerdings nehmen Sie dabei an, dass Ihnen die Suchmaschine von YouTube Videos liefert, die repräsentativ für alle Funk-Musikvideos auf YouTube sind. In der Realität werden einige beliebte Künstler in den Suchergebnissen jedoch überrepräsentiert sein (und wenn Sie in Brasilien leben, erhalten Sie eine Menge Videos zu »funk carioca«, die sich überhaupt nicht wie James Brown anhören). Wie aber sonst sollte man einen großen Datensatz sammeln?

Minderwertige Daten

Wenn Ihre Trainingsdaten voller Fehler, Ausreißer und Rauschen sind (z. B. wegen schlechter Messungen), ist es für das System schwieriger, die zugrunde liegenden Muster zu erkennen. Damit ist es weniger wahrscheinlich, dass Ihr System eine hohe Leistung erzielt. Meistens lohnt es sich, Zeit in das Säubern der Trainingsdaten zu investieren. Tatsächlich verbringen die meisten Data Scientists einen Großteil ihrer Zeit mit nichts anderem, beispielsweise:

- Wenn einige Datenpunkte deutliche Ausreißer sind, hilft es, diese einfach zu entfernen oder die Fehler manuell zu beheben.
- Wenn manche Merkmale lückenhaft sind (z. B. 5 % Ihrer Kunden ihr Alter nicht angegeben haben), müssen Sie sich entscheiden, ob Sie dieses Merkmal insgesamt ignorieren wollen oder die entsprechenden Datenpunkte entfernen,

die fehlenden Werte ergänzen (z.B. mit dem Median) oder ein Modell mit diesem Merkmal und eines ohne dieses Merkmal trainieren möchten.

Irrelevante Merkmale

Eine Redewendung besagt: Müll rein, Müll raus. Ihr System wird nur etwas erlernen können, wenn Ihre Trainingsdaten genug relevante Merkmale und nicht zu viele irrelevante enthalten. Ein für den Erfolg eines Machine-Learning-Projekts maßgeblicher Schritt ist, die Merkmale zum Trainieren gut auszuwählen. Zu diesem *Entwicklung von Merkmalen* genannten Vorgang gehören:

- *Auswahl von Merkmalen* (aus den vorhandenen die nützlichsten Merkmale für das Trainieren auswählen).
- *Extraktion von Merkmalen* (vorhandene Merkmale miteinander kombinieren, sodass ein nützlicheres entsteht – wie wir oben gesehen haben, helfen dabei Algorithmen zur Dimensionsreduktion).
- Erstellen neuer Merkmale durch das Erheben neuer Daten.

Nun, da wir viele Beispiele für schlechte Daten kennengelernt haben, schauen wir uns auch noch einige schlechte Algorithmen an.

Overfitting der Trainingsdaten

Sagen wir, Sie sind im Ausland unterwegs, und der Taxifahrer zockt Sie ab. Sie mögen versucht sein, zu sagen, dass *alle* Taxifahrer in diesem Land Betrüger seien. Menschen neigen häufig zu übermäßiger Verallgemeinerung, und Maschinen können leider in die gleiche Falle tappen, wenn wir nicht vorsichtig sind. Beim Machine Learning nennt man dies *Overfitting*: Dabei funktioniert das Modell auf den Trainingsdaten, kann aber nicht gut verallgemeinern.

Abbildung 1-22 zeigt ein Beispiel für ein polynomielles Modell höheren Grades für die Zufriedenheit, das die Trainingsdaten stark overfittet. Es erzielt zwar auf den Trainingsdaten eine höhere Genauigkeit als das einfache lineare Modell, aber würden Sie den Vorhersagen dieses Modells wirklich trauen?

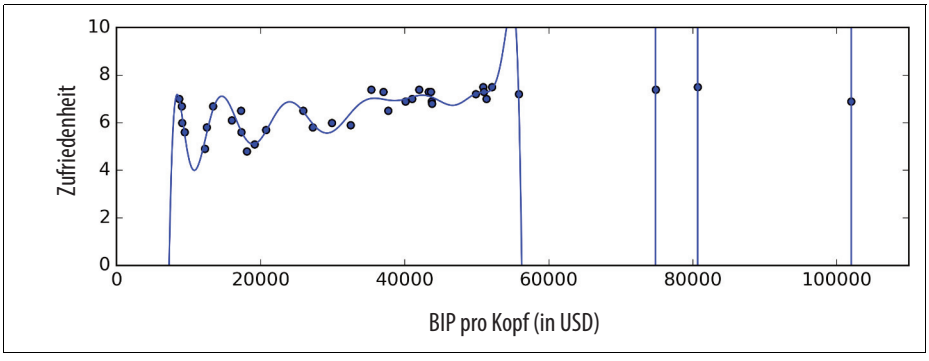


Abbildung 1-22: Overfitting der Trainingsdaten

Komplexe Modelle wie Deep-Learning-Netze können subtile Muster in den Daten erkennen. Wenn aber der Trainingsdatensatz verrauscht oder zu klein ist (wodurch die Stichprobe verrauscht ist), entdeckt das Modell Muster im Rauschen selbst. Diese Muster lassen sich natürlich nicht auf neue Daten übertragen. Nehmen wir beispielsweise an, Sie stellten Ihrem Modell für die Zufriedenheit viele weitere Merkmale zur Verfügung, darunter wenig informative wie den Namen des Lands. Ein komplexes Modell könnte dann herausfinden, dass alle Länder mit einer Zufriedenheit über 7 ein *w* im Namen haben: New Zealand (7,3), Norway (7,4), Sweden (7,2) und Switzerland (7,5). Wie sicher können Sie sich sein, dass diese *w*-Regel sich auf Rwanda oder Zimbabwe anwenden lässt? Natürlich trat dieses Muster in den Trainingsdaten rein zufällig auf, aber das Modell ist nicht in der Lage, zu entscheiden, ob ein Muster echt oder durch das Rauschen in den Daten bedingt ist.

Overfitting tritt auf, wenn das Modell angesichts der Menge an Trainingsdaten und der Menge an Rauschen zu komplex ist. Mögliche Lösungen sind:

- das Modell zu vereinfachen, indem man es durch eines mit weniger Parametern ersetzt (z.B. ein lineares Modell statt eines polynomiellen Modells höheren Grades), die Anzahl der Merkmale im Trainingsdatensatz verringert oder dem Modell Restriktionen auferlegt,
- mehr Trainingsdaten zu sammeln
- oder das Rauschen in den Trainingsdaten zu reduzieren (z.B. Datenfehler zu beheben und Ausreißer zu entfernen).

Einem Modell Restriktionen aufzuerlegen, um es zu vereinfachen und das Risiko für Overfitting zu reduzieren, wird als *Regularisierung* bezeichnet. Beispielsweise hat das oben definierte lineare Modell zwei Parameter, θ_0 und θ_1. Damit hat der Lernalgorithmus zwei *Freiheitsgrade*, mit denen das Modell an die Trainingsdaten angepasst werden kann: Sowohl die Höhe (θ_0) als auch die Steigung (θ_1) der Geraden lassen sich verändern. Wenn wir $\theta_1 = 0$ erzwingen würden, hätte der Algorithmus nur noch einen Freiheitsgrad, und es würde viel schwieriger, die Daten gut zu fitten: Die Gerade könnte sich nur noch nach oben oder unten bewegen, um so nah wie möglich an den Trainingsdatenpunkten zu landen. Sie würde daher in der Nähe des Mittelwerts landen. Wirklich ein sehr einfaches Modell! Wenn wir dem Modell erlauben, θ_1 zu verändern, aber einen kleinen Wert erzwingen, hat der Lernalgorithmus zwischen einem und zwei Freiheitsgrade. Das entstehende Modell ist einfacher als das mit zwei Freiheitsgraden, aber komplexer als das mit nur einem. Ihre Aufgabe ist es, die richtige Balance zwischen dem perfekten Fitten der Daten und einem möglichst einfachen Modell zu finden, sodass es gut verallgemeinert.

Abbildung 1-23 zeigt drei Modelle: Die gepunktete Linie steht für das ursprüngliche mit den als Kreis dargestellten Ländern trainierte Modell (ohne die als Quadrat dargestellten Länder), die gestrichelte Linie für unser zweites mit allen Ländern (Kreise und Quadrate) trainiertes Modell, und die durchgezogene Linie ist ein

Modell, das mit den gleichen Daten wie das erste Modell trainiert wurde, aber mit zusätzlicher Regularisierung. Sie sehen, dass die Regularisierung eine geringere Steigung erzwungen hat. Dieses Modell passt nicht so gut zu den Trainingsdaten wie das erste Modell, erlaubt aber eine bessere Verallgemeinerung auf neue Beispiele, die es während des Trainings nicht kannte (Quadrate).

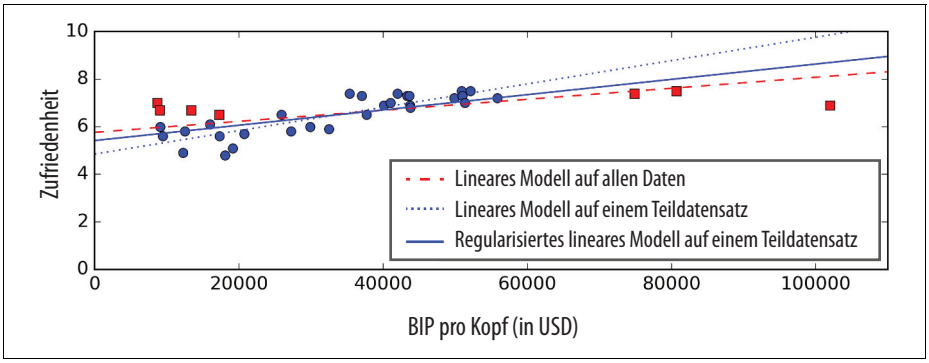


Abbildung 1-23: Regularisierung reduziert das Risiko für Overfitting.

Die Stärke der Regularisierung beim Lernen lässt sich über einen *Hyperparameter* kontrollieren. Ein Hyperparameter ist ein Parameter des Lernalgorithmus (nicht des Modells). Als solcher unterliegt er nicht dem Lernprozess selbst; er muss vor dem Training gesetzt werden und bleibt über den gesamten Trainingszeitraum konstant. Wenn Sie den Hyperparameter zur Regularisierung auf einen sehr großen Wert setzen, erhalten Sie ein beinahe flaches Modell (eine Steigung nahe null); Sie können sich sicher sein, dass der Lernalgorithmus die Trainingsdaten nicht overfittet, eine gute Lösung wird aber ebenfalls unwahrscheinlicher. Die Feineinstellung der Hyperparameter ist ein wichtiger Teil bei der Entwicklung eines Machine-Learning-Systems (im nächsten Kapitel lernen Sie ein detailliertes Beispiel kennen).

Underfitting der Trainingsdaten

Wie Sie sich denken können, ist *Underfitting* das genaue Gegenteil von Overfitting: Es tritt auf, wenn Ihr Modell zu einfach ist, um die in den Daten enthaltene Struktur zu erlernen. Beispielsweise ist ein lineares Modell der Zufriedenheit anfällig für Underfitting; die Realität ist einfach komplexer als unser Modell, sodass Vorhersagen selbst auf den Trainingsdaten zwangsläufig ungenau werden.

Die wichtigsten Möglichkeiten, dieses Problem zu beheben, sind:

- ein mächtigeres Modell mit mehr Parametern zu verwenden,
- dem Lernalgorithmus bessere Merkmale zur Verfügung zu stellen (Entwicklung von Merkmalen) oder
- die Restriktionen des Modells zu verringern (z.B. die Hyperparameter zur Regularisierung zu verringern).

Zusammenfassung

Inzwischen wissen Sie schon eine Menge über Machine Learning. Wir haben aber so viele Begriffe behandelt, dass Sie sich vielleicht ein wenig verloren vorkommen. Betrachten wir deshalb noch einmal das Gesamtbild:

- Beim Machine Learning geht es darum, Maschinen bei der Lösung einer Aufgabe zu verbessern, indem sie aus Daten lernen, anstatt explizit definierte Regeln zu erhalten.
- Es gibt viele unterschiedliche Arten von ML-Systemen: überwachte und unüberwachte, Batch- und Online-Learning, instanzbasierte und modellbasierte Systeme.
- In einem ML-Projekt sammeln Sie Daten in einem Trainingsdatensatz und speisen diesen in einen Lernalgorithmus ein. Wenn der Algorithmus auf einem Modell basiert, tunt er einige Parameter, um das Modell an die Trainingsdaten anzupassen (d.h., um gute Vorhersagen auf den Trainingsdaten selbst zu treffen). Danach ist es hoffentlich in der Lage, auch für neue Daten gute Vorhersagen zu treffen. Wenn der Algorithmus instanzbasiert ist, lernt er die Beispiele einfach auswendig und verallgemeinert auf neue Instanzen durch ein Ähnlichkeitsmaß, um sie mit den bekannten Instanzen zu vergleichen.
- Wenn der Trainingsdatensatz zu klein ist oder die Daten nicht repräsentativ, verrauscht oder durch irrelevante Merkmale verunreinigt sind, wird das System keine hohe Leistung erbringen (Müll rein, Müll raus). Schließlich darf Ihr Modell weder zu einfach (dann underfittet es) noch zu komplex sein (dann overfittet es).

Es gilt noch ein letztes wichtiges Thema zu behandeln: Sobald Sie ein Modell trainiert haben, sollten Sie nicht nur »hoffen«, dass es gut verallgemeinert, Sie sollten es auch evaluieren und, falls nötig, Feinabstimmungen vornehmen. Sehen wir einmal, wie das geht.

Testen und Validieren

Ein Modell mit neuen Datenpunkten auszuprobieren, ist tatsächlich die einzige Möglichkeit, zu erfahren, ob es gut auf neue Daten verallgemeinert. Sie können das Modell dazu in einem Produktivsystem einsetzen und beobachten, wie es funktioniert. Wenn sich Ihr Modell aber als definitiv untauglich herausstellt, werden sich Ihre Nutzer beschweren – nicht die beste Idee.

Eine bessere Alternative ist, Ihre Daten in zwei Datensätze zu unterteilen: den *Trainingsdatensatz* und den *Testdatensatz*. Wie die Namen vermuten lassen, trainieren Sie Ihr Modell mit dem Trainingsdatensatz und testen es mit dem Testdatensatz. Die Abweichung bei neuen Datenpunkten bezeichnet man als *Verallgemeinerungsfehler* (engl. *out-of-sample error*). Indem Sie Ihr Modell auf dem Testdatensatz eva-

luieren, erhalten Sie eine Schätzung dieser Abweichung. Der Wert verrät Ihnen, wie gut Ihr Modell auf zuvor nicht bekannten Datenpunkten abschneidet.

Wenn der Fehler beim Training gering ist (Ihr Modell also auf dem Trainingsdatensatz wenige Fehler begeht), der Verallgemeinerungsfehler aber groß, overfittet Ihr Modell die Trainingsdaten.

Es ist üblich, 80 % der Daten zum Trainieren zu nehmen und 20 % zum Testen *zurückzuhalten*. Das hängt aber auch von der Größe des Datensatzes ab: Enthält er 10 Millionen Instanzen, bedeutet ein Zurückhalten von 1 %, dass Ihr Testdatensatz aus 100.000 Instanzen besteht – vermutlich mehr als genug für eine gute Abschätzung des Verallgemeinerungsfehlers.

Hyperparameter anpassen und Modellauswahl

Das Evaluieren eines Modells ist einfach: Verwenden Sie einen Testdatensatz. Aber möglicherweise schwanken Sie zwischen zwei Typen von Modellen (z. B. einem linearen und einem polynomiellen Modell): Wie sollen Sie sich zwischen ihnen entscheiden? Sie können natürlich beide trainieren und mit dem Testdatensatz vergleichen, wie gut beide verallgemeinern.

Nehmen wir an, das lineare Modell verallgemeinert besser, aber Sie möchten durch Regularisierung dem Overfitting entgegenwirken. Die Frage ist: Wie wählen Sie den Wert des Regularisierungshyperparameters aus? Natürlich können Sie 100 unterschiedliche Modelle mit 100 unterschiedlichen Werten für diesen Hyperparameter trainieren. Angenommen, Sie finden einen optimalen Wert für den Hyperparameter, der den niedrigsten Verallgemeinerungsfehler liefert, z. B. 5 %. Sie bringen Ihr Modell daher in die Produktivumgebung, aber leider schneidet es nicht wie erwartet ab und produziert einen Fehler von 15 %. Was ist passiert?

Das Problem ist, dass Sie den Verallgemeinerungsfehler mithilfe des Testdatensatzes mehrfach bestimmt und das Modell und seine Hyperparameter so angepasst haben, dass es *für diesen Datensatz* das beste Modell hervorbringt. Damit erbringt das Modell bei neuen Daten wahrscheinlich keine gute Leistung.

Eine übliche Lösung dieses Problems nennt sich *Hold-out-Validierung*: Sie halten einfach einen Teil des Trainingsdatensatzes zurück, um die verschiedenen Modellkandidaten zu bewerten, und wählen den besten aus. Dieser neue zurückgehaltene Datensatz wird als *Validierungsdatensatz* (oder manchmal auch als *Entwicklungsdatensatz* oder *Dev Set*) bezeichnet. Genauer gesagt, trainieren Sie mehrere Modelle mit unterschiedlichen Hyperparametern auf dem Trainingsdatensatz und wählen das auf dem Validierungsdatensatz am besten abschneidende Modell und die Hyperparameter aus. Nach diesem Hold-out-Validierungsprozess trainieren Sie das beste Modell mit dem vollständigen Trainingsdatensatz (einschließlich des Validierungsdatensatzes) und erhalten so das endgültige Modell. Schließlich füh-

ren Sie einen einzelnen abschließenden Test an diesem finalen Modell mit dem Testdatensatz durch, um den Verallgemeinerungsfehler abzuschätzen.

Dieses Vorgehen funktioniert meist ziemlich gut. Ist aber der Validierungsdatensatz zu klein, wird die Modellbewertung ungenau sein – Sie landen eventuell unabsichtlich bei einem suboptimalen Modell. Ist der Validierungsdatensatz hingegen zu groß, wird der verbleibende Trainingsdatensatz viel kleiner sein als der vollständige Trainingsdatensatz. Warum ist das schlecht? Nun, da das finale Modell mit dem vollständigen Trainingsdatensatz trainiert werden wird, ist es nicht ideal, Modellkandidaten zu vergleichen, die mit einem viel kleineren Trainingsdatensatz trainiert wurden. Das wäre so, als würden Sie den schnellsten Sprinter auswählen, damit er an einem Marathon teilnimmt. Eine Möglichkeit, dieses Problem zu lösen, ist eine wiederholt durchgeführte *Kreuzvalidierung* mit vielen kleinen Validierungsdatensätzen. Jedes Modell wird einmal per Validierungsdatensatz geprüft, nachdem es mit dem Rest der Daten trainiert wurde. Durch ein Mitteln aller Bewertungen eines Modells können Sie dessen Genauigkeit deutlich besser bestimmen. Es gibt aber einen Nachteil: Die Trainingsdauer wird mit der Anzahl der Validierungsdatensätze multipliziert.

Datendiskrepanz

In manchen Fällen kommt man zwar leicht an große Mengen an Trainingsdaten, diese sind aber keine perfekte Repräsentation der im Produktivumfeld verwendeten Daten. Stellen Sie sich zum Beispiel vor, Sie wollten eine Mobil-App bauen, die Bilder von Blumen aufnimmt und automatisch deren Spezies bestimmt. Sie können problemlos Millionen von Blumenbildern aus dem Web laden, aber sie werden nicht perfekt die Bilder repräsentieren, die tatsächlich mit der App auf einem Smartphone aufgenommen werden. Vielleicht haben Sie nur 10.000 repräsentative Bilder (also solche, die mit der App aufgenommen wurden). In diesem Fall ist es am wichtigsten, daran zu denken, dass der Validierungsdatensatz und der Testdatensatz in Bezug auf die produktiv genutzten Daten so repräsentativ wie möglich sein müssen, daher sollten sie nur aus repräsentativen Bildern bestehen: Sie können sie mischen, die eine Hälfte in den Validierungsdatensatz und die andere in den Testdatensatz stecken (wobei Sie darauf achten müssen, keine Dubletten oder Nahezu-Dubletten in beiden Sets zu haben). Aber beobachten Sie nach dem Trainieren Ihres Modells mit den Bildern aus dem Web, dass die Genauigkeit des Modells beim Validierungsdatensatz enttäuschend ist, werden Sie nicht wissen, ob das daran liegt, dass das Modell bezüglich des Trainingsdatensatzes overfittet oder ob es einfach einen Unterschied zwischen den Bildern aus dem Web und denen aus der App gibt. Eine Lösung ist, einen Teil der Trainingsbilder (aus dem Web) in einen weiteren Datensatz zu stecken, den Andrew Ng als *Train-Dev-Set* bezeichnet. Nachdem das Modell trainiert wurde (mit dem Trainingsdatensatz, *nicht* mit dem Train-Dev-Set), können Sie es auf das Train-Dev-Set loslassen. Funktioniert es dort gut, hat das Modell kein Overfitting bezüglich des Trainingsdatensatzes. Ist die Genauig-

keit beim Validierungsdatensatz schlecht, muss das Problem aus der Datendiskrepanz entstanden sein. Sie können versuchen, die Lösung des Problems durch eine Vorverarbeitung der Webbilder anzugehen, sodass sie mehr wie die Bilder aus der App aussehen, und dann das Modell neu zu trainieren. Funktioniert das Modell hingegen schlecht mit dem Train-Dev-Set, muss es bezüglich des Trainingsdatensatzes overfittet sein, und Sie sollten versuchen, das Modell zu vereinfachen oder zu regularisieren sowie mehr Trainingsdaten zu erhalten und diese zu säubern.

Das No-Free-Lunch-Theorem

Ein Modell ist eine vereinfachte Version der Beobachtungen. Die Vereinfachung soll überflüssige Details eliminieren, die sich vermutlich nicht auf neue Datenpunkte verallgemeinern lassen. Um allerdings zu entscheiden, welche Daten zu verwerfen und welche beizubehalten sind, müssen Sie *Annahmen* treffen. Beispielsweise trifft ein lineares Modell die Annahme, dass die Daten grundsätzlich linear sind und die Distanz zwischen den Datenpunkten und der Geraden lediglich Rauschen ist, das man ignorieren kann.

In einem berühmten Artikel aus dem Jahr 1996 (*https://homl.info/8*)[11] zeigte David Wolpert, dass es keinen Grund gibt, ein Modell gegenüber einem anderen zu bevorzugen, wenn Sie absolut keine Annahmen über die Daten treffen. Dies nennt man auch das *No-Free-Lunch*-(NFL-)Theorem. Bei einigen Datensätzen ist das optimale Modell ein lineares Modell, während bei anderen ein neuronales Netz am besten geeignet ist. Es gibt kein Modell, das garantiert *a priori* besser funktioniert (daher der Name des Theorems). Der einzige Weg, wirklich sicherzugehen, ist, alle möglichen Modelle zu evaluieren. Da dies nicht möglich ist, treffen Sie in der Praxis einige wohlüberlegte Annahmen über die Daten und evaluieren nur einige sinnvoll ausgewählte Modelle. Bei einfachen Aufgaben könnten Sie beispielsweise lineare Modelle mit unterschiedlich starker Regularisierung auswerten, bei einer komplexen Aufgabe hingegen verschiedene neuronale Netze.

Übungen

In diesem Kapitel haben wir einige der wichtigsten Begriffe zum Machine Learning behandelt. In den folgenden Kapiteln werden wir uns eingehender damit beschäftigen und mehr Code schreiben, aber zuvor sollten Sie sicherstellen, dass Sie die folgenden Fragen beantworten können:

1. Wie würden Sie Machine Learning definieren?
2. Können Sie vier Arten von Aufgaben nennen, für die Machine Learning gut geeignet ist?

11 David Wolpert, »The Lack of A Priori Distinctions Between Learning Algorithms«, *Neural Computation* 8 (1996): 1341–1390.

3. Was ist ein gelabelter Trainingsdatensatz?
4. Was sind die zwei verbreitetsten Aufgaben beim überwachten Lernen?
5. Können Sie vier häufig anzutreffende Aufgaben für unüberwachtes Lernen nennen?
6. Was für einen Machine-Learning-Algorithmus würden Sie verwenden, um einen Roboter über verschiedene unbekannte Oberflächen laufen zu lassen?
7. Welche Art Algorithmus würden Sie verwenden, um Ihre Kunden in unterschiedliche Gruppen einzuteilen?
8. Würden Sie die Aufgabe, Spam zu erkennen, als überwachte oder unüberwachte Lernaufgabe einstufen?
9. Was ist ein Onlinelernsystem?
10. Was ist Out-of-Core-Lernen?
11. Welche Art Lernalgorithmus beruht auf einem Ähnlichkeitsmaß, um Vorhersagen zu treffen?
12. Worin besteht der Unterschied zwischen einem Modellparameter und einem Hyperparameter eines Lernalgorithmus?
13. Wonach suchen modellbasierte Lernalgorithmen? Welche Strategie führt am häufigsten zum Erfolg? Wie treffen sie Vorhersagen?
14. Können Sie vier der wichtigsten Herausforderungen beim Machine Learning benennen?
15. Welches Problem liegt vor, wenn Ihr Modell auf den Trainingsdaten eine sehr gute Leistung erbringt, aber schlecht auf neue Daten verallgemeinert? Nennen Sie drei Lösungsansätze.
16. Was ist ein Testdatensatz, und warum sollte man einen verwenden?
17. Was ist der Zweck eines Validierungsdatensatzes?
18. Was ist das Train-Dev-Set, wann brauchen Sie es, und wie verwenden Sie es?
19. Was kann schiefgehen, wenn Sie Hyperparameter mithilfe der Testdaten einstellen?

Lösungen zu diesen Übungsaufgaben finden Sie in Anhang A.

KAPITEL 2

Ein Machine-Learning-Projekt von A bis Z

In diesem Kapitel werden Sie ein Beispielprojekt vom Anfang bis zum Ende durchleben. Wir nehmen dazu an, Sie seien ein frisch angeheuerter Data Scientist in einer Immobilienfirma.[1] Dies sind die wichtigsten Schritte, die es zu bearbeiten gilt:

1. Betrachte das Gesamtbild.
2. Beschaffe die Daten.
3. Erkunde und visualisiere die Daten, um daraus Erkenntnisse zu gewinnen.
4. Bereite die Daten für Machine-Learning-Algorithmen vor.
5. Wähle ein Modell aus und trainiere es.
6. Verfeinere das Modell.
7. Präsentiere die Lösung.
8. Nimm das System in Betrieb, beobachte und warte es.

Der Umgang mit realen Daten

Wenn Sie Machine Learning gerade erst erlernen, experimentieren Sie am besten mit realen Daten, nicht mit künstlich generierten Datensätzen. Glücklicherweise stehen Tausende frei verfügbarer Datensätze aus allen möglichen Fachgebieten zur Auswahl. Hier sind einige Quellen, unter denen Sie passende Daten finden können:

- Beliebte Archive mit frei verfügbaren Daten:
  - UC Irvine Machine Learning Repository (*http://archive.ics.uci.edu/ml/*)
  - Datensätze von Kaggle (*https://www.kaggle.com/datasets*)
  - Datensätze von Amazon AWS (*https://registry.opendata.aws/*)

1 Dieses Beispielprojekt ist frei erfunden; das Ziel ist, die wichtigsten Schritte eines Machine-Learning-Projekts zu illustrieren, nicht etwas über den Immobilienmarkt zu erfahren.

- Metaseiten (Listen von Datenarchiven):
  - Data Portals (*http://dataportals.org/*)
  - OpenDataMonitor (*http://opendatamonitor.eu/*)
  - Quandl (*http://quandl.com/*)
- Andere Seiten, die beliebte offene Datenarchive auflisten:
  - Die Wikipedia-Seite mit Machine-Learning-Datensätzen (*https://homl.info/9*)
  - Quora.com (*https://homl.info/10*)
  - subreddit zu Datensätzen (*https://www.reddit.com/r/datasets*)

Für dieses Kapitel suchen wir uns einen Datensatz zu Immobilienpreisen in Kalifornien aus dem StatLib Repository aus (siehe Abbildung 2-1).[2] Dieser Datensatz basiert auf Informationen aus der kalifornischen Volkszählung von 1990. Er ist nicht gerade aktuell (in der San Francisco Bay konnte man sich damals noch ein nettes Häuschen leisten), bietet aber viele Eigenschaften, anhand deren sich gut lernen lässt. Wir werden deshalb so tun, als wären die Daten aktuell. Wir haben aus didaktischen Gründen auch ein zusätzliches Merkmal hinzugefügt und einige Merkmale entfernt.

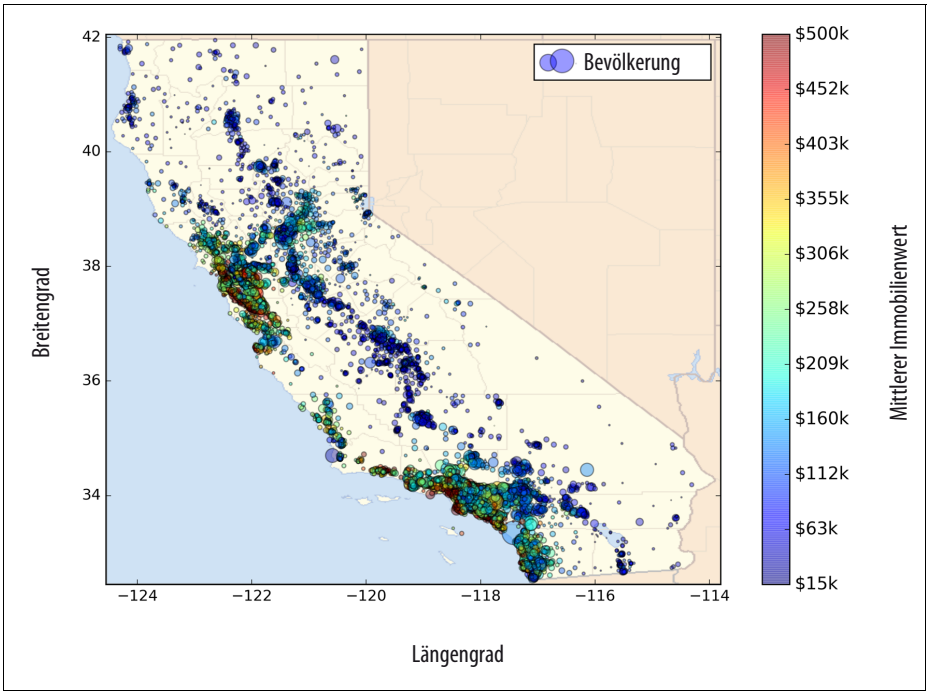


Abbildung 2-1: Immobilienpreise in Kalifornien

2 Der ursprüngliche Datensatz erschien in: R. Kelley Pace and Ronald Barry, »Sparse Spatial Autoregressions«, *Statistics & Probability Letters* 33, no. 3 (1997): 291–297.

Betrachte das Gesamtbild

Willkommen beim Machine-Learning-Immobilienkonzern! Ihre erste Aufgabe ist, ein Modell der Immobilienpreise in Kalifornien mithilfe der Daten aus der kalifornischen Volkszählung zu erstellen. Diese Daten enthalten für jede Blockgruppe in Kalifornien Metriken wie die Bevölkerung, das mittlere Einkommen, die mittleren Immobilienpreise und so weiter. Blockgruppen sind die kleinste geografische Einheit, für die das US Census Bureau Daten veröffentlicht (eine Blockgruppe hat typischerweise eine Bevölkerung von 600 bis 3.000 Menschen). Wir werden diese einfach »Bezirke« nennen.

Ihr Modell sollte aus diesen Daten lernen und in der Lage sein, den mittleren Immobilienpreis in einem beliebigen Bezirk aus allen übrigen Metriken vorherzusagen.

Da Sie ein gut organisierter Data Scientist sein möchten, ist Ihr erster Arbeitsschritt, Ihre Checkliste für Machine-Learning-Projekte zu zücken. Sie können mit der Liste in Anhang B beginnen; diese sollte für die meisten Machine-Learning-Projekte annehmbar funktionieren. Sie sollten sie aber an Ihre Bedürfnisse anpassen. In diesem Kapitel werden wir viele Punkte dieser Checkliste abarbeiten, aber auch einige selbsterklärende oder in späteren Kapiteln besprochene überspringen.

Die Aufgabe abstecken

Die allererste Frage an Ihren Vorgesetzten ist, was denn eigentlich das Geschäftsziel sei; ein Modell zu erstellen, ist vermutlich nicht das eigentliche Ziel. In welcher Weise möchte Ihre Firma das Modell voraussichtlich nutzen und davon profitieren? Von der Antwort hängt ab, wie Sie Ihre Aufgabenstellung formulieren, welche Algorithmen Sie auswählen, welches Qualitätsmaß Sie zum Auswerten des Modells verwenden und wie viel Aufwand Sie in die Optimierung stecken sollten.

Ihr Vorgesetzter antwortet, dass die Ausgabe Ihres Modells (eine Vorhersage des mittleren Immobilienpreises eines Bezirks) zusammen mit anderen *Signalen*[3] in ein anderes Machine-Learning-System eingespeist werden soll (siehe Abbildung 2-2). Dieses nachgeschaltete System soll entscheiden, ob sich Investitionen in einer bestimmten Gegend lohnen oder nicht. Die Richtigkeit dieser Entscheidungen wirkt sich direkt auf die Einnahmen aus.

3 Eine in ein Machine-Learning-System eingegebene Information wird oft nach Claude Shannons Informationstheorie, die er an den Bell Labs entwickelt hat, um die Telekommunikation zu verbessern, als *Signal* bezeichnet: Ein hoher Quotient von Signal und Hintergrundrauschen ist wünschenswert.

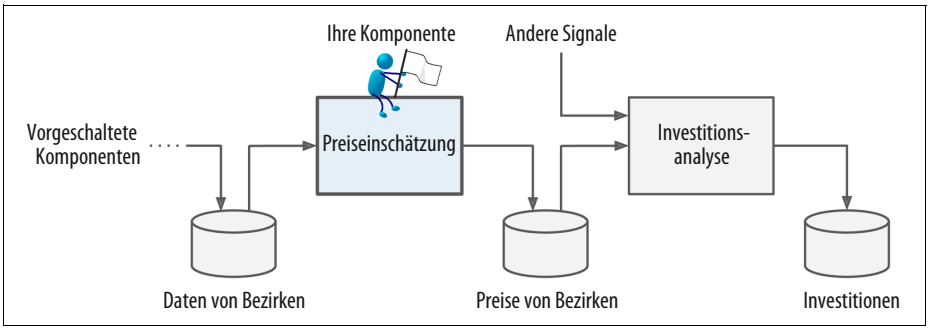


Abbildung 2-2: Eine Machine-Learning-Pipeline für Immobilieninvestitionen

Pipelines

Eine Abfolge von *Komponenten* zur Datenverarbeitung nennt man eine *Pipeline*. Pipelines sind in Machine-Learning-Systemen sehr häufig, weil dabei eine Menge Daten zu bearbeiten und viele Datentransformationen anzuwenden sind.

Diese Komponenten werden üblicherweise asynchron ausgeführt. Jede Komponente liest eine große Datenmenge ein, verarbeitet sie und schiebt die Ergebnisse in einen anderen Datenspeicher. Etwas später liest die nächste Komponente der Pipeline diese Daten ein, produziert ihre eigene Ausgabe und so weiter. Jede Komponente ist einigermaßen eigenständig: Als Schnittstelle zwischen den Komponenten dient der Datenspeicher. Dadurch ist das System recht einfach zu erfassen (mithilfe eines Datenflussdiagramms), und mehrere Teams können sich auf unterschiedliche Komponenten konzentrieren. Wenn außerdem eine Komponente ausfällt, können die nachgeschalteten Komponenten oft normal weiterarbeiten (zumindest für eine Weile), indem sie einfach die letzte Ausgabe der ausgefallenen Komponente verwenden. Dadurch ist diese Architektur recht robust.

Andererseits kann eine ausgefallene Komponente eine ganze Weile unbemerkt bleiben, falls das System nicht angemessen überwacht wird. Die Daten veralten dann, und die Leistung des Gesamtsystems sinkt.

Die nächste Frage, die Sie Ihrem Chef stellen sollten, ist, was für eine Lösung bereits verwendet wird (falls überhaupt). Häufig erhalten Sie dabei einen Referenzwert für die Leistung sowie Hinweise zum Lösen der Aufgabe. Ihr Vorgesetzter antwortet Ihnen, dass die Immobilienpreise der Bezirke im Moment von Experten manuell geschätzt werden: Ein Team sammelt aktuelle Informationen über einen Bezirk, und wenn es den mittleren Immobilienpreis nicht ermitteln kann, werden diese mithilfe komplexer Regeln geschätzt.

Dieses Verfahren ist sowohl kosten- als auch zeitintensiv, und die Schätzungen sind nicht besonders gut; wenn ein Team den mittleren Immobilienpreis herausfindet, stellt es häufig fest, dass es mit seiner Schätzung um mehr als 10% danebenlag.

Deshalb möchte das Unternehmen ein Modell trainieren, um den mittleren Immobilienpreis eines Bezirks aus anderen Angaben über den Bezirk vorherzusagen. Die Daten aus der Volkszählung könnten eine großartige, für diesen Zweck gut nutzbare Datenquelle sein, da sie den mittleren Immobilienpreis und andere Daten für Tausende Bezirke enthalten.

Mit all diesen Informationen sind Sie nun so weit, Ihr System zu entwerfen. Erstens müssen Sie Ihre Aufgabe abstecken: Handelt es sich um überwachtes Lernen, unüberwachtes Lernen oder Reinforcement Learning? Ist es eine Klassifikationsaufgabe, eine Regressionsaufgabe oder etwas anderes? Sollten Sie Techniken aus dem Batch-Learning oder Online-Learning verwenden? Bevor Sie weiterlesen, nehmen Sie sich einen Moment Zeit, und versuchen Sie, sich diese Fragen selbst zu beantworten.

Haben Sie die Antworten gefunden? Schauen wir einmal: Es ist ganz klar eine typische überwachte Lernaufgabe, da Ihnen Lernbeispiele mit *Labels* zur Verfügung stehen (jeder Datenpunkt enthält die erwartete Ausgabe, d.h. den mittleren Immobilienpreis eines Bezirks). Es ist außerdem eine typische Regressionsaufgabe, da Sie einen Zahlenwert vorhersagen sollen. Genauer gesagt, handelt es sich um eine *multiple Regressionsaufgabe*, da das System mehrere Eigenschaften zum Treffen einer Vorhersage heranziehen wird (die Bevölkerung eines Bezirks, das mittlere Einkommen und so weiter). Gleichzeitig ist es auch eine eine *univariate Regressionsaufgabe*, da wir nur versuchen, für jeden Bezirk einen einzelnen Wert vorherzusagen. Würden wir versuchen, mehrere Werte pro Bezirk vorherzusagen, handelte es sich um eine *multivariate Regressionsaufgabe*. Schließlich gibt es keinen kontinuierlichen Strom neuer Daten in das System. Es gibt keinen besonderen Grund, sich auf schnell veränderliche Daten einzustellen, und der Datensatz ist so klein, dass er im Speicher Platz findet. Daher reicht gewöhnliches Batch-Learning völlig aus.

Wenn die Datenmenge riesig wäre, könnten Sie das Batch-Learning entweder auf mehrere Server verteilen (z.B. mit der *MapReduce*-Technik) oder stattdessen eine Technik zum Online-Learning verwenden.

Wähle ein Qualitätsmaß aus

Der nächste Schritt besteht darin, ein geeignetes Qualitätsmaß auszuwählen. Ein typisches Qualitätsmaß für Regressionsaufgaben ist die Wurzel der mittleren quadratischen Abweichung (RMSE). Sie entspricht der Größe des Fehlers, den das System im Mittel bei Vorhersagen macht, wobei großen Fehlern ein höheres Gewicht beigemessen wird. Formel 2-1 zeigt die mathematische Formel zur Berechnung des RMSE.

Formel 2-1: Wurzel der mittleren quadratischen Abweichung (RMSE)

$$\mathrm{RMSE}(\mathbf{X}, h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\left(h(\mathbf{x}^{(i)}) - y^{(i)}\right)^2}$$

Schreibweisen

In dieser Gleichung werden mehrere im Machine Learning übliche Schreibweisen verwendet, die wir im gesamten Buch wiedersehen werden:

- m ist die Anzahl Datenpunkte im Datensatz, für den der RMSE bestimmt wird.
  - Wenn Sie beispielsweise den RMSE für einen Validierungsdatensatz mit 2.000 Bezirken auswerten, dann gilt $m = 2000$.
- $\mathbf{x}^{(i)}$ ist ein Vektor der Werte aller Merkmale (ohne das Label) des i. Datenpunkts im Datensatz, und $y^{(i)}$ ist das dazugehörige Label (der gewünschte Ausgabewert für diesen Datenpunkt).
  - Wenn der erste Bezirk beispielsweise bei –118,29° Länge und 33,91° nördlicher Breite liegt, 1.416 Einwohner mit einem mittleren Einkommen von 38.372 USD hat und der mittlere Immobilienpreis 156.400 USD beträgt (die übrigen Merkmale ignorieren wir noch), dann gilt:

    $$\mathbf{x}^{(1)} = \begin{pmatrix} -118,29 \\ 33,91 \\ 1416 \\ 38372 \end{pmatrix}$$

    und:

    $$y^{(1)} = 156400$$

- $\mathbf{X}$ ist eine Matrix mit den Werten sämtlicher Merkmale (ohne Labels) für alle Datenpunkte im Datensatz. Pro Datenpunkt gibt es eine Zeile, und die i. Zeile entspricht der transponierten Form von $\mathbf{x}^{(i)}$, auch als $(\mathbf{x}^{(i)})^{\mathrm{T}}$ geschrieben.[4]
  - Beispielsweise sieht die Matrix $\mathbf{X}$ für den oben beschriebenen ersten Bezirk folgendermaßen aus:

    $$\mathbf{X} = \begin{pmatrix} (\mathbf{x}^{(1)})^{\mathrm{T}} \\ (\mathbf{x}^{(2)})^{\mathrm{T}} \\ \vdots \\ (\mathbf{x}^{(1999)})^{\mathrm{T}} \\ (\mathbf{x}^{(2000)})^{\mathrm{T}} \end{pmatrix} = \begin{pmatrix} -118.29 & 33.91 & 1416 & 38372 \\ \vdots & \vdots & \vdots & \vdots \end{pmatrix}$$

- h ist die Vorhersagefunktion Ihres Systems, auch *Hypothese* genannt. Wenn Ihr System den Merkmalsvektor eines Datenpunkts $\mathbf{x}^{(i)}$ erhält, gibt es den Vorhersagewert $\hat{y}^{(i)} = h(\mathbf{x}^{(i)})$ für diesen Datenpunkt aus.
  - Sagt Ihr System beispielsweise im ersten Bezirk einen mittleren Immobilienpreis von 158.400 USD vorher, so ist $\hat{y}^{(1)} = h(\mathbf{x}^{(1)}) = 158400$. Der Vorhersagefehler für diesen Bezirk ist dann $\hat{y}^{(1)} - y^{(1)} = 2000$.

4 Der Transponierungsoperator wandelt einen Spaltenvektor in einen Zeilenvektor um und umgekehrt.

- RMSE($\mathbf{X}$,h) ist die auf den Beispieldaten mit Ihrer Hypothese h gemessene Kostenfunktion.

Wir verwenden kursive Kleinbuchstaben für Skalare (wie m oder $y^{(i)}$) und Funktionen (wie h), fett gedruckte Kleinbuchstaben für Vektoren (wie $\mathbf{x}^{(i)}$) und fett gedruckte Großbuchstaben für Matrizen (wie $\mathbf{X}$).

Obwohl der RMSE bei Regressionsaufgaben grundsätzlich das Qualitätsmaß erster Wahl ist, sollten Sie in manchen Situationen eine andere Funktion vorziehen. Nehmen wir an, es gäbe viele Ausreißer unter den Bezirken. In diesem Fall könnten Sie den *mittleren absoluten Fehler* (MAE, auch als mittlere absolute Abweichung bezeichnet; siehe Formel 2-2) berücksichtigen:

Formel 2-2: mittlerer absoluter Fehler

$$\mathrm{MAE}(\mathbf{X}, h) = \frac{1}{m}\sum_{i=1}^{m}\left| h(\mathbf{x}^{(i)}) - y^{(i)} \right|$$

Sowohl RMSE als auch MAE quantifizieren den Abstand zwischen zwei Vektoren: dem Vektor aller Vorhersagen und dem Vektor mit den Zielwerten. Dabei sind unterschiedliche Abstandsmaße oder Normen möglich:

- Die Wurzel einer quadratischen Summe (RMSE) entspricht dem *euklidischen Abstand*: Dies ist der Ihnen vertraute Abstand. Sie wird auch als ℓ_2*-Norm* oder $\|\cdot\|_2$ (oder einfach $\|\cdot\|$) bezeichnet.
- Das Berechnen einer Summe von Absolutwerten (MAE) entspricht der ℓ_1*-Norm*, geschrieben als $\|\cdot\|_1$. Diese nennt man bisweilen auch *Manhattan-Metrik*, da sie den Abstand zwischen zwei Punkten in einer Stadt angibt, in der man sich nur entlang rechtwinkliger Häuserblöcke bewegen kann.
- Allgemeiner ist die ℓ_k*-Norm* eines Vektors $\mathbf{v}$ mit n Elementen als $\|\mathbf{v}\|_k = \left(|v_0|^k + |v_1|^k + \ldots + |v_n|^k\right)^{\frac{1}{k}}$ definiert. ℓ_0 gibt einfach nur die Anzahl der Elemente ungleich null im Vektor wieder, und ℓ_∞ berechnet den größten Absolutwert im Vektor.
- Je höher der Index einer Norm ist, umso stärker berücksichtigt er große Werte und vernachlässigt kleinere. Deshalb ist der RMSE empfindlicher für Ausreißer als der MAE. Sind Ausreißer aber exponentiell selten (wie in einer Glockenkurve), funktioniert der RMSE sehr gut und ist grundsätzlich vorzuziehen.

Überprüfe die Annahmen

Es ist eine gute Angewohnheit, die bisher (von Ihnen oder von anderen) getroffenen Annahmen aufzuschreiben und zu überprüfen; damit können Sie größere Schwierigkeiten früh erkennen. Die von Ihrem System vorhergesagten Preise für einzelne Bezirke werden in ein nachgeschaltetes Machine-Learning-System eingegeben. Wir nehmen an, dass die Preise als solche verwendet werden. Was aber, wenn das nachgeschaltete System stattdessen die Preise in Kategorien einteilt (z. B. »günstig«, »mittel« oder »teuer«) und anstelle der Preise dann diese Kategorien verwendet? In dem Fall wäre es überhaupt nicht wichtig, den Preis genau vorherzusagen; Ihr System müsste lediglich die Kategorie richtig bestimmen. Ist das der Fall, sollte die Aufgabe als Klassifikation beschrieben werden, nicht als Regression. Zu dieser Art von Erkenntnissen möchte man nicht erst nach monatelanger Arbeit an einem Regressionssystem gelangen.

Nachdem Sie mit dem für das nachgeschaltete System zuständigen Team gesprochen haben, sind Sie sich glücklicherweise sicher, dass Sie tatsächlich die Preise und keine Kategorien benötigen. Großartig! Damit sind wir gut aufgestellt und haben grünes Licht, um mit dem Programmieren zu beginnen!

Beschaffe die Daten

Es ist Zeit, sich die Hände schmutzig zu machen. Sie können sich gern Ihren Laptop nehmen und die folgenden Codebeispiele in einem Jupyter-Notebook nachvollziehen. Das vollständige Jupyter-Notebook finden Sie unter *https://github.com/ageron/handson-ml2*.

Erstelle eine Arbeitsumgebung

Zuerst benötigen Sie eine Python-Installation. Python ist vermutlich bereits auf Ihrem System installiert. Falls nicht, finden Sie es unter *https://www.python.org/*.[5]

Als Nächstes müssen Sie sich ein Arbeitsverzeichnis für Ihren Machine-Learning-Code und die Datensätze erstellen. Öffnen Sie eine Kommandozeile und geben Sie die folgenden Befehle dort ein (nach dem $-Prompt):

```
$ export ML_PATH="$HOME/ml"      # Sie können den Pfad ändern, wenn Sie möchten.
$ mkdir -p $ML_PATH
```

Sie benötigen ein paar Python-Module: Jupyter, NumPy, pandas, Matplotlib und Scikit-Learn. Wenn Jupyter bei Ihnen bereits mit all diesen Modulen läuft, können Sie beruhigt bei »Die Daten herunterladen« auf Seite 48 fortfahren. Sollte noch etwas fehlen, gibt es mehrere Möglichkeiten, diese Module (und ihre Paketabhän-

5 Empfohlen ist die neueste Version von Python 3. Python 2.7+ kann auch funktionieren, aber da es mittlerweile überholt ist, beenden alle großen Wissenschaftsbibliotheken den Support dafür, und Sie sollten so schnell wie möglich zu Python 3 wechseln.

gigkeiten) zu installieren. Sie können die Paketverwaltung Ihres Systems verwenden (z.B. apt-get unter Ubuntu oder MacPorts und HomeBrew unter macOS), eine wissenschaftliche Python-Distribution wie Anaconda installieren und dessen Paketverwaltung nutzen oder einfach das in Python eingebaute Paketverwaltungstool pip einsetzen, das (seit Python 2.7.9) standardmäßig Teil der binären Python-Distributionen ist.[6] Mit dem folgenden Befehl können Sie überprüfen, ob pip installiert ist:

```
$ python3 -m pip --version
pip 19.0.2 from [...]/lib/python3.6/site-packages (python 3.6)
```

Sie sollten sicherstellen, dass Sie eine aktuelle Version von pip haben. Um das pip-Modul zu aktualisieren, geben Sie Folgendes ein (die genaue Version mag eine andere sein):[7]

```
$ python3 -m pip install --user -U pip
Collecting pip
[...]
Successfully installed pip-19.0.2
```

Erstellen einer isolierten Umgebung

Falls Sie in einer isolierten Umgebung arbeiten möchten (was sehr empfehlenswert ist, um Konflikte zwischen Modulversionen in unterschiedlichen Projekten zu vermeiden), sollten Sie mit dem folgenden pip-Befehl das Programm virtualenv[8] installieren: (Auch hier gilt wieder: Wollen Sie virtualenv für alle Benutzer auf Ihrem Rechner installieren, entfernen Sie `--user` und lassen diesen Befehl mit Administratorrechten laufen.)

```
$ python3 -m pip install --user -U virtualenv
Collecting virtualenv
[...]
Successfully installed virtualenv
```

Nun können Sie eine isolierte Python-Umgebung erstellen:

```
$ cd $ML_PATH
$ virtualenv my_env
Using base prefix '[...]'
New python executable in [...]/ml/my_env/bin/python3.6
Also creating executable in [...]/ml/my_env/bin/python
Installing setuptools, pip, wheel...done.
```

6 Wir zeigen hier die Installationsschritte mit pip auf der bash-Konsole eines Linux- oder macOS-Systems. Eventuell müssen Sie die Befehle an Ihr System anpassen. Unter Windows empfehlen wir die Installation von Anaconda.

7 Wollen Sie pip nicht nur für sich, sondern für alle Benutzer auf Ihrem Rechner aktualisieren, sollten Sie die Option `--user` weglassen und sicherstellen, dass Sie Administratorrechte haben (beispielsweise durch das Präfix `sudo` vor dem Befehl unter Linux oder macOS).

8 Es gibt auch Alternativen wie venv (sehr ähnlich wie virtualenv und schon in der Standardbibliothek enthalten), virtualenv-wrapper (bietet zusätzliche Funktionalität über virtualenv hinaus), pyenv (erlaubt einen einfachen Wechsel zwischen Python-Versionen) und pipenv (ein großartiges Packaging-Tool von denselben Autoren wie die beliebte Bibliothek `requests`, das auf pip und virtualenv aufbaut).

Jedes Mal, wenn Sie diese Umgebung aktivieren möchten, öffnen Sie einfach eine Kommandozeile und geben dort ein:

```
$ cd $ML_PATH
$ source my_env/bin/activate # unter Linux oder macOS
$ .\my_env\Scripts\activate  # unter Windows
```

Un diese Umgebung zu deaktivieren, geben Sie **deactivate** ein. Solange diese Umgebung aktiv ist, wird jedes von pip installierte Paket in dieser isolierten Umgebung installiert. Python hat nur auf diese Pakete Zugriff (wenn Sie auch die auf dem gesamten System installierten Pakete nutzen möchten, sollten Sie beim Erstellen der Umgebung die Option `--system-site-packages` angeben). Weitere Informationen dazu finden Sie in der Dokumentation zu virtualenv.

Nun können Sie sämtliche benötigten Module und ihre Paketabhängigkeiten mit einem einfachen pip-Befehl installieren (verwenden Sie keine virtualenv, benötigen Sie die Option `--user` oder Administratorrechte):

```
$ python3 -m pip install --upgrade jupyter matplotlib numpy pandas scipy scikit-
learn
Collecting jupyter
  Downloading jupyter-1.0.0-py2.py3-none-any.whl
Collecting matplotlib
  [...]
```

Haben Sie eine virtualenv installiert, müssen Sie sie für Jupyter registrieren und ihr einen Namen geben:

```
$ $ python3 -m ipykernel install --user --name=python3
```

Nun können Sie mit folgendem Befehl Jupyter starten:

```
$ jupyter notebook
[I 15:24 NotebookApp] Serving notebooks from local directory: [...]/ml
[I 15:24 NotebookApp] 0 active kernels
[I 15:24 NotebookApp] The Jupyter Notebook is running at: http://localhost:8888/
[I 15:24 NotebookApp] Use Control-C to stop this server and shut down all
kernels (twice to skip confirmation).
```

Nun läuft in Ihrer Kommandozeile ein Jupyter-Server auf Port 8888. Sie können diesen Server besuchen, indem Sie in Ihrem Browser die Adresse *http://localhost:8888/* eingeben (normalerweise passiert das automatisch beim Starten des Servers). Sie sollten ein leeres Arbeitsverzeichnis sehen (das lediglich das Verzeichnis *env* enthält, falls Sie die obige Anleitung zum Installieren von virtualenv befolgt haben).

Erstellen Sie nun ein neues Python-Notebook mit dem Button »New« ❶ und wählen Sie die gewünschte Python-Version aus ❷[9] (siehe Abbildung 2-3). Dabei passieren drei Dinge: Erstens wird eine neue Datei namens *Untitled.ipynb* für das Notebook in Ihrem Arbeitsverzeichnis angelegt, zweitens wird ein Python-Kernel zum Ausführen

9 Jupyter kann mit mehreren Python-Versionen und sogar vielen anderen Sprachen wie R oder Octave umgehen.

des Notebooks gestartet, und drittens wird das Notebook in einem neuen Unterfenster geöffnet. Zu Beginn sollten Sie das Notebook in »Housing« umbenennen ❶ (siehe Abbildung 2-4), indem Sie auf »Untitled« klicken und den neuen Namen eingeben (damit wird auch die Datei automatisch zu *Housing.ipynb* umbenannt).

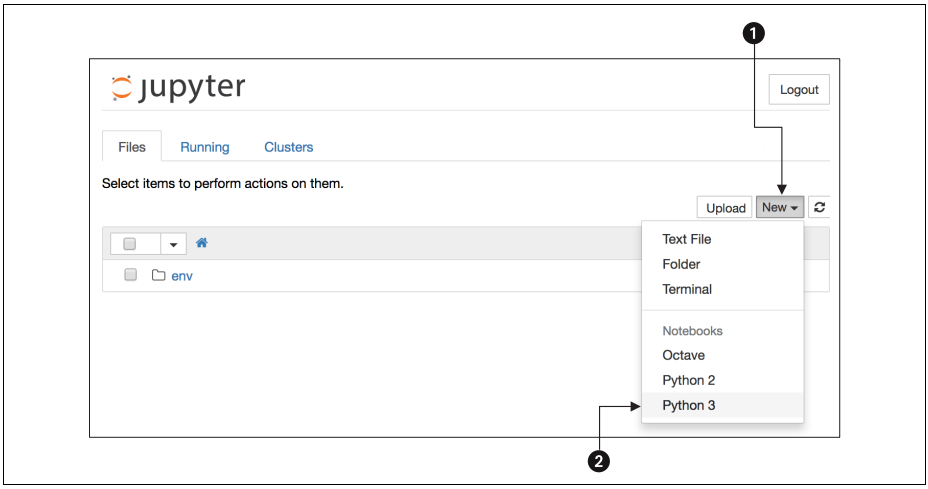


Abbildung 2-3: Ihre Arbeitsumgebung in Jupyter

Ein Notebook besteht aus einer Abfolge von Zellen. Jede Zelle kann ausführbaren Code oder formatierten Text enthalten. Im Moment enthält das Notebook nur eine leere Codezelle mit dem Label *In [1]:*. Schreiben Sie in die Zelle **`print("Hello world!")`** ❷ und drücken Sie dann auf den Play-Button ❸ (siehe Abbildung 2-4) oder drücken Sie Umschalt-Enter. Damit wird der Inhalt der aktuellen Zelle an den Python-Kernel des Notebooks geschickt, der diesen ausführt und eine Ausgabe zurückgibt. Das Ergebnis wird unterhalb der Zelle dargestellt, und weil wir uns am unteren Ende des Notebooks befinden, wird automatisch eine neue leere Zelle hinzugefügt. Zum Erlernen weiterer Grundlagen finden Sie im Hilfemenü von Jupyter eine Tour durch die Benutzeroberfläche.

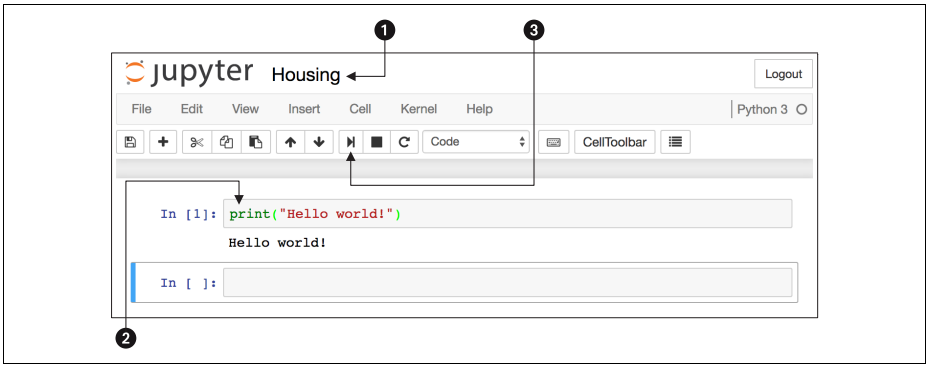


Abbildung 2-4: Python-Notebook mit Hello-World-Befehl

Die Daten herunterladen

In einer typischen Arbeitsumgebung wären Ihre Daten in einer relationalen Datenbank (oder einem anderen Datenspeicher) und über viele Tabellen, Dokumente oder Dateien verteilt. Um auf sie zuzugreifen, müssten Sie zuerst Zugriffsrechte und Passwörter erhalten[10] und sich mit dem Datenmodell vertraut machen. In diesem Projekt sind die Dinge jedoch deutlich einfacher: Sie laden *housing.tgz* herunter, eine einzelne komprimierte Datei, in der sämtliche Daten als kommaseparierte Datei (CSV) namens *housing.csv* vorliegen.

Sie könnten Ihren Browser zum Herunterladen nutzen und anschließend `tar xzf housing.tgz` zum Entpacken und Extrahieren der CSV-Datei eingeben, es ist aber besser, dazu eine kleine Funktion zu schreiben. Eine solche Funktion ist besonders dann nützlich, wenn sich die Daten regelmäßig ändern, weil Sie so die jeweils neuesten Daten mit einem selbst geschriebenen Skript herunterladen können (Sie könnten dieses automatisch in regelmäßigen Abständen ausführen lassen). Den Prozess der Datenbeschaffung zu automatisieren, hilft außerdem, wenn Sie den Datensatz auf mehreren Maschinen installieren möchten.

Mit der folgenden Funktion können Sie die Daten herunterladen:[11]

```
import os
import tarfile
import urllib

DOWNLOAD_ROOT = "https://raw.githubusercontent.com/ageron/handson-ml2/master/"
HOUSING_PATH = os.path.join("datasets", "housing")
HOUSING_URL = DOWNLOAD_ROOT + "datasets/housing/housing.tgz"

def fetch_housing_data(housing_url=HOUSING_URL, housing_path=HOUSING_PATH):
    os.makedirs(housing_path, exist_ok=True)
    tgz_path = os.path.join(housing_path, "housing.tgz")
    urllib.request.urlretrieve(housing_url, tgz_path)
    housing_tgz = tarfile.open(tgz_path)
    housing_tgz.extractall(path=housing_path)
    housing_tgz.close()
```

Wenn Sie nun `fetch_housing_data()` aufrufen, wird das Verzeichnis *datasets/housing* in Ihrer Arbeitsumgebung erstellt, die Datei *housing.tgz* wird heruntergeladen, und die Datei *housing.csv* wird in dieses Verzeichnis entpackt.

Nun laden Sie die Daten mit pandas. Auch diesmal sollten Sie eine kleine Funktion zum Laden der Daten schreiben:

```
import pandas as pd

def load_housing_data(housing_path=HOUSING_PATH):
```

10 Sie müssten eventuell auch gesetzliche Vorgaben berücksichtigen, z. B. Felder mit persönlichen Daten, die niemals in ungeschützte Datenspeicher kopiert werden dürfen.

11 In einem echten Projekt würden Sie den Code in einer Python-Datei ablegen. Im Moment können Sie ihn aber auch in Ihr Jupyter-Notebook schreiben.

```
    csv_path = os.path.join(housing_path, "housing.csv")
    return pd.read_csv(csv_path)
```

Diese Funktion liefert ein pandas-DataFrame-Objekt mit sämtlichen Daten.

Wirf einen kurzen Blick auf die Datenstruktur

Schauen wir uns die ersten fünf Zeilen des DataFrames mit der Methode `head()` an (siehe Abbildung 2-5).

```
In [5]: housing = load_housing_data()
        housing.head()
```

Out[5]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrooms	populatio
0	-122.23	37.88	41.0	880.0	129.0	322.0
1	-122.22	37.86	21.0	7099.0	1106.0	2401.0
2	-122.24	37.85	52.0	1467.0	190.0	496.0
3	-122.25	37.85	52.0	1274.0	235.0	558.0
4	-122.25	37.85	52.0	1627.0	280.0	565.0

Abbildung 2-5: Die ersten fünf Zeilen im Datensatz

Jede Zeile steht für einen Bezirk. Es gibt zehn Merkmale (Sie können die ersten sechs im Screenshot sehen): `longitude`, `latitude`, `housing_median_age`, `total_rooms`, `total_bedrooms`, `population`, `households`, `median_income`, `median_house_value` und `ocean_proximity`.

Die Methode `info()` hilft, schnell eine Beschreibung der Daten zu erhalten. Dies sind insbesondere die Anzahl der Zeilen, der Typ jedes Attributs und die Anzahl der Werte ungleich null (siehe Abbildung 2-6).

```
In [6]: housing.info()
        <class 'pandas.core.frame.DataFrame'>
        RangeIndex: 20640 entries, 0 to 20639
        Data columns (total 10 columns):
        longitude             20640 non-null float64
        latitude              20640 non-null float64
        housing_median_age    20640 non-null float64
        total_rooms           20640 non-null float64
        total_bedrooms        20433 non-null float64
        population            20640 non-null float64
        households            20640 non-null float64
        median_income         20640 non-null float64
        median_house_value    20640 non-null float64
        ocean_proximity       20640 non-null object
        dtypes: float64(9), object(1)
        memory usage: 1.6+ MB
```

Abbildung 2-6: Informationen zu Immobilien

Im Datensatz gibt es 20.640 Datenpunkte. Damit ist er für Machine-Learning-Verhältnisse eher klein, für den Anfang ist das aber ausgezeichnet! Beachten Sie, dass das Merkmal `total_bedrooms` nur 20.433 Werte ungleich null hat, es gibt also 207 Bezirke ohne diese Angabe. Darum werden wir uns später kümmern müssen.

Bis auf das Feld `ocean_proximity` sind sämtliche Merkmale numerisch. Dessen Typ ist `object`, und es könnte beliebige Python-Objekte enthalten. Da Sie aber diese Daten aus einer CSV-Datei geladen haben, muss es sich dabei natürlich um Text handeln. Beim Betrachten der ersten fünf Zeilen haben Sie möglicherweise bemerkt, dass sich die Werte in der Spalte `ocean_proximity` wiederholen. Es handelt sich dabei also um ein kategorisches Merkmal. Sie können mit der Methode `value_counts()` herausfinden, welche Kategorien es gibt und wie viele Bezirke zu jeder Kategorie gehören:

```
>>> housing["ocean_proximity"].value_counts()
<1H OCEAN     9136
INLAND        6551
NEAR OCEAN    2658
NEAR BAY      2290
ISLAND           5
Name: ocean_proximity, dtype: int64
```

Betrachten wir auch die anderen Spalten. Die Methode `describe()` fasst die numerischen Merkmale zusammen (siehe Abbildung 2-7).

```
In [8]: housing.describe()
```

Out[8]:

	longitude	latitude	housing_median_age	total_rooms	total_bedrc
count	20640.000000	20640.000000	20640.000000	20640.000000	20433.000(
mean	-119.569704	35.631861	28.639486	2635.763081	537.870553
std	2.003532	2.135952	12.585558	2181.615252	421.38507(
min	-124.350000	32.540000	1.000000	2.000000	1.000000
25%	-121.800000	33.930000	18.000000	1447.750000	296.00000(
50%	-118.490000	34.260000	29.000000	2127.000000	435.00000(
75%	-118.010000	37.710000	37.000000	3148.000000	647.00000(
max	-114.310000	41.950000	52.000000	39320.000000	6445.00000(

Abbildung 2-7: Zusammenfassung aller numerischen Merkmalen

Die Zeilen `count`, `mean`, `min` und `max` sind selbsterklärend. Beachten Sie, dass die leeren Werte ignoriert werden (z. B. beträgt `count` bei `total_bedrooms` 20433, nicht 20640). Die Zeile `std` enthält die *Standardabweichung*, die die Streuung der Werte angibt.[12]

12 Die Standardabweichung wird im Allgemeinen mit σ angegeben (dem griechischen Buchstaben Sigma) und ist die Quadratwurzel der *Varianz*, der durchschnittlichen quadratischen Abweichung vom Mittelwert. Wenn ein Merkmal der sehr häufigen glockenförmigen *Normalverteilung* folgt (auch *Gaußverteilung* genannt), gilt die »68-95-99,7«-Regel: Etwa 68 % der Werte liegen innerhalb von 1σ des Mittelwerts, 95 % innerhalb von 2σ und 99,7 % innerhalb von 3σ.

Die Zeilen mit 25 %, 50 % und 75 % zeigen die entsprechenden *Perzentile*: Ein Perzentil besagt, dass ein bestimmter prozentualer Anteil der Beobachtungen unterhalb eines Werts liegt. Beispielsweise haben 25 % der Bezirke ein `housing_median_age` unter 18, 50 % liegen unter 29, und 75 % liegen unter 37. Diese nennt man oft das 25. Perzentil (oder 1. *Quartil*), den Median und das 75. Perzentil (oder 3. Quartil).

Eine andere Möglichkeit, schnell einen Eindruck von den Daten, die wir sehen, zu erhalten, ist, für jedes numerische Merkmal ein Histogramm zu plotten. Ein Histogramm zeigt die Anzahl Datenpunkte (auf der vertikalen Achse), die in einem bestimmten Wertebereich (auf der horizontalen Achse) liegen. Sie können diese entweder für jedes Merkmal einzeln plotten oder die Methode `hist()` für den gesamten Datensatz aufrufen (wie im folgenden Codebeispiel) und für jedes numerische Merkmal ein Histogramm erhalten (siehe Abbildung 2-8):

```
%matplotlib inline   # nur im Jupyter-Notebook
import matplotlib.pyplot as plt
housing.hist(bins=50, figsize=(20,15))
plt.show()
```

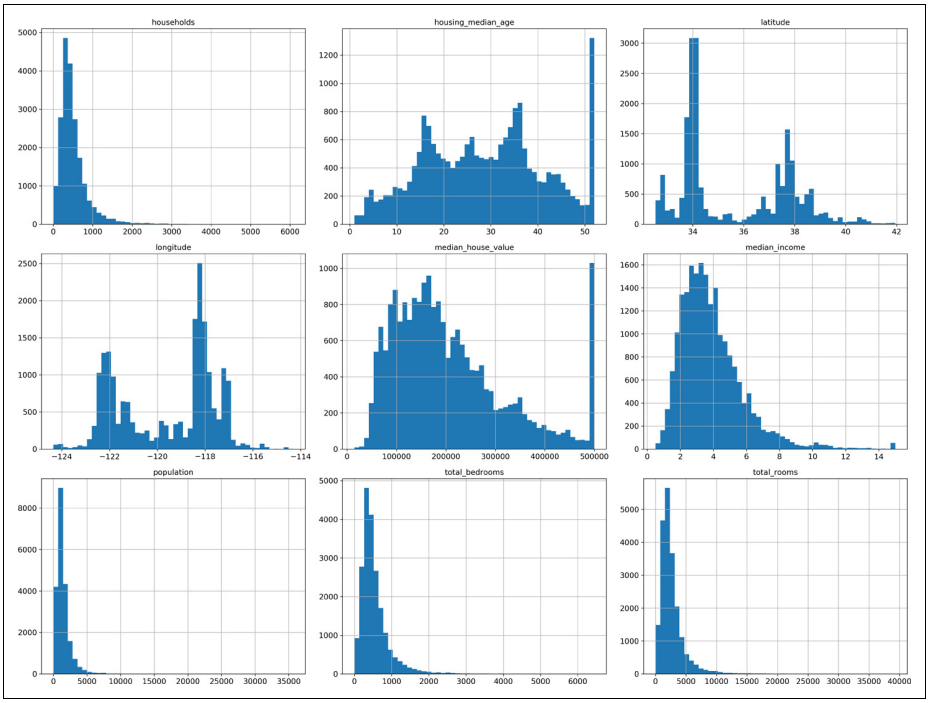


Abbildung 2-8: Ein Histogramm für jedes numerische Attribut

Die Methode `hist()` verwendet Matplotlib, das wiederum ein nutzerabhängiges grafisches Backend zum Zeichnen auf den Bildschirm verwendet. Bevor Sie also etwas plotten können, müssen Sie Matplotlib darüber informieren, welches Backend es verwenden soll. Die einfachste Möglichkeit ist, in Jupyter das magische Kommando `%matplotlib inline` einzusetzen. Dieses weist Jupyter an, Matplotlib zu nutzen, sodass Jupyter als Backend verwendet wird. Die Diagramme werden dann im Notebook selbst dargestellt, wobei Jupyter sie automatisch generiert, sobald eine Zelle ausgeführt wird.

In diesen Histogrammen gibt es einiges zu sehen:

1. Erstens sieht das mittlere Einkommen nicht nach Werten in US-Dollar aus (USD). Nachdem Sie sich mit dem Team, das die Daten erhoben hat, in Verbindung gesetzt haben, erfahren Sie, dass die Daten skaliert wurden und für höhere mittlere Einkommen nach oben bei 15 (genau 15,0001) und für geringere mittlere Einkommen nach unten bei 0,5 (genau 0,4999) abgeschnitten wurden. Die Zahlen stehen ungefähr für 10.000 USD (eine 3 bedeutet also etwa 30.000 USD). Es ist im Machine Learning durchaus üblich, mit solchen vorverarbeiteten Merkmalen zu arbeiten, was nicht notwendigerweise ein Problem darstellt. Sie sollten aber versuchen, nachzuvollziehen, wie die Daten berechnet wurden.
2. Das mittlere Alter und der mittlere Wert von Gebäuden wurden ebenfalls gekappt. Letzteres könnte sich als ernstes Problem herausstellen, da Ihre Zielgröße (Ihr Label) betroffen ist. Ihre Machine-Learning-Algorithmen könnten dann lernen, dass es keine Preise jenseits dieser Obergrenze gibt. Sie müssen mit Ihrem Team (dem Team, das die Ausgabe Ihres Systems nutzen möchte) klären, ob das ein Problem darstellt oder nicht. Wenn Ihnen erklärt wird, dass auch jenseits von 500.000 USD präzise Vorhersagen nötig sind, haben Sie zwei Alternativen:
   a. Für die nach oben begrenzten Bezirke korrekte Labels zu sammeln.
   b. Die entsprechenden Bezirke aus dem Trainingsdatensatz zu entfernen (auch aus dem Testdatensatz, da Ihr System nicht als schlechter eingestuft werden sollte, wenn es Werte jenseits von 50.000 USD vorhersagt).
3. Diese Attribute haben sehr unterschiedliche Wertebereiche. Wir werden das weiter unten in diesem Kapitel besprechen, wenn wir uns dem Skalieren von Merkmalen widmen.
4. Schließlich sind viele der Histogramme *rechtsschief*: Sie erstrecken sich viel weiter vom Median nach rechts als nach links. Dadurch wird das Erkennen von Mustern für einige Machine-Learning-Algorithmen schwieriger. Wir werden später versuchen, diese Merkmale zu einer annähernd glockenförmigen Verteilung zu transformieren.

Hoffentlich haben Sie nun einen besseren Eindruck von den Daten, mit denen Sie sich beschäftigen.

Warten Sie! Bevor Sie sich die Daten weiter ansehen, sollten Sie einen Testdatensatz erstellen, beiseitelegen und nicht hineinschauen.

Erstelle einen Testdatensatz

Es mag sich seltsam anhören, an dieser Stelle einen Teil der Daten freiwillig beiseitezulegen. Schließlich haben wir gerade erst einen kurzen Blick auf die Daten geworfen, und Sie sollten bestimmt noch weiter analysieren, bevor Sie sich für einen Algorithmus entscheiden, oder? Das ist zwar richtig, aber Ihr Gehirn ist ein faszinierendes System zur Mustererkennung. Es ist daher äußerst anfällig für Overfitting: Wenn Sie sich die Testdaten ansehen, könnten Sie auf ein interessantes Muster im Datensatz stoßen, das Sie zur Auswahl eines bestimmten Machine-Learning-Modells veranlasst. Wenn Sie den Fehler der Verallgemeinerung anhand des Testdatensatzes schätzen, wird Ihr Schätzwert zu optimistisch ausfallen, und Sie würden in der Folge ein System starten, das die erwartete Vorhersageleistung nicht erfüllt. Dies nennt man auch das *Data- Snooping*-Bias.

Einen Testdatensatz zu erstellen, ist theoretisch einfach: Wählen Sie zufällig einige Datenpunkte aus, meist 20% des Datensatzes (oder weniger, wenn Ihr Datensatz sehr groß ist), und legen Sie diese beiseite:

```
import numpy as np

def split_train_test(data, test_ratio):
    shuffled_indices = np.random.permutation(len(data))
    test_set_size = int(len(data) * test_ratio)
    test_indices = shuffled_indices[:test_set_size]
    train_indices = shuffled_indices[test_set_size:]
    return data.iloc[train_indices], data.iloc[test_indices]
```

Sie können diese Funktion anschließend folgendermaßen verwenden:[13]

```
>>> train_set, test_set = split_train_test(housing, 0.2)
>>> len(train_set)
16512
>>> len(test_set)
4128
```

Das funktioniert, ist aber noch nicht perfekt: Wenn Sie dieses Programm erneut ausführen, erzeugt es einen anderen Testdatensatz! Sie (oder Ihre Machine-Learning-Algorithmen) werden mit der Zeit den kompletten Datensatz als Gesamtes sehen, was Sie ja genau vermeiden möchten.

Eine Lösungsmöglichkeit besteht darin, den Testdatensatz beim ersten Durchlauf zu speichern und in späteren Durchläufen zu laden. Eine andere Möglichkeit ist,

13 Enthält in diesem Buch ein Codebeispiel einen Mix aus Code und Ausgaben – wie in diesem Fall –, ist es zur besseren Lesbarkeit wie im Python-Interpreter formatiert: Den Codezeilen ist >>> (oder ... für eingerückte Blöcke) vorangestellt, die Ausgabezeilen haben keinen Präfix.

den Seed-Wert des Zufallsgenerators festzulegen (z. B. mit `np.random.seed(42)`)[14], bevor Sie `np.random.permutation()` aufrufen, sodass jedes Mal die gleichen durchmischten Indizes generiert werden.

Allerdings scheitern beide Lösungsansätze, sobald Sie einen aktualisierten Datensatz erhalten. Um auch danach über eine stabile Trennung zwischen Trainings- und Testdatensatz zu verfügen, können Sie als Alternative einen eindeutigen Identifikator verwenden, um zu entscheiden, ob ein Datenpunkt in den Testdatensatz aufgenommen werden soll oder nicht (vorausgesetzt, die Datenpunkte haben eindeutige unveränderliche Identifikatoren). Sie könnten beispielsweise aus dem Identifikator eines Datenpunkts einen Hash berechnen und den Datenpunkt in den Testdatensatz aufzunehmen, falls der Hash kleiner oder gleich 20% des maximalen Hashwerts ist. Damit stellen Sie sicher, dass der Testdatensatz über mehrere Durchläufe konsistent ist, selbst wenn Sie ihn aktualisieren. Ein neuer Datensatz enthält auf diese Weise 20% der neuen Datenpunkte, aber keinen der Datenpunkte, die zuvor im Trainingsdatensatz waren.

Hier folgt eine mögliche Implementierung:

```
from zlib import crc32

def test_set_check(identifier, test_ratio):
    return crc32(np.int64(identifier)) & 0xffffffff < test_ratio * 2**32

def split_train_test_by_id(data, test_ratio, id_column):
    ids = data[id_column]
    in_test_set = ids.apply(lambda id_: test_set_check(id_, test_ratio))
    return data.loc[~in_test_set], data.loc[in_test_set]
```

Leider gibt es im Immobiliendatensatz keine Identifikatorspalte. Die Lösung ist, den Zeilenindex als ID zu nutzen:

```
housing_with_id = housing.reset_index()   # fügt die Spalte `index` hinzu
train_set, test_set = split_train_test_by_id(housing_with_id, 0.2, "index")
```

Wenn Sie den Zeilenindex als eindeutigen Identifikator verwenden, müssen Sie sicherstellen, dass die neuen Daten am Ende des Datensatzes angehängt werden und nie eine Zeile gelöscht wird. Falls das nicht möglich ist, können Sie immer noch versuchen, einen eindeutigen Identifikator aus den stabilsten Merkmalen zu entwickeln. Beispielsweise werden geografische Länge und Breite garantiert für die nächsten paar Millionen Jahre stabil bleiben, daher könnten Sie diese folgendermaßen zu einer ID kombinieren:[15]

14 Sie werden häufig sehen, dass der Seed-Wert auf 42 gesetz wird. Diese Zahl hat keine besondere Bedeutung, außer dass sie die Antwort auf die ultimative Frage nach dem Leben, dem Universum und dem ganzen Rest ist.

15 Die Koordinatenangaben sind recht grob, daher werden viele Bezirke eine identische ID erhalten und somit im gleichen Teildatensatz landen (Test oder Training). Damit haben wir unglücklicherweise ein Bias in der Auswahl.

```
housing_with_id["id"] = housing["longitude"] * 1000 + housing["latitude"]
train_set, test_set = split_train_test_by_id(housing_with_id, 0.2, "id")
```

Scikit-Learn enthält einige Funktionen, die Datensätze auf unterschiedliche Weise in Teildatensätze aufteilen. Die einfachste Funktion darunter ist `train_test_split()`, die so ziemlich das Gleiche tut wie die oben definierte Funktion `split_train_test()`, aber einige zusätzliche Optionen bietet. Erstens gibt es den Parameter `random_state`, der den Seed-Wert des Zufallszahlengenerators festlegt. Und zweitens können Sie mehrere Datensätze mit einer identischen Anzahl Zeilen übergeben, die anhand der gleichen Indizes aufgeteilt werden (das ist sehr nützlich, z. B. wenn Sie ein separates DataFrame mit den Labels haben):

```
from sklearn.model_selection import train_test_split

train_set, test_set = train_test_split(housing, test_size=0.2, random_state=42)
```

Bisher haben wir ausschließlich zufallsbasierte Methoden zur Stichprobenauswahl betrachtet. Wenn Ihr Datensatz groß genug ist (insbesondere im Vergleich zur Anzahl der Merkmale), ist daran nichts auszusetzen. Ist er aber nicht groß genug, besteht das Risiko, ein erhebliches Stichproben-Bias zu verursachen. Wenn ein Umfrageunternehmen 1.000 Personen anruft, um diesen Fragen zu stellen, wählt es nicht einfach nur zufällig 1.000 Probanden aus dem Telefonbuch aus. Beispielsweise besteht die Bevölkerung der USA aus 51,3 % Frauen und 48,7 % Männern, also sollte eine gut aufgebaute Studie in den USA dieses Verhältnis auch in der Stichprobe repräsentieren: 513 Frauen und 487 Männer. Dies bezeichnet man als *stratifizierte Stichprobe*: Die Bevölkerung wird in homogene Untergruppen, die *Strata*, aufgeteilt, und aus jedem Stratum wird die korrekte Anzahl Datenpunkte ausgewählt. Damit ist garantiert, dass der Testdatensatz die Gesamtbevölkerung angemessen repräsentiert. Würde die Stichprobe rein zufällig ausgewählt, gäbe es eine 12%ige Chance, dass die Stichprobe verzerrt ist und entweder weniger als 49 % Frauen oder mehr als 54 % Frauen im Datensatz enthalten sind. In beiden Fällen wären die Ergebnisse mit einem erheblichen Bias behaftet.

Nehmen wir an, Experten hätten Ihnen in einer Unterhaltung erklärt, dass das mittlere Einkommen ein sehr wichtiges Merkmal zur Vorhersage des mittleren Immobilienpreises ist. Sie möchten sicherstellen, dass der Testdatensatz die unterschiedlichen im Datensatz enthaltenen Einkommensklassen gut repräsentiert. Da das mittlere Einkommen ein stetiges numerisches Merkmal ist, müssen Sie zuerst ein kategorisches Merkmal für das Einkommen generieren. Betrachten wir das Histogramm des Einkommens etwas genauer (siehe Abbildung 2-8): Die meisten mittleren Einkommen liegen bei 1,6 bis 6 (also 15.000 bis 60.000 USD), aber einige mittlere Einkommen liegen deutlich über 6. Es ist wichtig, dass Ihr Datensatz für jedes Stratum eine genügende Anzahl Datenpunkte enthält, andernfalls liegt ein Bias für die Schätzung der Wichtigkeit des Stratums vor. Das heißt, Sie sollten nicht zu viele Strata haben, und jedes Stratum sollte groß genug sein. Der folgende Code nutzt die Funktion `pd.cut()`, um ein kategorisches Merkmal für das Einkom-

men mit fünf Kategorien zu erzeugen (mit den Labels 1 bis 5): Kategorie 1 reicht von 0 bis 1,5 (also weniger als 15.000 USD), Kategorie 2 von 1,5 bis 3 und so weiter:

```
housing["income_cat"] = pd.cut(housing["median_income"],
                               bins=[0., 1.5, 3.0, 4.5, 6., np.inf],
                               labels=[1, 2, 3, 4, 5])
```

Diese Einkommenskategorien sind in Abbildung 2-9 dargestellt:

```
housing["income_cat"].hist()
```

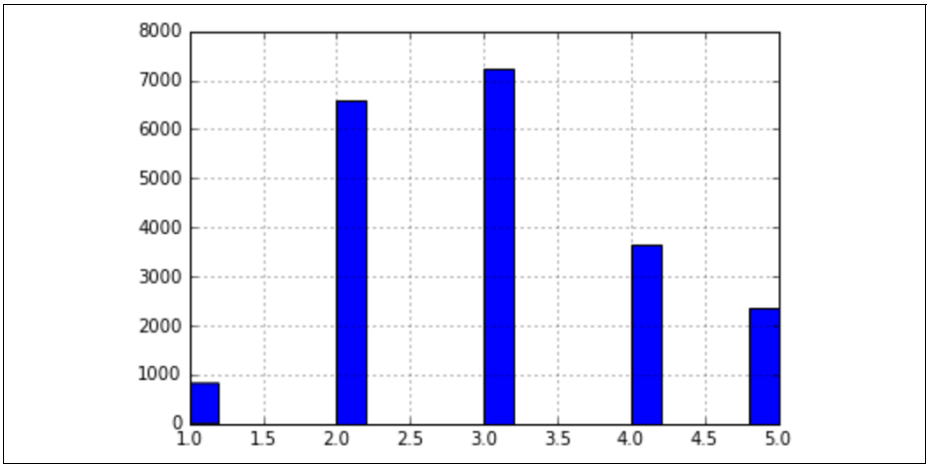


Abbildung 2-9: Histogramm der Einkommenskategorien

Nun sind wir so weit, eine stratifizierte Stichprobe anhand der Einkommenskategorie zu ziehen. Dazu können Sie die Klasse `StratifiedShuffleSplit` aus Scikit-Learn verwenden:

```
from sklearn.model_selection import StratifiedShuffleSplit

split = StratifiedShuffleSplit(n_splits=1, test_size=0.2, random_state=42)
for train_index, test_index in split.split(housing, housing["income_cat"]):
    strat_train_set = housing.loc[train_index]
    strat_test_set = housing.loc[test_index]
```

Prüfen wir, ob das wie erwartet funktioniert hat. Zunächst einmal können Sie sich die Anteile der Einkommenskategorien im Testdatensatz ansehen:

```
>>> strat_test_set["income_cat"].value_counts() / len(strat_test_set)
3    0.350533
2    0.318798
4    0.176357
5    0.114583
1    0.039729
Name: income_cat, dtype: float64
```

Mit einer ähnlichen Codezeile lassen sich die Anteile der Einkommenskategorien im vollständigen Datensatz bestimmen. In Abbildung 2-10 werden die Anteile der Einkommenskategorien im gesamten Datensatz, im als stratifizierte Stichprobe

generierten Testdatensatz und in einem rein zufälligen Testdatensatz miteinander verglichen. Wie Sie sehen, sind die Anteile der Einkommenskategorien in der stratifizierten Stichprobe beinahe die gleichen wie im Gesamtdatensatz, während der als zufällige Stichprobe erzeugte Testdatensatz recht verzerrt ist.

Nun sollten Sie das Merkmal `income_cat` entfernen, damit die Daten wieder in ihrem Ursprungszustand sind:

```
for set_ in (strat_train_set, strat_test_set):
    set_.drop("income_cat", axis=1, inplace=True)
```

Wir haben uns aus gutem Grund eine Menge Zeit für das Erstellen des Testdatensatzes genommen – ist es doch ein oft vernachlässigter, aber entscheidender Teil eines Machine-Learning-Projekts. Viele der hier vorgestellten Ideen werden noch nützlich sein, sobald wir die Kreuzvalidierung besprechen. Nun ist es an der Zeit, mit dem nächsten Abschnitt fortzufahren: dem Erkunden der Daten.

	Overall	Random	Stratified	Rand. %error	Strat. %error
1.0	0.039826	0.040213	0.039738	0.973236	-0.219137
2.0	0.318847	0.324370	0.318876	1.732260	0.009032
3.0	0.350581	0.358527	0.350618	2.266446	0.010408
4.0	0.176308	0.167393	0.176399	-5.056334	0.051717
5.0	0.114438	0.109496	0.114369	-4.318374	-0.060464

Abbildung 2-10: Vergleich des Stichproben-Bias einer stratifizierten und einer zufälligen Stichprobe

Erkunde und visualisiere die Daten, um Erkenntnisse zu gewinnen

Bisher haben wir nur einen kurzen Blick auf die Daten geworfen, um ein allgemeines Verständnis von der Art der verarbeiteten Daten zu erhalten. Nun ist unser Ziel, etwas mehr in die Tiefe zu gehen.

Zunächst sollten Sie sicherstellen, dass Sie den Testdatensatz beiseitegelegt haben und nur noch den Trainingsdatensatz erkunden. Wenn Ihr Trainingsdatensatz sehr groß ist, sollten Sie eine Stichprobe ziehen, um sich die Arbeit beim Erkunden zu erleichtern und sie zu beschleunigen. In unserem Fall ist der Datensatz recht klein, sodass Sie direkt mit den vollständigen Daten arbeiten können. Erzeugen wir eine Kopie, mit der Sie experimentieren können, ohne den Trainingsdatensatz zu beschädigen:

```
housing = strat_train_set.copy()
```

Visualisieren geografischer Daten

Weil uns geografische Informationen zur Verfügung stehen (Breite und Länge), sollten wir einen Scatterplot sämtlicher Bezirke erzeugen, um uns die Daten anzusehen (siehe Abbildung 2-11):

```
housing.plot(kind="scatter", x="longitude", y="latitude")
```

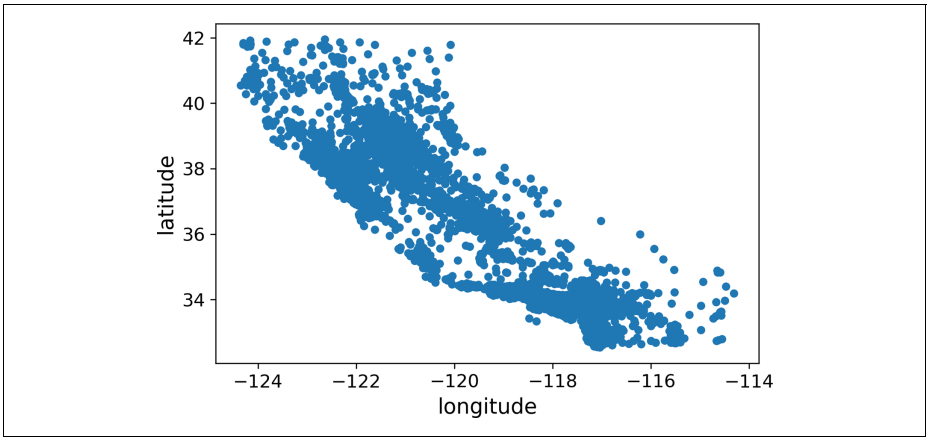


Abbildung 2-11: Ein geografischer Scatterplot der Daten

In Ordnung, die Abbildung sieht wie Kalifornien aus, aber abgesehen davon ist es schwierig, irgendein Muster zu erkennen. Setzen wir den Parameter `alpha` auf 0,1, wird es viel einfacher, die Orte mit einer hohen Dichte an Datenpunkten zu erkennen (siehe Abbildung 2-12):

```
housing.plot(kind="scatter", x="longitude", y="latitude", alpha=0.1)
```

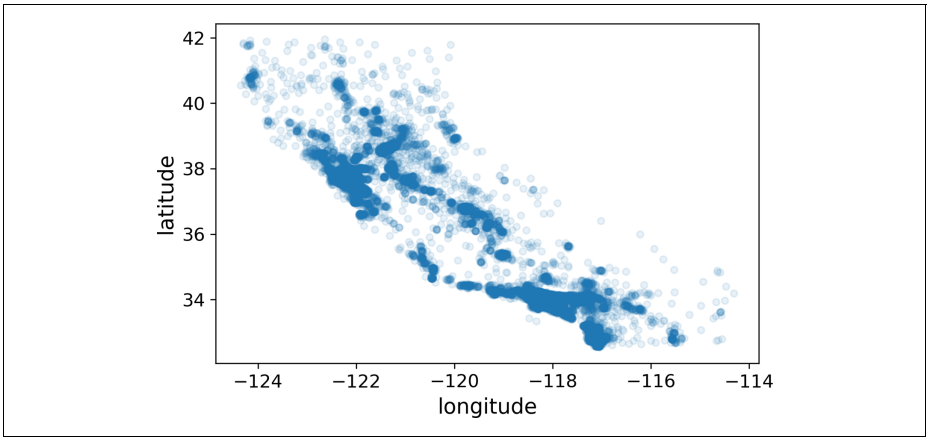


Abbildung 2-12: Eine bessere Darstellung von Gebieten mit hoher Dichte