



Uli SCHELL

# MASCHINELLES LERNEN MIT R

Daten aufbereiten  
und verarbeiten  
mit H2O und Keras



Beispiele zum Download:  
[plus.hanser-fachbuch.de](https://plus.hanser-fachbuch.de)

HANSER

HANSER

Uli Schell

# **Maschinelles Lernen mit R**

Daten aufbereiten und verarbeiten mit H2O und Keras

## **Ihr Plus – digitale Zusatzinhalte!**

Auf unserem Download-Portal  
finden Sie zu diesem Titel  
kostenloses Zusatzmaterial.

Geben Sie auf [plus.hanser-  
fachbuch.de](https://plus.hanser-fachbuch.de) einfach diesen  
Code ein:

**plus-xNea5-U69cd**

Der Autor:

*Prof. Dr. Uli Schell*, Hochschule Kaiserslautern

Kontakt: [uli.schell@hs-kl.de](mailto:uli.schell@hs-kl.de)

Alle in diesem Buch enthaltenen Informationen, Verfahren und Darstellungen wurden nach bestem Wissen zusammengestellt und mit Sorgfalt getestet. Dennoch sind Fehler nicht ganz auszuschließen. Aus diesem Grund sind die im vorliegenden Buch enthaltenen Informationen mit keiner Verpflichtung oder Garantie irgendeiner Art verbunden. Autor und Verlag übernehmen infolgedessen keine juristische Verantwortung und werden keine daraus folgende oder sonstige Haftung übernehmen, die auf irgendeine Art aus der Benutzung dieser Informationen - oder Teilen davon - entsteht. Ebenso übernehmen Autor und Verlag keine Gewähr dafür, dass beschriebene Verfahren usw. frei von Schutzrechten Dritter sind. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Buch berechtigt deshalb auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

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.

Dieses Werk ist urheberrechtlich geschützt.

Alle Rechte, auch die der Übersetzung, des Nachdruckes und der Vervielfältigung des Buches, oder Teilen daraus, vorbehalten. Kein Teil des Werkes darf ohne schriftliche Genehmigung des Verlages in irgendeiner Form (Fotokopie, Mikrofilm oder ein anderes Verfahren) - auch nicht für Zwecke der Unterrichtsgestaltung - reproduziert oder unter Verwendung elektronischer Systeme verarbeitet, vervielfältigt oder verbreitet werden.

© 2022 Carl Hanser Verlag München

[www.hanser-fachbuch.de](http://www.hanser-fachbuch.de)

Lektorat: Sylvia Hasselbach

Copy editing: Jürgen Dubau, Freiburg/Elbe

Layout: le-tex publishing services GmbH

Umschlagdesign: Marc Müller-Bremer, [www.rebranding.de](http://www.rebranding.de), München

Umschlagrealisation: Max Kostopoulos, unter Verwendung von Grafiken von

© shutterstock.com/Taleseedum

Datenbelichtung, Druck und Bindung: Eberl & Koesel GmbH & Co. KG, Altusried-Krugzell

Ausstattung patentrechtlich geschützt. Kösel FD 351, Patent-Nr. 0748702

Print-ISBN: 978-3-446-47165-8

E-Book-ISBN: 978-3-446-47244-0

ePub-ISBN: 978-3-446-47323-2

# Inhalt

## Titelei

## Impressum

## Inhalt

## Vorwort

## Zusatzmaterial online

## Teil I Einstieg

### 1 Einleitung

#### 1.1 Informatik und künstliche Intelligenz

#### 1.2 Expertensysteme

## **1.3 Maschinelles Lernen**

### **1.3.1 Überwachtes Lernen**

### **1.3.2 Unüberwachtes Lernen**

### **1.3.3 Verstärkendes Lernen**

## **1.4 Methoden und Werkzeuge für das maschinelle Lernen**

## **1.5 Zielsetzungen dieses Buchs und Vorgehensweise**

# **2 Einführung in R und RStudio**

## **2.1 Installation unter Windows**

## **2.2 Installation unter Ubuntu Linux**

## **2.3 Die RStudio-Oberfläche**

### **2.3.1 Das Konsolenfenster**

### **2.3.2 Das Source-Fenster**

### **2.3.3 Das Files-/Packages-Fenster**

#### **2.3.3.1 Die Files-Registerkarte**

#### **2.3.3.2 Die Plot-Registerkarte**

[2.3.3.3 Die Packages-Registerkarte](#)

[2.3.3.4 Das Environment-Fenster](#)

[2.3.4 Daten einlesen und speichern](#)

## [\*\*2.4 Zuweisungen, Variablen und elementare Datentypen\*\*](#)

## [\*\*2.5 Zusammengesetzte Datentypen\*\*](#)

[2.5.1 Vektoren](#)

[2.5.1.1 Union, intersect, setdiff](#)

[2.5.1.2 Umwandeln des Datentyps, Faktorisierung](#)

[2.5.2 Matrizen](#)

[2.5.3 Frames](#)

[2.5.3.1 Hinzufügen von Spalten](#)

[2.5.3.2 Hinzufügen von Zeilen](#)

[2.5.3.3 Auswahl von Spalten](#)

[2.5.3.4 Auswahl von Zeilen und Spalten](#)

[2.5.3.5 Löschen von Spalten oder Zeilen](#)

[2.5.4 tibbles](#)



2.5.5 Listen

2.5.6 Zeichenketten

2.5.7 Datum und Zeit

## **2.6 Programmablaufsteuerung**

2.6.1 Blöcke

2.6.2 Bedingte Ausführung

2.6.3 Schleifen

2.6.4 Apply()

2.6.5 Gefahren und Benachrichtigungen

2.6.6 Funktionen

2.6.7 Der Pipe-Operator

## **2.7 Debugging**

## **2.8 Programmierstil**

2.8.1 Variablen- und Funktionsbenennung

2.8.2 Abstände

2.8.3 Blöcke

[2.8.4 Abschnitte](#)

## [2.9 Formatierhilfen](#)

[2.9.1 Styler](#)

[2.9.2 Lintr](#)

## [2.10 Zum Nachschlagen](#)

## [2.11 Einstiegsaufgabe](#)

# [Teil II Datenvorbereitung](#)

## [3 Daten visualisieren und vorbereiten](#)

### [3.1 Plots mit ggplot2](#)

[3.1.1 Plot-Architekturen](#)

[3.1.2 Erster Plot mit ggplot2](#)

[3.1.3 Aufbau eines Plots](#)

[3.1.4 Ebenen](#)

[3.1.5 Geome](#)

[3.1.6 Stats](#)

### 3.1.7 Themes

## 3.2 Datenaufbereitung

## 3.3 Strukturierung der Daten

### 3.3.1 Werte in Spaltenüberschriften

### 3.3.2 Mehrere Werte in einer Spalte

### 3.3.3 Imputation und Grafikdarstellung

### 3.3.4 Überwachung von Datentypen/Bereichsgrenzen

## 3.4 Datenaufbereitung an Zeitreihen – ein Beispiel

### 3.4.1 Erster Überblick

### 3.4.2 Messfehler

### 3.4.3 Überprüfen der zeitlichen Abfolge

### 3.4.4 Zeitserie generieren

#### 3.4.4.1 Bestimmung der mittleren Tagestemperatur

#### 3.4.4.2 Bestimmung der mittleren Jahrestemperatur

#### 3.4.4.3 Imputation – Ersetzen von NA-Werten

## 3.5 Zwei kleine Checklisten

## 3.6 Aufgaben

# 4 Datenplausibilität

## 4.1 Hypothesen-Betrachtung

## 4.2 Allgemeine Kenngrößen als Hilfsmittel

### 4.2.1 Mittelwert und Median

### 4.2.2 Varianz

### 4.2.3 Momente höherer Ordnung

## 4.3 Grafische Hilfsmittel

### 4.3.1 Histogramme

### 4.3.2 Box-Plots

### 4.3.3 Summenhäufigkeiten und QQ-Plots

## 4.4 R-Packages zum Erkennen von Ausreißern

### 4.4.1 Das Package „Outliers“

#### 4.4.1.1 scores()

#### 4.4.1.2 Dixon()

#### 4.4.1.3 Cochran.test()

4.4.1.4 Grubbs-Test

4.4.1.5 Weitere Funktionen

## 4.5 Datenplausibilität bei mehreren Variablen und weitere Funktionen

4.6 Aufgaben

# Teil III Statistische Lernmodelle

## 5 Regression

### 5.1 Regression mit einer unabhängigen Variablen

5.1.1 Methode der kleinsten Fehlerquadrate

5.1.2 Homoskedastizität

5.1.3 Modellvalidierung

5.1.3.1 Residuen

5.1.4 Vorbeugende Wartung

5.1.4.1 Das Modellbeispiel

5.1.4.2 Etwas mehr Residuenanalyse

5.1.4.3 Vorhersagen

5.1.5 Erweiterung der Regression auf nichtlineare Funktionen

5.1.6 Kreuzvalidierung

## 5.2 Regression mit mehreren unabhängigen Variablen

5.2.1 Der Boston-Datensatz

5.2.2 Durchführung der Regression

5.2.3 Modellvalidierung

5.2.4 Regressionen robuster machen

5.2.4.1 Regularisierung

5.2.4.2 M-Schätzer

5.2.4.3 Weitere Alternativen

## 5.3 Aufgaben

# 6 Klassifikation

6.1 Logistische Regression

6.2 Der Perceptron-Algorithmus

6.3 Support Vector Machines

6.3.1 Der Iris-Datensatz

## **6.4 Entscheidungsbaumverfahren**

6.4.1 Entscheidungsbäume

6.4.2 Der Iris-Datensatz

6.4.3 Bagging

6.4.4 Random Forests

6.4.5 Boosted Regression Trees

## **6.5 Naive Bayes-Klassifikatoren**

6.5.1 Multinomiale naive Bayes-Klassifikatoren

6.5.2 Das spam-Beispiel

6.5.3 Likelihood

6.5.4 Gauß'sche naive Bayes-Klassifizierer

## **6.6 KNN Nächste-Nachbarn-Klassifikation**

## **6.7 Modellbewertung: Devianzen und universellere Kenngrößen**

6.7.1 Receiver Operating Characteristic, ROC und AUC

6.7.2 Die  $R^2$ -Metrik

6.7.3 Eine generalisierte Metrik

6.7.3.1 Pseudo- $R^2$

6.7.3.2 Devianzen

6.8 Aufgaben

## 7 Objekte clustern, Merkmale reduzieren und Zeitreihen zerlegen

7.1 K-means-Clustering

7.2 Korrelationen und Merkmalsreduktion durch Hauptkomponentenanalyse

7.2.1 Der beste Standpunkt

7.2.2 Kovarianzen

7.2.3 Kovarianzmatrizen

7.2.4 Anwendungen

7.2.4.1 Eine kleine Weinprobe

7.2.4.2 Der Boston-Datensatz

7.2.4.3 Erweiterungen der Hauptkomponentenanalyse



7.2.4.4 Ausreißerererkennung

## **7.3 Zeitreihen**

7.3.1 Komponentenmodelle

7.3.2 Glättungsverfahren bei Zeitreihen

7.3.2.1 Gleitender Durchschnitt

7.3.2.2 Exponentielle Glättung

7.3.2.3 Holt-Winters-Glättung

7.3.2.4 Weitere Glättungsmethoden

7.3.3 AR-Modelle

7.3.3.1 Autokorrelationsfunktionen

7.3.3.2 Partielle Autokorrelationsfunktion

7.3.4 MA-Modelle

7.3.5 ARMA- und ARIMA-Modelle

## **7.4 Aufgaben**

# **Teil IV Lernen mit neuronalen Netzen**

## **8 Neuronale Netze**

## 8.1 Das Perceptron

## 8.2 Layers

## 8.3 Aktivierungsfunktionen

## 8.4 Warum das Stapeln linearer Funktionen nicht sinnvoll ist

## 8.5 Der Regelkreis eines Lernvorgangs

### 8.5.1 Verlustfunktionen und Metriken

### 8.5.2 Epochen und Batches

### 8.5.3 Gradienten, lokale Gradienten und Autodiff

#### 8.5.3.1 Manuelle Bestimmung des Gradienten eines Perceptrons

#### 8.5.3.2 Berechnung des Gradienten mit CAS oder Differenzenquotienten

#### 8.5.3.3 Berechnung des Gradienten mit Forward-Mode Autodiff

#### 8.5.3.4 Berechnung des Gradienten mit Backpropagation Autodiff

### 8.5.4 Der Optimizer

8.5.4.1 Stochastisches Gradientenverfahren (SGD)

8.5.4.2 Momenten-Update

8.5.4.3 Nesterov-Moment

8.5.4.4 Adagrad

8.5.4.5 Rmsprop

8.5.4.6 Adadelta

8.5.4.7 Adam

8.5.4.8 ADAMax

8.5.4.9 Was tun, wenn ...?

## 8.6 Regularisierung und Dropouts

8.6.1 Regularisierung

8.6.2 Dropouts

## 8.7 Aufgaben

# 9 H2O

9.1 Das Unternehmen H2O

9.2 Installation und erste Schritte

## **9.3 Univariate lineare Regression**

### **9.3.1 Generalisierte lineare Modelle**

#### **9.3.1.1 Lambda-Suche**

#### **9.3.1.2 Grid-Suche**

## **9.4 Entscheidungsbäume, Random Forests und Gradient Boosting**

## **9.5 Neuronale Netze**

## **9.6 Der Boston-Datensatz**

### **9.6.1 AutoML**

### **9.6.2 Explain**

#### **9.6.2.1 Residuenanalyse**

#### **9.6.2.2 Variablenwichtigkeit**

#### **9.6.2.3 Heatmap der Variablenwichtigkeit**

#### **9.6.2.4 Modell-Korrelation**

#### **9.6.2.5 Partielles Abhängigkeitsdiagramm**

## **9.7 Iris**

9.7.1 Stacked Ensemble

## **9.8 MNIST**

9.8.1 Der Standarddatensatz

9.8.2 Bewertung

## **9.9 Aufgaben**

# **10 Keras/Tensorflow**

## **10.1 Einrichtung und Nutzung von Keras**

10.1.1 Dimensionen und Tensoren

10.1.2 Normalisierung

## **10.2 Boston**

10.2.1 Normalisierung

10.2.2 Modelldefinition

10.2.3 Compilierung

10.2.4 Fit

## **10.3 Diagnosemöglichkeiten und Optimierung**

[10.3.1 Eine kleine Regression](#)

[10.3.2 Speichern und Laden des Modells und der Gewichte](#)

[10.3.3 Auslesen des Modells](#)

[10.3.4 Callbacks](#)

[10.3.5 TensorBoard](#)

## **[10.4 Convolutional Networks](#)**

[10.4.1 Faltung und Feature Learning](#)

[10.4.1.1 Diskrete 2D-Faltung](#)

[10.4.1.2 Aufbau von Mustern](#)

[10.4.1.3 Pooling Layers](#)

[10.4.2 Komposition der Layer](#)

[10.4.3 MNIST](#)

[10.4.4 ImageNet](#)

## **[10.5 Transferlernen](#)**

[10.5.1 Modelldefinition](#)

[10.5.2 Datenbereitstellung](#)

[10.5.3 Augmentation](#)

10.5.4 Lernvorgänge

## **10.6 Recurrent Networks**

10.6.1 Simple Recurrent Networks

10.6.2 LSTM

10.6.3 Wettervorhersage

## **10.7 Aufgaben**

# **Teil V Anhang**

## **A Basiswissen Statistik**

### **A.1 Beschreibende Statistik**

A.1.1 Mittelwert und Median

A.1.2 Varianz

A.1.3 Momente höherer Ordnung

A.1.4 Histogramme, Summenhäufigkeiten und Quantile

A.1.5 Box-Plots

### **A.2 Schließende Statistik**

A.2.1 Normalverteilung

A.2.2 t-(Student-)Verteilung

A.2.3 Chi-Quadrat-Verteilung

A.2.4 F-Verteilung

Literatur



# Vorwort

Künstliche Intelligenz ist einer der Haupt-Innovationstreiber der Gegenwart. Hätten Sie noch vor wenigen Jahren damit gerechnet, dass 2021 das erste Gesetz zum autonomen Fahren in Deutschland in Kraft treten würde? Ohne künstliche Intelligenz und maschinelles Lernen wäre das nicht möglich geworden.

Maschinelles Lernen – dieser Begriff weckt vielfältige Assoziationen. Zum einen erhofft man sich, dass damit der Alltag bequemer und sicherer gemacht wird. Zum anderen befürchtet man, dass immer noch intelligentere Geräte immer autonomer agieren. Vielleicht machen sie sich irgendwann sogar noch selbständig und stellen Unfug an? Um das besser beurteilen zu können, lohnt sich die Beschäftigung mit diesem Thema.

Und da liegt wahrscheinlich das eigentliche Problem: Es gibt noch wenig Einstiegsliteratur in das Thema „Maschinelles Lernen“, aber die technologischen Entwicklungen nehmen keine Rücksicht darauf. So kommt die Angst auf, dass man diesen rasanten Entwicklungen nicht mehr folgen kann, man fühlt sich „abgehängt“ und verschließt womöglich noch die Augen davor.

Ich denke, das muss nicht sein. Sie sehen das wahrscheinlich auch so, denn Sie haben sich dazu entschlossen, in dieses Thema einzusteigen, weil Sie gerade dieses Buch lesen.

Es war mir ein Ansporn, Ihnen diesen Weg mit vielen Beispielen zu ebneten und Ihnen möglichst viele Möglichkeiten einer Vertiefung zu bieten.

Allen Gesprächspartnern, die mich hierbei mit vielen guten Anregungen unterstützt haben, möchte ich an dieser Stelle sehr herzlich danken, insbesondere Frau Sylvia Hasselbach, Frau Irene Weilhart, Herrn Dr. Jochen Hirschle und Herrn Jürgen Dubau.

Trotz großer Sorgfalt lässt sich der eine oder andere Fehler nicht vermeiden. Wenn Sie also Kritik, Anmerkungen oder auch Wünsche haben, bitte ich Sie um eine Mail an [uli.schell@gmx.de](mailto:uli.schell@gmx.de), damit ich mich darum kümmern kann.

Ich wünsche Ihnen, dass Ihnen der Einstieg in das maschinelle Lernen gut gelingt, dass dieses Buch einen Beitrag dazu leisten konnte, und vor allem: dass Sie dabei Ihre Freude haben!

*Uli Schell*, Dezember 2021

# Zusatzmaterial online

Auf unserem Download-Portal finden Sie zu diesem Titel:

- die Codebeispiele aus dem Buch zum direkten Aufruf
- die Daten zu den Buchbeispielen und Aufgaben
- den originalen MNIST-Datensatz

Geben Sie auf

<https://plus.hanser-fachbuch.de>

diesen Code ein:

plus-xNea5-U69cd

**TEIL I**

**Einstieg**

# 1 Einleitung



## Fragen, die dieses Kapitel beantwortet:

- Wie können wir künstliche Intelligenz beschreiben?
- Was ist Lernen?
- Was ist maschinelles Lernen?
- Wozu Statistik?
- Warum R und nicht Python?

In diesem Kapitel werden die Begriffe künstliche Intelligenz, Expertensysteme und maschinelles Lernen gegeneinander abgegrenzt. Weiterhin werden die Zielsetzungen und die Vorgehensweise in diesem Buch erläutert.

## 1.1 Informatik und künstliche Intelligenz

Seit vielen Jahrhunderten bemüht man sich, Berechnungsmethoden so zu beschreiben, dass sie fehlerfrei zu einem Ergebnis führen:

- Schon vor 2300 Jahren hatte Euklid bereits ein Verfahren zum Bestimmen des größten gemeinsamen Teilers beschrieben.
- Al-Khwarizmi hatte vor über 1200 Jahren Rechenregeln zum Lösen quadratischer Gleichungen formuliert. Sein Name stand Pate zum Begriff des Algorithmus [Bau09].
- Die Arbeiten von Leibniz zum Dualsystem zum Ende des 17. Jahrhunderts bildete die Basis heutiger Rechenanlagen [Wik21c].

Eine Vielzahl von weiteren Ansätzen waren ihrer Zeit weit voraus, konnten aber wegen fehlender Umsetzungsmöglichkeiten nicht realisiert werden wie z. B. die Analytical Engine von Babbage [Wik21b].

Als in der Mitte des letzten Jahrhunderts erste elektronische Anlagen zur Durchführung der Informationsverarbeitung zur Verfügung standen, war eine neue Epoche eingeläutet: Der Begriff „Informatik“ kam im Jahr 1957 auf [Wik21d].

Bereits 14 Jahre zuvor erschien ein Artikel von McCulloch und Pitts [MP43], in dem Betrachtungen zu nervlichen Aktivitäten und deren Logik angestellt werden. Dieser Artikel gilt als erster Beitrag zum Thema „künstliche Intelligenz“. Als Geburtsstunde der künstlichen Intelligenz (KI oder AI, Artificial Intelligence) wird eine Konferenz angesehen, die im Jahre 1956 am Dartmouth College in Hanover, New Hampshire (USA), stattfand [RN12]. John McCarthy, der Organisator der Konferenz, gilt als Gründungsvater dieser neuen Disziplin.