Einleitung

Aus den Nachrichten und den sozialen Medien ist Ihnen vermutlich bekannt, dass das Machine Learning zu einer der spannendsten Technologien der heutigen Zeit geworden ist. Große Unternehmen wie Google, Facebook, Apple, Amazon, IBM und viele andere investieren aus gutem Grund kräftig in die Erforschung des Machine Learnings und dessen Anwendung. Auch wenn man manchmal den Eindruck bekommt, dass »Machine Learning« als leeres Schlagwort gebraucht wird, handelt es sich doch zweifellos nicht um eine Modeerscheinung. Dieses spannende Fachgebiet eröffnet viele neue Möglichkeiten und ist im Alltag schon nicht mehr wegzudenken. Denken Sie an die virtuellen Assistenten von Smartphones, Produktempfehlungen für Kunden in Onlineshops, das Verhindern von Kreditkartenbetrug, Spamfilter in E-Mail-Programmen oder die Erkennung und Diagnose von Krankheitssymptomen – die Liste ließe sich beliebig lang fortsetzen.

Einstieg in Machine Learning

Wenn Sie zu einem Praktiker des Machine Learnings und einem besseren Problemlöser werden möchten oder vielleicht sogar eine Laufbahn in der Erforschung des Machine Learnings anstreben, dann ist dies das richtige Buch für Sie. Für einen Neuling können die dem Machine Learning zugrunde liegenden theoretischen Konzepte zunächst einmal erdrückend wirken. In den vergangenen Jahren sind aber viele praxisorientierte Bücher mit leistungsfähigen Lernalgorithmen erschienen, die Ihnen den Start erleichtern.

Theorie und Praxis

Die Verwendung praxisorientierter Codebeispiele dient einem wichtigen Zweck: Konkrete Beispiele verdeutlichen die allgemeinen Konzepte, indem das Erlernte unmittelbar in die Tat umgesetzt wird. Allerdings darf man dabei nicht vergessen, dass mit großer Macht auch immer große Verantwortung einhergeht! Neben der unmittelbaren Erfahrung, Machine Learning mithilfe der Programmiersprache Python und auf Python beruhenden Lernbibliotheken in die Tat umzusetzen, stellt das Buch auch die den Machine-Learning-Algorithmen zugrunde liegenden mathematischen Konzepte vor, die für den erfolgreichen Einsatz von Machine Learning unverzichtbar sind. Das Buch ist also kein rein praktisch orientiertes Werk, sondern ein Buch, das die erforderlichen Details der Konzepte des Machine Learnings

erörtert, die Funktionsweise von Lernalgorithmen und ihre Verwendung verständlich, aber dennoch informativ erklärt und – was noch wichtiger ist – das zeigt, wie man die häufigsten Fehler vermeidet.

Warum Python?

Bevor wir uns eingehender mit Machine Learning befassen, müssen wir die wichtigste Frage beantworten: Warum Python? Die Antwort ist ganz einfach: Python ist leistungsfähig, aber dennoch sehr leicht erlernbar. Python ist auf dem Gebiet der Data Science zur verbreitetsten Programmiersprache geworden, weil sie es uns ermöglicht, die lästigen Aspekte des Programmierens zu vergessen, und eine Umgebung bereitstellt, in der wir unsere Ideen schnell umsetzen und Konzepte direkt zur Anwendung bringen können.

Erkundung des Fachgebiets Machine Learning

Wenn Sie bei Google Scholar den Suchbegriff *machine learning* eingeben, erhalten Sie als Resultat eine riesige Zahl (ca. 3.250.000) von Treffern. Nun können wir in diesem Buch natürlich nicht sämtliche Einzelheiten der in den letzten 60 Jahren entwickelten Algorithmen und Anwendungen erörtern. Wir werden uns jedoch auf eine spannende Tour begeben, die alle wichtigen Themen und Konzepte umfasst, damit Sie eine gründliche Einführung erhalten. Sollte Ihr Wissensdurst auch nach der Lektüre noch nicht gestillt sein, steht Ihnen eine Vielzahl weiterer hilfreicher Ressourcen zur Verfügung, die Sie nutzen können, um die entscheidenden Fortschritte auf diesem Fachgebiet zu verfolgen.

Wir, die Autoren, können aus eigener Erfahrung sagen, dass wir durch die Beschäftigung mit dem Machine Learning zu besseren Wissenschaftlern, Denkern und Problemlösern geworden sind. In diesem Buch möchten wir unsere diesbezüglichen Erkenntnisse mit Ihnen teilen. Wissen wird durch Lernen erworben, das wiederum einen gewissen Eifer erfordert, und erst Übung macht den sprichwörtlichen Meister.

Der vor Ihnen liegende Weg ist manchmal nicht ganz einfach, und einige der Themenbereiche sind deutlich schwieriger als andere, aber wir hoffen dennoch, dass Sie die Gelegenheit nutzen und sich auf den Lohn der Mühe konzentrieren. Im weiteren Verlauf des Buches werden Sie Ihrem Repertoire eine ganze Reihe leistungsfähiger Techniken hinzufügen können, die dabei helfen, auch die schwierigsten Aufgaben auf datengesteuerte Weise zu bewältigen.

An wen richtet sich das Buch?

Falls Sie sich schon ausführlich mit der Theorie des Machine Learnings beschäftigt haben, zeigt Ihnen dieses Buch, wie Sie Ihre Kenntnisse in die Praxis umsetzen können. Wenn Sie bereits entsprechende Techniken eingesetzt haben, aber

deren Funktionsweise besser verstehen möchten, kommen Sie hier ebenfalls auf Ihre Kosten.

Und wenn Ihnen das Thema Machine Learning noch völlig neu ist, haben Sie umso mehr Grund, sich zu freuen, denn ich kann Ihnen versprechen, dass dieses Verfahren Ihre Denkweise über Ihre in Zukunft zu lösenden Aufgaben verändern wird – und ich möchte Ihnen zeigen, wie Sie Problemstellungen in Angriff nehmen, indem Sie die den Daten innewohnende Kraft freisetzen. Wenn Sie herausfinden möchten, wie Sie Python verwenden können, um die entscheidenden Fragen zu Ihren Daten zu beantworten, greifen Sie einfach zu diesem Buch. Ob Sie völliger Neuling sind oder Ihre Kenntnisse der Data Science vertiefen möchten: Dieses Buch ist eine unentbehrliche Informationsquelle und unbedingt lesenswert.

Zum Inhalt des Buches

Kapitel 1, Wie Computer aus Daten lernen können, führt Sie in die wichtigsten Teilbereiche des Machine Learnings ein, mit denen sich verschiedene Probleme in Angriff nehmen lassen. Darüber hinaus werden die grundlegenden Schritte beim Entwurf eines typischen Machine-Learning-Modells erörtert, auf die wir in den nachfolgenden Kapiteln zurückgreifen.

Kapitel 2, Lernalgorithmen für die Klassifikation trainieren, geht zurück zu den Anfängen des Machine Learnings und stellt binäre Perzeptron-Klassifizierer und adaptive lineare Neuronen vor. Dieses Kapitel ist eine behutsame Einführung in die Grundlagen der Klassifikation von Mustern und konzentriert sich auf das Zusammenspiel von Optimierungsalgorithmen und Machine Learning.

Kapitel 3, Machine-Learning-Klassifikatoren mit scikit-learn verwenden, beschreibt die wichtigsten Klassifikationsalgorithmen des Machine Learnings und stellt praktische Beispiele vor. Dabei kommt eine der beliebtesten und verständlichsten Open-Source-Bibliotheken für Machine Learning zum Einsatz: scikit-learn.

Kapitel 4, Gut geeignete Trainingsdatenmengen: Datenvorverarbeitung, erläutert die Handhabung der gängigsten Probleme unverarbeiteter Datenmengen, wie z.B. fehlende Daten. Außerdem werden verschiedene Ansätze zur Ermittlung der informativsten Merkmale einer Datenmenge vorgestellt. Des Weiteren erfahren Sie, wie sich Variablen unterschiedlichen Typs als geeignete Eingabe für Lernalgorithmen einsetzen lassen.

Kapitel 5, Datenkomprimierung durch Dimensionsreduktion, beschreibt ein wichtiges Verfahren zur Reduzierung der Merkmalsanzahl eines Datenbestands durch Aufteilung in kleinere Mengen unter Beibehaltung eines Großteils der nützlichsten und charakteristischsten Informationen. Hier wird der Standardansatz zur Dimensionsreduktion durch die Analyse der Hauptkomponenten erläutert und mit überwachten und nichtlinearen Transformationsverfahren verglichen.

Kapitel 6, Bewährte Verfahren zur Modellbewertung und Hyperparameter-Optimierung, erörtert die Einschätzung der Aussagekraft von Vorhersagemodellen. Darüber hinaus kommen verschiedene Bewertungskriterien der Modelle sowie Verfahren zur Feinabstimmung der Lernalgorithmen zur Sprache.

Kapitel 7, Kombination verschiedener Modelle für das Ensemble Learning, führt Sie in die verschiedenen Konzepte zur effektiven Kombination diverser Lernalgorithmen ein. Sie erfahren, wie Sie Ensembles einrichten, um die Schwächen einzelner Klassifizierer zu überwinden, was genauere und verlässlichere Vorhersagen liefert.

Kapitel 8, Machine Learning zur Analyse von Stimmungslagen nutzen, erläutert die grundlegenden Schritte zur Transformierung von Textdaten in eine für Lernalgorithmen sinnvolle Form, um so die Meinung von Menschen anhand der von ihnen verfassten Texte vorherzusagen.

Kapitel 9, Einbettung eines Machine-Learning-Modells in eine Webanwendung, führt vor, wie Sie das Lernmodell des vorangehenden Kapitels Schritt für Schritt in eine Webanwendung einbetten können.

Kapitel 10, Vorhersage stetiger Zielvariablen durch Regressionsanalyse, erörtert grundlegende Verfahren zur Modellierung linearer Beziehungen zwischen Zielvariablen und Regressanden, um auch stetige Werte vorhersagen zu können. Nach der Vorstellung der linearen Modelle kommen auch Polynom-Regression und baumbasierte Ansätze zur Sprache.

Kapitel 11, Verwendung von Daten ohne Label: Clusteranalyse, konzentriert sich auf einen anderen Teilbereich des Machine Learnings, nämlich auf das unüberwachte Lernen. Wir werden drei unterschiedlichen Familien von Clustering-Algorithmen zugehörige Verfahren anwenden, um Objektgruppen aufzuspüren, die einen gewissen Ähnlichkeitsgrad aufweisen.

Kapitel 12, Implementierung eines künstlichen neuronalen Netzes, erweitert das in Kapitel 2 vorgestellte Konzept der Gradient-basierten Optimierung, um leistungsfähige, mehrschichtige neuronale Netze in Python zu erstellen, die auf dem verbreiteten Backpropagation-Algorithmus beruhen.

Kapitel 13, Parallelisierung des Trainings neuronaler Netze mit TensorFlow, baut auf den in den vorausgehenden Kapiteln erworbenen Kenntnissen auf, um Ihnen einen praxisorientierten Leitfaden für ein effizienteres Training neuronaler Netze (NN) an die Hand zu geben. Der Schwerpunkt dieses Kapitels liegt dabei auf TensorFlow 2.0, einer quelloffenen Python-Bibliothek, die die Verwendung mehrerer Kerne moderner Grafikprozessoren (GPUs) ermöglicht und die es gestattet, mithilfe von Bausteinen der benutzerfreundlichen Keras-API tiefe NN zu erstellen.

Kapitel 14, Die Funktionsweise von TensorFlow im Detail, stellt die fortgeschritteneren Konzepte und Funktionalitäten von TensorFlow 2.0 vor. TensorFlow ist eine äußerst umfassende und ausgeklügelte Bibliothek. Dieses Kapitel betrachtet die grundle-

genden Konzepte des Kompilierens von Code zu statischen Graphen zwecks schnellerer Berechnung und der Definition trainierbarer Modellparameter. Darüber hinaus kommen Themen wie das Trainieren tiefer NN mithilfe von TensorFlows Keras-API sowie die vorgefertigten Schätzer zur Sprache.

Kapitel 15, Bildklassifikation mit Deep Convolutional Neural Networks, stellt neuronale Netzarchitekturen vor, die bei maschinellem Sehen und der Bilderkennung aufgrund der gegenüber klassischen Ansätzen überlegenen Leistung zu einem neuen Standard geworden sind, nämlich konvolutionale neuronale Netze (*Convolutional Neural Networks*, CNN). Dieses Kapitel zeigt, wie man Faltungsschichten als Merkmalsextraktoren zur Klassifikation von Bildern verwenden kann.

Kapitel 16, Modellierung sequenzieller Daten durch rekurrente neuronale Netze, stellt eine weitere verbreitete neuronale Netzarchitektur für Deep Learning vor, die besonders gut für die Verarbeitung von Text, anderen sequenziellen Daten und Zeitreihen geeignet ist. In diesem Kapitel werden wir verschiedene rekurrente neuronale Netzarchitekturen auf Textdaten anwenden. Als Aufwärmübung betrachten wir zunächst eine Stimmungsanalyse von Filmbewertungen. Anschließend wird erörtert, wie ein rekurrentes NN anhand der Informationen aus Büchern völlig neue Texte erzeugen kann.

Kapitel 17, Synthetisieren neuer Daten mit Generative Adversarial Networks, stellt eine verbreitete Form eines NN vor, das dazu verwendet werden kann, neue, realistisch wirkende Bilder zu erzeugen. Das Kapitel enthält zunächst eine kurze Einführung in Autoencoder, einen bestimmten Typ eines NN, das zur Datenkomprimierung verwendet werden kann. Anschließend wird erläutert, wie man den Decoder-Teil eines Autoencoders mit einem zweiten NN kombiniert, das zwischen echten und erzeugten Bildern unterscheiden kann. Indem Sie zwei NN miteinander wetteifern lassen, werden Sie ein GAN (Generative Adversarial Network) implementieren, das neue Bilder von scheinbar handgeschriebenen Ziffern erzeugen kann. Nachdem die grundlegenden Konzepte von GAN vorgestellt wurden, endet das Kapitel mit einer Beschreibung von Verfahren, die das Training von GAN stabilisieren können, wie beispielsweise die Verwendung der Wasserstein-Metrik als Distanzmaß.

Kapitel 18, Entscheidungsfindung in komplexen Umgebungen per Reinforcement Learning, beschreibt ein Teilgebiet des Machine Learnings, das typischerweise beim Trainieren von Robotern und anderen autonomen System zum Einsatz kommt. Das Kapitel enthält zunächst eine Einführung in Reinforcement Learning (RL), damit Ihnen die Interaktionen von Agenten und Umgebungen, Belohnungssysteme und das Konzept, aus Erfahrungen zu lernen, vertraut sind. Das Kapitel stellt die beiden Hauptkategorien des RL vor, nämlich modellbasierte und modellfreie RL-Systeme. Nachdem Sie grundlegende Ansätze für Algorithmen kennengelernt haben, wie Monte-Carlo-Verfahren und Temporal-Difference-Algorithmen, werden Sie einen Agenten implementieren und trainieren, der sich mithilfe eines Q-Learning-Algo-

rithmus in einer Grid-World-Umgebung bewegt. Abschließend wird ein Deep-Q-Learning-Algorithmus vorgestellt, der eine Variante des Q-Learnings unter Verwendung tiefer NN ist.

Was Sie benötigen

Zum Ausführen der Codebeispiele ist die Python-Version 3.7.0 oder neuer auf macOS, Linux oder Microsoft Windows erforderlich. Wir werden häufig von Python-Bibliotheken Gebrauch machen, die für wissenschaftliche Berechnungen unverzichtbar sind, z.B. von SciPy, NumPy, scikit-learn, Matplotlib und pandas.

Im ersten Kapitel finden Sie Hinweise und Tipps zur Einrichtung Ihrer Python-Umgebung und dieser elementaren Bibliotheken. In den verschiedenen Kapiteln werden wir dann der Python-Umgebung weitere Bibliotheken hinzufügen: die NLTK-Bibliothek für die Verarbeitung natürlicher Sprache (Kapitel 8), das Web-Framework Flask (Kapitel 9) und schließlich TensorFlow, um neuronale Netze effizient auf GPUs zu trainieren (Kapitel 13 bis 18).

Codebeispiele herunterladen

Die Codebeispiele können Sie auf GitHub unter `https://github.com/rasbt/python-machine-learningbook-3rd-edition` oder über die Verlagsseite `http://www.mitp.de/0213` herunterladen. Dort sind auch farbige Abbildungen zu finden.

Konventionen im Buch

In diesem Buch werden verschiedene Textarten verwendet, um zwischen Informationen unterschiedlicher Art zu unterscheiden. Nachstehend finden Sie einige Beispiele und deren Bedeutungen.

Schlüsselwörter oder Code werden im Fließtext wie folgt dargestellt:

»Ein bereits installiertes Paket kann mit der Option `--upgrade` aktualisiert werden.«

Codeblöcke sehen so aus:

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> y = df.iloc[0:100, 4].values
>>> y = np.where(y == 'Iris-setosa', -1, 1)
>>> X = df.iloc[0:100, [0, 2]].values
>>> plt.scatter(X[:50, 0], X[:50, 1],
...             color='red', marker='x', label='setosa')
```

```
>>> plt.scatter(X[50:100, 0], X[50:100, 1],
...             color='blue', marker='o', label='versicolor')
>>> plt.xlabel('Länge des Kelchblatts')
>>> plt.ylabel('Länge des Blütenblatts')
>>> plt.legend(loc='upper left')
>>> plt.show()
```

Benutzereingaben oder Ausgaben auf der Kommandozeile werden in nicht proportionaler Schrift gedruckt:

```
> dot -Tpng tree.dot -o tree.png
```

Neue Ausdrücke und *wichtige Begriffe* werden kursiv gedruckt. Auf dem Bildschirm auswählbare oder anklickbare Bezeichnungen, wie z.B. Menüpunkte oder Schaltflächen, werden in der Schriftart KAPITÄLCHEN gedruckt: »Nach einem Klick auf die Schaltfläche ABBRECHEN in der unteren rechten Ecke wird der Vorgang abgebrochen.«

Hinweis

Warnungen oder Hinweise erscheinen in einem Kasten wie diesem.

Tipp

Und so werden Tipps und Tricks dargestellt.