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 aus dem 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, die Erkennung und Diagnose von Krankheitssymptomen – die Liste ließe sich beliebig lang fortsetzen.

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.

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.

Wenn Sie bei Google Scholar den Suchbegriff *machine learning* eingeben, erhalten Sie als Resultat eine riesige Zahl (ca. 1.800.000) 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.

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 können, indem Sie die den Daten innewohnende Kraft freisetzen.

Bevor wir uns eingehender mit dem Machine Learning befassen, soll aber zunächst noch Ihre vermutlich vordringlichste Frage beantwortet werden: Warum Python? Die Antwort ist einfach: Es ist leistungsfähig und doch leicht zu erlernen. Python ist zur beliebtesten Programmiersprache im Bereich Data Science geworden, weil man sich die lästigen Teile bei der Programmierung erspart und eine Umgebung bereitsteht, in der sich Ideen und Konzepte sofort umsetzen lassen.

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, was 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.

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 Klassifizierung 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 Klassifizierung von Mustern und konzentriert sich auf das Zusammenspiel von Optimierungsalgorithmen und Machine Learning.

Kapitel 3, Machine-Learning-Klassifizierer mit scikit-learn verwenden, beschreibt die wichtigsten Klassifizierungsalgorithmen 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-Abstimmung, 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 Vor-

stellung der linearen Modelle kommen auch Polynom-Regression und baumbasierte Ansätze zur Sprache.

Kapitel 11, Verwendung nicht gekennzeichneter Daten: 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 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 an die Hand zu geben. Der Schwerpunkt dieses Kapitels liegt dabei auf TensorFlow, einer quelloffenen Python-Bibliothek, die die Verwendung mehrerer Kerne moderner Grafikprozessoren ermöglicht.

Kapitel 14, Die Funktionsweise von TensorFlow im Detail, behandelt TensorFlow ausführlicher und erläutert die grundlegenden Konzepte von Berechnungsgraphen und Sitzungen. Darüber hinaus kommen Themen wie das Abspeichern und Visualisieren der Graphen neuronaler Netze zur Sprache, was sich im verbleibenden Teil des Buches als sehr nützlich erweisen wird.

Kapitel 15, Bildklassifizierung mit tiefen konvolutionalen neuronalen Netzen, stellt neuronale Netzarchitekturen vor, die bei maschinellem Sehen und der Bilderkennung zu einem neuen Standard geworden sind, nämlich konvolutionale neuronale Netze. Dieses Kapitel erörtert die grundlegenden Konzepte konvolutionaler Schichten als Merkmalsextraktoren und zeigt die Anwendung einer konvolutionalen neuronalen Netzarchitektur zur Klassifizierung von Bildern, die eine nahezu perfekte Klassifizierung erzielt.

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 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 und erzeugen anschließend völlig neue Texte.

Was Sie benötigen

Zum Ausführen der Codebeispiele ist die Python-Version 3.6.0 oder neuer auf macOS, Linux oder Microsoft Windows erforderlich. Wir werden häufig von für

wissenschaftliche Berechnungen unverzichtbaren Python-Bibliotheken Gebrauch machen, 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), die Seaborn-Bibliothek zur Visualisierung statistischer Daten (Kapitel 10) und schließlich TensorFlow, um neuronale Netze effizient mit grafischen Symbolen zu trainieren.

Für wen ist das Buch gedacht?

Wenn Sie wissen möchten, wie Sie Python einsetzen können, um wichtige Fragen über Ihre Daten zu beantworten, sind Sie hier genau richtig. Ob Sie nun Anfänger sind oder Ihre Kenntnisse auf dem Gebiet der Data Science vertiefen möchten: Dieses Buch ist eine unentbehrliche Informationsquelle und unbedingt lesenswert.

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, Datenbanktabellen-, Twitter-, Datei-, Ordner-, Datei- und Pfadnamen sowie URLs und Usereingaben werden im Fließtext wie folgt dargestellt:

»Durch die Einstellung `out_file=None` weisen wir die Daten direkt der Variablen `doc_data` zu, ohne sie erst in eine temporäre Datei `tree.dot` auf die Festplatte zu schreiben.«

Codeblöcke sehen so aus:

```
>>> from sklearn.neighbors import KNeighborsClassifier
>>> knn = KNeighborsClassifier(n_neighbors=5, p=2,
...                            metric='minkowski')
>>> knn.fit(X_train_std, y_train)
>>> plot_decision_regions(X_combined_std, y_combined,
...                       classifier=knn, test_idx=range(105,150))
>>> plt.xlabel('petal length [standardized]')
>>> plt.ylabel('petal width [standardized]')
>>> plt.show()
```

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

```
pip3 install graphviz
```

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.«

Vorsicht

Warnungen oder wichtige Hinweise erscheinen in einem Kasten wie diesem.

Tipp

Und so werden Tipps und Tricks dargestellt.

Codebeispiele herunterladen

Die Codebeispiele zum Buch finden Sie auf GitHub unter `https://github.com/rasbt/python-machine-learning-book-2nd-edition`.

Farbige Abbildungen

Alle in diesem Buch verwendeten Diagramme und Screenshots stehen unter `www.mitp.de/733` zusätzlich in einer farbigen Variante zum Download zur Verfügung.