Mathe-Basics für Data Scientists

Lineare Algebra, Statistik und Wahrscheinlichkeitsrechnung für die Datenanalyse

DAS VORWORT

Einführung

In den letzten zehn Jahren hat das Interesse an angewandter Mathematik und Statistik in unserem Arbeits- und Lebensalltag zugenommen. Warum ist das so? Hat es mit dem gestiegenen Interesse an *Data Science* zu tun, das von der Harvard Business Review als der »sexiest Job des 21. Jahrhunderts« deklariert wurde (*https://oreil.ly/GslO6*)? Oder ist es das Versprechen, dass maschinelles Lernen und *künstliche Intelligenz* unser Leben verändern werden? Liegt es daran, dass die Schlagzeilen mit Studien, Umfragen und Forschungsergebnissen überschwemmt werden, wir aber nicht wissen, wie wir solche Behauptungen überprüfen sollen? Oder ist es das Versprechen, dass »selbstfahrende« Autos und Roboter in naher Zukunft Arbeitsplätze automatisieren werden?

Ich behaupte, dass die Disziplinen Mathematik und Statistik aufgrund der zunehmenden Verfügbarkeit von Daten das breite Interesse der Öffentlichkeit geweckt haben und wir Mathematik, Statistik und maschinelles Lernen benötigen, um diese Daten sinnvoll zu nutzen. Ja, wir haben wissenschaftliche Werkzeuge, maschinelles Lernen und andere Instrumente der Automatisierung, die wie die Sirenen nach uns rufen. Diesen »Blackboxes«, Geräten und Programmen vertrauen wir blindlings; wir verstehen sie zwar nicht, aber wir nutzen sie trotzdem.

Man ist zwar geneigt zu glauben, dass Computer schlauer sind als wir (und dieser Gedanke wird häufig vermarktet), aber die Realität ist weit davon entfernt. Diese Diskrepanz kann auf vielen Ebenen heikel sein. Wollen Sie wirklich, dass ein Algorithmus oder eine KI eine Strafe verhängt oder ein Fahrzeug steuert, dass aber niemand, nicht einmal der Entwickler, erklären kann, wie es zu einer bestimmten Entscheidung gekommen ist? Erklärbarkeit ist die nächste Stufe der statistischen Datenverarbeitung und der KI. Diese kann erst dann beginnen, wenn wir die Blackbox öffnen und die Mathematik enthüllen.

Vielleicht fragen Sie sich auch, wieso eine Entwicklerin nicht wissen kann, wie ihr eigener Algorithmus funktioniert. Darauf gehen wir in der zweiten Hälfte des Buchs ein, wenn wir Techniken des maschinellen Lernens erörtern und aufzeigen, warum wir die Mathematik hinter den von uns erstellten Blackboxes verstehen müssen.

Ein weiterer Punkt ist, dass der Grund für die massenhafte Datenerfassung vor allem in den vernetzten Geräten und ihrer Präsenz in unserem Alltag liegt. Wir nutzen das Internet nicht mehr nur über einen Desktop- oder Laptop-Computer. Wir haben es jetzt auch in unseren Smartphones, Autos und Haushaltsgeräten dabei. Dies hat in den letzten zwei Jahrzehnten einen fast unmerklichen Wandel in Gang gesetzt. Daten haben sich von einem operativen Werkzeug zu etwas entwickelt, das für Ziele gesammelt und analysiert wird, die noch gar nicht genau umrissen sind. Eine Smartwatch sammelt ständig Daten über unsere Herzfrequenz, die Atmung, die zu Fuß zurückgelegte Strecke und andere Kennwerte. Dann lädt sie die Daten hoch in eine Cloud, um sie zusammen mit anderen Usern zu analysieren. Unsere Fahrgewohnheiten werden von computergesteuerten Autos erfasst und von den Herstellern genutzt, um Daten für selbstfahrende Fahrzeuge zusammenzutragen. Sogar »intelligente Zahnbürsten« finden ihren Weg in die Drogerien. Sie zeichnen die Putzgewohnheiten auf und speichern diese Daten in der Cloud. Es lässt sich darüber streiten, ob die Daten der intelligenten Zahnbürste nützlich und wichtig sind!

Dieses ganze Datensammeln durchdringt jeden Winkel unseres Lebens. Das kann erdrückend sein, und man könnte ein ganzes Buch über Bedenken hinsichtlich Datenschutz und Ethik schreiben. Aber die Verfügbarkeit von Daten bietet auch die Möglichkeit, Mathematik und Statistik auf neue Art und Weise zu nutzen und sich außerhalb des akademischen Umfelds stärker zu engagieren. Wir können mehr über die menschliche Erfahrung lernen, das Design und die Anwendung von Produkten verbessern sowie kommerzielle Strategien optimieren. Wenn Sie die in diesem Buch vorgestellten Ideen verstehen, werden Sie den Wert erschließen können, der in unserer Daten sammelnden Infrastruktur steckt. Das bedeutet nicht, dass Daten und statistische Werkzeuge ein Allheilmittel sind, um alle Probleme dieser Welt zu lösen, aber sie bieten uns neue Werkzeuge, die wir nutzen können. Manchmal ist es ebenso wertvoll, bestimmte Datenprojekte als Irrwege anzuerkennen und sich klarzumachen, dass die Anstrengungen an anderer Stelle besser investiert sind.

Die zunehmende Verfügbarkeit von Daten hat bewirkt, dass Data Science und maschinelles Lernen zu gefragten Berufen geführt haben. Wir definieren die Mathe-Basics als einen Umgang mit Wahrscheinlichkeit, linearer Algebra, Statistik und maschinellem Lernen. Wenn Sie eine Karriere in den Bereichen Data Science, maschinelles Lernen oder Softwaretechnik anstreben, sind diese Themen unerlässlich. Ich werde nur so viel Hochschulmathematik, Analysis und Statistik einbringen, wie es für das Verständnis der Blackbox-Bibliotheken – mit denen Sie arbeiten werden – unerlässlich ist.

Mit diesem Buch möchte ich Leserinnen und Lesern verschiedene Bereiche der Mathematik, Statistik und des maschinellen Lernens näherbringen, die auf reale Probleme anwendbar sind. Die ersten vier Kapitel befassen sich mit grundlegenden mathematischen Konzepten, darunter praktische Analysis, Wahrscheinlichkeit, lineare Algebra und Statistik. Die letzten drei Kapitel sind dem maschinellen Lernen gewid-

met. Der ultimative Zweck eines Lehrbuchs über maschinelles Lernen ist es, alles zu integrieren, was wir lernen, und praktische Einblicke in die Verwendung von maschinellem Lernen und statistischen Bibliotheken über ein Blackbox-Verständnis hinaus zu demonstrieren.

Um den Beispielen zu folgen, benötigen Sie lediglich einen Windows-/Mac-/Linux-Computer und eine Python-3-Umgebung Ihrer Wahl. Die wichtigsten Python-Bibliotheken, die Sie benötigen werden, sind `numpy`, `scipy`, `sympy` und `sklearn`. Wenn Sie mit Python nicht vertraut sind: Keine Angst, es handelt sich um eine einfache und übersichtliche Programmiersprache, die sich der Unterstützung durch riesige Lernressourcen erfreut. Unter anderem empfehle ich folgende Quellen:

Einführung in Data Science: Grundprinzipien der Datenanalyse mit Python, 2. Auflage, von Joel Grus (O'Reilly)
: Das zweite Kapitel dieses Buchs ist der beste Crashkurs in Python, der mir je untergekommen ist. Selbst wenn Sie noch nie zuvor ein Programm geschrieben haben, gelingt es Joel auf fantastische Art und Weise, Ihnen in kürzester Zeit einen effektiven Einstieg in Python zu ermöglichen. Zudem ist es ein großartiges Buch, das man im Regal stehen hat und mit dem man sein mathematisches Wissen anwenden kann!

Python for the Busy Java Developer von Deepak Sarda (Apress)
: Wenn Sie als Softwareentwicklerin oder Softwareentwickler von einem statisch typisierten, objektorientierten Programmierhintergrund kommen, sollten Sie zu diesem Buch greifen. Als jemand, der mit Java zu programmieren begonnen hat, schätze ich es sehr, wie Deepak die Features von Python insbesondere für Java-Entwickler erörtert. Sollten Sie mit .NET, C++ oder anderen C-ähnlichen Sprachen gearbeitet haben, werden Sie Python mit diesem Buch aller Voraussicht nach effektiv lernen.

Dieses Buch wird Sie nicht zu einem Experten machen oder Ihnen einen Doktortitel verleihen. Nach Möglichkeit vermeide ich mathematische Ausdrücke, die mit griechischen Buchstaben gespickt sind, und werde sie stattdessen in einfacher Sprache beschreiben. Aber dieses Buch wird Ihnen den Umgang mit Mathematik und Statistik erleichtern und Ihnen das *wesentliche* Wissen vermitteln, damit Sie sich in diesen Bereichen erfolgreich bewegen können. Meiner Ansicht nach besteht der Weg zum Erfolg nicht darin, ein profundes Spezialwissen zu einem Thema zu haben, sondern sich mit mehreren Themen auseinanderzusetzen und praktische Kenntnisse zu erwerben. Das ist das Ziel dieses Buchs, und Sie werden gerade genug lernen, um angriffslustig zu sein und diese einst so schwer fassbaren kritischen Fragen zu stellen.

Fangen wir also an!