Thomas Bartz-Beielstein
Eva Bartz *Hrsg.*

Online Machine Learning

Eine praxisorientierte Einführung

Springer Vieweg

Online Machine Learning

Thomas Bartz-Beielstein · Eva Bartz
(Hrsg.)

Online Machine Learning

Eine praxisorientierte Einführung

Springer Vieweg

Hrsg.
Thomas Bartz-Beielstein
Institute for Data Science, Engineering, and Analytics
TH Köln
Gummersbach, Deutschland

Eva Bartz
Bartz & Bartz GmbH
Gummersbach, Deutschland

ISBN 978-3-658-42504-3 ISBN 978-3-658-42505-0 (eBook)
https://doi.org/10.1007/978-3-658-42505-0

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

Planung/Lektorat: David Imgrund
Springer Vieweg ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Das Papier dieses Produkts ist recycelbar.

Vorwort

Dieses Buch beschäftigt sich mit dem spannenden, zukunftsträchtigen Thema des Online Machine Learning (OML). Es gliedert sich in drei Teile:

Zunächst beschäftigen wir uns ausführlich mit den theoretischen Grundlagen von OML. Wir beschreiben, was OML ist, und fragen, wie man es mit Batch Machine Learning (BML) vergleichen kann und welche Kriterien für einen aussagekräftigen Vergleich man entwickeln sollte. Im zweiten Teil stellen wir Überlegungen zur Praxis an und belegen diese im dritten Teil mit konkreten praktischen Anwendungen.

Warum OML? Es geht unter anderem um den entscheidenden Zeitvorteil. Das können Monate, Wochen, Tage, Stunden oder auch nur Sekunden sein. Dieser Zeitvorteil kann entstehen, wenn die Künstliche Intelligenz (KI) Daten fortlaufend, also online auswerten kann. Sie nicht darauf warten muss, bis ein kompletter Datensatz zur Verfügung steht, sondern bereits eine einzelne Beobachtung zur Aktualisierung des Modells verwendet werden kann. Hat OML noch andere Vorteile außer dem offensichtlichen Zeitvorteil? Wenn ja, welche? Wir fragen: Gibt es Grenzen des BML, die OML überwindet?

Es muss genau untersucht werden, zu welchem Preis man sich diese Vorteile durch OML verschafft. Wie hoch ist der Speicherbedarf im Vergleich zu herkömmlichen Verfahren? Speicherbedarf bedeutet auch finanzielle Kosten, z. B. durch höheren Energiebedarf. Ist OML eventuell energiesparend und damit nachhaltiger, also Green IT? Erhält man vergleichbar gute Ergebnisse? Leidet die Güte (Performanz), werden die Ergebnisse ungenauer? Um diese Fragen verlässlich zu beantworten, geben wir im Theorieteil zunächst eine verständliche Einführung in OML, die sich sowohl für Anfänger als auch für Fortgeschrittene eignet. Dann begründen wir die von uns gefundenen Kriterien, die wir für die Vergleichbarkeit von OML und BML heranziehen, nämlich eine gut nachvollziehbare Darstellung von Güte, Zeit- und Speicherbedarf.

Im zweiten Teil beschäftigen wir uns mit der Frage, genau wie OML in der Praxis eingesetzt werden kann. Zu Wort kommen Experten aus der Praxis, die von den Anforderungen an die amtliche Statistik berichten. Wir begründen Empfehlungen für den praktischen Einsatz von OML. Wir stellen umfassend die Softwarepakete vor, die derzeit

für OML zur Verfügung stehen, insbesondere „river“[1], und bieten mit Sequential Parameter Optimization Toolbox for River (spotRiver) eine von uns eigens für OML entwickelte Software an. Wir beschäftigen uns ausführlich mit den besonderen Problemen, die bei Datenströmen auftreten können. Hier sei das für Datenströme zentrale Problem der Drift genannt. Wir behandeln die Erklärbarkeit von KI-Modellen, die Interpretierbarkeit und Reproduzierbarkeit. Diese Aspekte können zu höherer Akzeptanz von KI beitragen, wie sie in kommenden Regulierungen für KI-Systeme gefordert wird.

Im Anwendungsteil präsentieren wir zwei ausführliche Studien, davon eine mit einem großen Datensatz mit einer Million Daten. Wir belegen, wann OML besser funktioniert als BML. Besonders interessant ist die Studie zum Hyperparameter-Tuning von OML. Hier zeigen wir, wie OML durch die Optimierung von Hyperparametern deutlich besser funktionieren kann.

Notebooks
Ergänzender Programmcode zu den Anwendungen und Beispielen aus diesem Buch ist in sogenannten „Jupyter-Notebooks“ im GitHub-Repository https://github.com/sn-code-inside/online-machine-learning zu finden. Die Notebooks sind kapitelweise organisiert.

Die Unternehmensberatung Bartz & Bartz GmbH[2] hat den Grundstein für dieses Buch gelegt, als sie 2022 einen Auftrag aus einer Ausschreibung des Statistischen Bundesamtes zugeschlagen bekam[3]. Das Statistische Bundesamt wollte wissen, ob es für den Schatz an Daten und die Auswertung im öffentlichen Auftrag Sinn macht, OML jetzt schon einzusetzen (siehe hierzu die Ausführungen in Kap. 7). Unser leicht ernüchterndes Ergebnis der Expertise war: Es eröffnen sich interessante Perspektiven für die Zukunft, aber momentan bietet sich ein Einsatz noch nicht unmittelbar an. Teils gibt es in der Praxis fachliche und organisatorische Hürden, Prozesse so anzupassen, dass die Vorteile von OML wirklich zur Geltung kommen können. Teils sind OML-Verfahren und Implementierungen noch nicht ausgereift genug.

Das Thema hat uns so fasziniert, dass wir uns entschlossen haben, es weiter zu verfolgen. Prof. Dr. Thomas Bartz-Beielstein hat die Frage nach der Praxisrelevanz von OML mit in die TH Köln genommen und dort seine seit Jahren laufende Forschung zu dem Bereich weiter voran getrieben. Die Forschergruppe am Institut für Data Science, Engineering, and Analytics (IDE+A)[4] konnte unter seiner Anleitung Software so weit entwickeln, dass wir glauben, die Tauglichkeit ein ganzes Stück voran gebracht zu haben. So haben

[1] https://riverml.xyz/
[2] https://bartzundbartz.de
[3] https://destatis.de
[4] https://www.th-koeln.de/idea

wir die damals entstandene Expertise der Bartz & Bartz GmbH mit der Forschung an der TH Köln kombiniert, woraus dieses Buch entstanden ist.

Insgesamt eignet sich das Buch gleichermaßen als Handbuch zum Nachschlagen für Experten, die sich mit OML beschäftigen, als Lehrbuch für Anfänger, die sich mit OML beschäftigen wollen, und als wissenschaftliche Publikation für Wissenschaftler, die sich mit OML beschäftigen, da es den neuesten Stand der Forschung wiedergibt. Es kann aber auch quasi als OML-Consulting dienen, denn Entscheider und Praktiker können anhand unserer Ausführungen OML maßgeschneidert für ihre Bedürfnisse anpassen, für ihre Anwendung einsetzen und fragen, ob die Vorteile von OML eventuell die Kosten aufwiegen.

Um nur einige Beispiele aus der militärischen und zivilen Praxis zu nennen:

- Sie verwenden hochmoderne Sensorsysteme, um Hochwasser vorauszusagen. Hier kann eine schnellere Vorhersage Leben retten.
- Sie müssen terroristische Angriffe abwehren und setzen dazu Unterwassersensorik ein. Hier kann es entscheidend sein, dass die KI schneller „erkennt“, ob es sich um harmlose Wassersportler handelt.
- Sie sind verantwortlich für die Beobachtung des Luftraums. Aufklärungsdrohnen können beispielsweise effizienter eingesetzt werden, wenn sie mit ganz aktuellen KI-Datenauswertungen programmiert und trainiert werden können.
- Sie müssen sehr zügig die Produktion von Gütern der kritischen Infrastruktur, wie Impfstoff, Schutzkleidung oder medizinische Apparaturen, anpassen. Hier kann es sinnvoll sein, den gesamten Produktionsablauf samt einzusetzender Rohstoffe so aktuell wie möglich zu gestalten. Dazu kann eine Echtzeitauswertung und Übersetzung in Bedarfe anhand der Krankenhausbettenbelegung oder Krankschreibungen dienen.
- Sie müssen als Zahlungsdienstleister Betrugsversuche quasi in Echtzeit erkennen.

Abschließend stellen wir fest: OML wird bald praxistauglich, es lohnt sich, sich jetzt schon damit zu beschäftigen. In diesem Buch werden schon einige Werkzeuge vorgestellt, die in Zukunft die Praxis von OML erleichtern werden. Ein vielversprechender Durchbruch steht zu erwarten, weil die Praxis zeigt, dass aufgrund der großen Datenmengen, die in der Praxis anfallen, das bisherige BML nicht mehr ausreicht. OML ist die Lösung, um die Datenströme in Echtzeit auszuwerten, zu verarbeiten und Ergebnisse zu liefern, die für die Praxis relevant sind.

Im Einzelnen behandelt das Buch folgende Themen: Kapitel 1 beschreibt die Motivation für dieses Buch und die Zielsetzung. Es beschreibt die Nachteile und Grenzen von BML und die Notwendigkeit für OML. Kapitel 2 gibt eine Übersicht und Bewertung von Verfahren und Algorithmen mit speziellem Fokus auf Supervised Learning (Klassifikation und Regression). Kapitel 3 beschreibt Verfahren zur Drifterkennung. Die Aktualisierbarkeit der OML-Verfahren wird in Kap. 4 behandelt. Kapitel 5 erläutert Verfahren zur Bewertung von OML-Verfahren. Kapitel 6 beschäftigt sich mit den besonderen

Anforderungen an OML. Mögliche OML-Anwendungen werden in Kap. 7 dargestellt und von Experten der amtlichen Statistik beurteilt. Die Verfügbarkeit der Algorithmen in Softwarepaketen, im Speziellen für R und Python, wird in Kap. 8 dargestellt.

Der benötigte Rechenaufwand bei der Aktualisierung der OML-Modelle, auch im Vergleich zu einem algorithmisch ähnlichen Offline-Verfahren (BML), wird experimentell in Kap. 9 untersucht. Dort wird auch darauf eingegangen, inwiefern die Modellgüte beeinträchtigt werden könnte, insbesondere im Vergleich zu ähnlichen Offline-Verfahren. Kapitel 10 beschreibt das Hyperparameter-Tuning für OML. Kapitel 11 präsentiert eine Zusammenfassung und gibt wichtige Empfehlungen für die Praxis.

Gummersbach Eva Bartz
April 2023

Inhaltsverzeichnis

Autorenverzeichnis

Thomas Bartz-Beielstein Institute for Data Science, Engineering, and Analytics, TH Köln, Gummersbach, Deutschland

Eva Bartz Bartz & Bartz GmbH, Gummersbach, Deutschland

Florian Dumpert Statistisches Bundesamt, Wiesbaden, Deutschland

Steffen Moritz Statistisches Bundesamt, Wiesbaden, Deutschland

Einleitung: Vom Batch Machine Learning zum Online Machine Learning

1

Thomas Bartz-Beielstein

Inhaltsverzeichnis

Zusammenfassung

Batch Machine Learning (BML), das auch als „Offline Machine Learning“ bezeichnet wird, stößt bei sehr großen Datenmengen an seine Grenzen. Dies betrifft insbesondere den verfügbaren Speicher, das Behandeln von Drift in Datenströmen und die Verarbeitung neuer, unbekannter Daten. Online Machine Learning (OML) ist eine Alternative zu BML, die die Grenzen von BML überwindet. In diesem Kapitel werden die grundlegenden Begriffe und Konzepte von OML vorgestellt, wodurch die Unterschiede zum BML sichtbar werden.

T. Bartz-Beielstein (✉)
Institute for Data Science, Engineering, and Analytics, TH Köln, Gummersbach, Deutschland
E-Mail: thomas.bartz-beielstein@th-koeln.de

T. Bartz-Beielstein und E. Bartz (Hrsg.), *Online Machine Learning*,
https://doi.org/10.1007/978-3-658-42505-0_1

1.1 Datenströme

Das Volumen der aus verschiedenen Quellen generierten Daten hat in den letzten Jahren enorm zugenommen. Technologische Fortschritte haben die kontinuierliche Erfassung von Daten ermöglicht. Zur Beschreibung von Big Data wurden anfänglich die „drei Vs“ (Volume, Velocity und Variety) als Kriterien verwendet[1]: Volume bezeichnet hierbei die große Menge an Daten, Velocity die hohe Geschwindigkeit, mit der die Daten generiert werden, und Variety die große Vielfalt der Daten.

Die in diesem Buch betrachteten Datenströme (Streamingdaten) stellen eine noch größere Herausforderung für Machine Learning (ML)-Algorithmen dar als Big Data. Zu den drei Big-Data-Vs kommen noch weitere Herausforderungen hinzu, die sich insbesondere aus der Flüchtigkeit (Volatilität) und der Möglichkeit, dass abrupte Änderungen („Drift“) auftreten können, ergeben.

Definition 1.1 (Streamingdaten)
Streamingdaten sind Daten, die in einem kontinuierlichen Datenstrom erzeugt werden. Sie sind lose strukturiert, flüchtig (nur einmal verfügbar), immer „fließend“ und beinhalten unvorhersehbare, teilweise abrupte, Änderungen. Streamingdaten sind eine Teilmenge von Big Data mit den folgenden Eigenschaften:

- Volumen: Streamingdaten werden in sehr großen Mengen erzeugt.
- Velocity: Streamingdaten werden in sehr hoher Geschwindigkeit erzeugt.
- Variety: Streamingdaten sind in sehr unterschiedlichen Formaten verfügbar. Diese Eigenschaft bezeichnen wir als „vertikale Vielfalt“.
- Variability: Streamingdaten sind strukturlos und variieren im Laufe der Zeit. Beispielsweise kann graduell oder abrupt Drift auftreten. Diese Eigenschaft bezeichnen wir als „horizontale Vielfalt“.
- Volatilität: Streamingdaten sind flüchtig und nur einmal verfügbar.

Beispiel Streamingdaten

Bei verschiedenen täglichen Transaktionen, z. B. beim Onlineshopping, beim Onlinebanking oder beim Onlinehandel mit Aktien, werden sehr viele Daten erzeugt. Hinzu kommen Sensordaten, Social-Media-Daten, Daten aus Betriebsüberwachungen und Daten aus dem Internet der Dinge, um nur einige Beispiele zu nennen. ◄

Streamingdaten erfordern Analysen in Echtzeit oder nahezu in Echtzeit. Da der Datenstrom ständig produziert wird und nie endet, ist es nicht möglich, diese enormen Datenmengen

[1] Die drei Vs wurden im Laufe der Zeit durch Hinzunahme von Veracity und Value zu den „fünf Vs“ erweitert.