Sarah SCHÖNBRODT, Karlsruhe & Martin FRANK, Karlsruhe

Digitales Lernmaterial zur Netflix Challenge

Zahlreiche Technologien und Anwendung, die wir tagtäglich nutzen basieren auf Methoden aus den Bereichen Data Science und Maschinellem Lernen. Bei der Beantwortung von Problemstellungen aus den genannten Bereichen, wie bspw. die Vorhersage von Nutzerpräferenzen auf Streaming-Plattformen, stellt die mathematische Modellierung mittels linearer Algebra ein wesentliches Werkzeug dar. Data Science Probleme bieten damit eine gute Möglichkeit, mathematischen Modellierungsunterricht zu schülernahen, aktuellen Problemstellungen zu gestalten. So kann und soll nicht nur die Modellierungskompetenz, sondern auch der Umgang mit Daten geschult werden – vor allem in der heutigen Zeit ist ein verantwortungsvoller, kritischer Umgang mit Daten wesentlich. Einige gewinnbringende Ansätze zur Implementierung von Lerneinheiten aus dem Bereich Data Science im Unterricht werden bereits von dem Paderborner Projekt ProDaBi vorangetrieben (vgl. Opel et al., 2019). Zudem wurde in Sube (2019) und Schönbrodt (2019) an verschiedenen realen Fragestellungen herausgearbeitet, wie Modellierungsunterricht zu Data Science Problemen gestaltet werden kann. Wir haben eine Unterrichtsreihe entwickelt, bei der Schüler/innen der Sek. II durch die Bearbeitung eines realen Problems einen Einblick in Methoden aus Data Science und Maschinellem Lernen erlangen.

Die Problemstellung – authentisch und relevant

In der entwickelten Unterrichtsreihe setzen sich die Schüler/innen problemorientiert mit Empfehlungssystemen und der sogenannten Netflix Challenge auseinander. Netflix, Amazon und Co setzen bei der Kundenbindung vor allem auf eins: individuelle Empfehlungen für neue Produkte, Filme etc. auszusprechen. Dazu werden Empfehlungssysteme entwickelt, die vorhersagen, was den Nutzer/innen gefallen könnte. Um das eigene Empfehlungssystem zu verbessern, rief Netflix 2006 einen Wettbewerb aus: Das Team, was im Vergleich zum bestehenden System von Netflix mindestens 10% genauer vorhersagen konnte, welche Filme einem Nutzer gefallen, hatte Chancen auf eine Million USD (vgl. Feuerverger et al., 2012). In der Unterrichtsreihe arbeiten die Schüler/innen mit dem originalen Datensatz der Netflix Challenge. Auf digitalen Arbeitsblättern (realisiert als Jupyter Notebooks) erkunden sie den Datensatz und sammeln Ideen für die Entwicklung eines Empfehlungssystems. Anschließend erarbeiten sie ein mathematisches Modell und wenden dieses auf den Netflixdatensatz an. Das entwickelte Lernmaterial zeigt exemplarisch wie datenlastige Probleme aus dem Alltag der Lernenden aufbereitet und im Rahmen von Modellierungsprojekten, sowohl im Distanzlernen als auch in Präsenz, durchgeführt werden können. Das Material liegt auf einer Cloud-Plattform (www.cammp.online/214.php) für den Unterrichtseinsatz bereit und kann im Webbrowser bearbeitet werden. Durch digitales Differenzierungsmaterial und individuelle, automatisierte Rückmeldungen zu den Lösungen der

In: Kerstin Hein, Cathleen Heil, Silke Ruwisch & Susanne Prediger (Hrsg.). Beiträge zum Mathematikunterricht 2021. Münster: WTM Verlag. https://doi.org/10.37626/GA9783959871846.0
Online unter https://eldorado.tu-dortmund.de/handle/2003/30630

Lernenden kann das Material in heterogenen Lerngruppen eingesetzt werden. Wesentliche Modellierungsschritte der Unterrichtsreihe werden nachfolgend kurz beschrieben.

Die Leitfrage der Unterrichtsreihe lautet: „Wie kann bestmöglich vorhergesagt werden, welche Bewertung ein User für einen Film abgeben würde, den er noch nicht bewertet bzw. gesehen hat?“ Basierend auf den Vorhersagen sollen den Usern Filmvorschläge unterbreitet werden.

Lerneinheit 1: Daten verstehen und analysieren

Ausgangspunkt der Unterrichtsreihe ist der Netflixdatensatz. Dieser besteht aus 17.700 Filmen, 480.189 Usern und 100.480.507 Bewertungen von eins (schlechteste) bis fünf (beste Bewertung). Von den Filmen sind Titel und Erscheinungsjahr bekannt (vgl. Feuerverger et al., 2012).

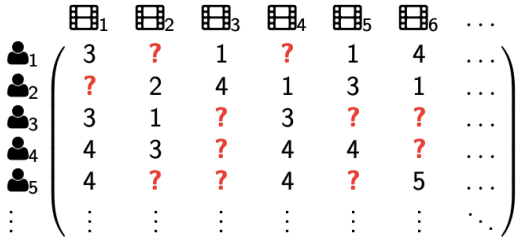

Abb. 1: Rating-Matrix bzw. Rating-Tabelle, die die Bewertungen der User enthält

Die Lernenden untersuchen den Datensatz mithilfe verschiedener Darstellungsformen, wie der Rating-Matrix bzw. Rating-Tabelle (vgl. Abb. 1) oder Streudiagrammen. Dabei beantworten sie Fragen wie „Bei welchem Film weichen die Bewertungen am wenigsten vom Mittelwert ab?“. Sie entscheiden, welche Darstellungsform für die Beantwortung welcher Frage geeignet ist. Bei der Einführung der tabellarischen Darstellung der Bewertungsdaten (vgl. Abb. 1) wird der Begriff einer Matrix zunächst vermieden, da er keine Voraussetzung für die verständige Bearbeitung des Lernmaterials sein sollte.

Lerneinheit 2: Entwicklung eines mathematischen Modells

Die Lernenden erarbeiten eigene Ideen, um mittels der bekannten Bewertungen in der Rating-Tabelle die fehlenden, unbekannten Bewertungen vorherzusagen. In bisherigen Durchführungen nannten die Lernenden diverse Ansätze, darunter:

- „Ähnliche User finden und dann Filme finden, die sie auch gut fanden.“,
- „Interessenbereiche der Nutzer, wie zum Beispiel das Genre feststellen.“
- „Genre-abhängig schauen, wie bewertet wurde und Filme aus Genre vorschlagen.“

Anschließend steigen die Lernenden an kleinen Rating-Tabellen in die Entwicklung eines ausgewählten mathematischen Modells ein: der Zerlegung der Rating-Tabelle in eine User-Eigenschafts-Tabelle und eine Film-Eigenschafts-Tabelle (vgl. Abb. 2). Die User-Tabelle gibt an, wie sehr ein User gewisse Eigenschaften, wie beispielsweise die Genres Action oder Comedy, mag. Die Movie-Tabelle gibt

an, wie stark diese Eigenschaften bei den einzelnen Filmen ausgeprägt sind. Die Schüler/innen entwickeln eigenständig eine Formel, mit der sich die bekannten Bewertungen aus den Zeilen der User-Tabelle und den Spalten der Movie-Tabelle ergeben und übertragen dies auf die Berechnung von unbekannten Bewertungen. Die Berechnung der Bewertungen ist dabei nichts anderes als das Skalarprodukt aus Zeilenvektor der User-Tabelle und Spaltenvektor der Movie-Tabelle. Der Kern des beschriebenen Modells ist eine Matrix-Faktorisierung. Diese ist auch der ausschlaggebende Bestandteil des Gewinnermodels der Netflix Challenge (vgl. Koren et al., 2009). Damit sind nicht nur die Problemstellung, sondern auch die verwendeten mathematischen Methoden authentisch.

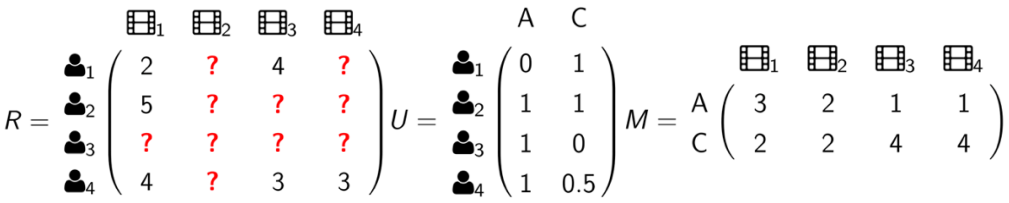

Abb. 2 Rating-Tabelle *R* und User-Tabelle U bzw. Movie-Tabelle M. Die User- bzw. Movie-Tabelle spiegeln die Interessen der User bzw. die Ausprägung der Filme hinsichtlich der Eigenschaften „Action“ (A) und „Comedy“ (C) wider.

Tatsächlich enthält der Netflixdatensatz keine Informationen zu den Ausprägungen der Filme hinsichtlich der Genres. Da die händische Festlegung der Ausprägungen bei großen Datensätzen nicht denkbar ist, müssen User- und Movie-Tabelle geeignet berechnet werden. Motiviert durch diese Tatsache bestimmen die Lernenden für eine 2x2-Rating-Tabelle eine Zerlegung, bei der die zuvor hergeleitete Formel (das Skalarprodukt), jeweils die bekannte Bewertung aus der Rating-Tabelle ergibt.

Anmerkung: Ein anderer Modellansatz, zu dem ebenfalls Lernmaterial erprobt wurde, ist die Modellierung von Ähnlichkeiten – entweder zwischen Usern oder zwischen Filmen. Bei der Wahl der Ähnlichkeitsmaße können die Lernenden kreativ werden, lernen überdies aber auch gängige Ähnlichkeitsmaße wie die Kosinus-Ähnlichkeit und die Pearson-Korrelation kennen (vgl. Sarwar et al., 2001)

Lerneinheit 3: Fehlermaß und Optimierung

Die Berechnung einer Zerlegung „per Hand“ ist für große Rating-Matrizen nicht denkbar. Ziel ist es, dem Computer die Berechnung zu überlassen. Dazu definieren die Lernenden zunächst ein Fehlermaß, mit dem sie (und das Computerverfahren) bewerten können, ob eine Zerlegung ausreichend gut ist. Anschließend wird ein Optimierungsverfahren angewandt, welches eine Zerlegung berechnet, die zu einem möglichst kleinen Fehler auf den bekannten Daten führt. Das Verfahren wird von den Lernenden nicht selbst entwickelt. Stattdessen kommt ein Optimierungspaket der verwendeten Programmiersprache, in unserem Fall Julia, als Black-Box zum Einsatz.

Lerneinheit 4: Anwendung auf den Netflixdatensatz und Diskussion

Da bisher nur bewertet wurde, wie gut das Modell zur Repräsentation bekannter Bewertungen geeignet ist, aber unbekannte Bewertungen vorhergesagt werden sollen, wird eine Strategie des überwachten Maschinellen Lernens genutzt: Die Unterteilung der bekannten Bewertungsdaten in solche, die zur Berechnung einer Zerlegung verwendet werden (Trainingsdaten) und solche, die genutzt werden, um zu validieren, wie gut die Vorhersage auf „unbekannten" Daten (Testdaten) funktioniert. In einer Abschlussdiskussion werden abschließend Grenzen des entwickelten mathematischen Modells diskutiert. Es wird u. a. kritisch reflektiert, welche Möglichkeiten der Manipulation von Empfehlungssystemen existieren und inwieweit sog. Filterblasen problematisch sein könnten. Die persönliche Meinung und die Erfahrungen der Lernenden mit Empfehlungssystemen werden dabei insbesondere einbezogen.

Erfahrungen und Fazit

Bei der Bearbeitung des Lernmaterials kommen zahlreiche schulmathematische Inhalte zum Einsatz: Mittelwerte, Standardabweichung, Vektoren, das Skalarprodukt und Funktionen. Die Problemstellung ist nicht nur äußerst schülernah, sondern real und authentisch und liefert den Lernenden eine Antwort auf die Frage, "wozu es Mathematik eigentlich braucht." Die Lernenden sammeln Erfahrungen im Umgang mit großen Datenmengen und erhalten einen Einblick in Strategien des Maschinellen Lernens. Das Material wurde bereits mit mehr als 90 Schüler/innen ab Klasse 10 im Rahmen von Unterrichtsreihen (á 3-4 Doppelstunden) oder Projekttagen eingesetzt. Bei den Durchführungen stachen die diversen Ideen der Lernenden zur Entwicklung eines Empfehlungssystems und die vielfältigen Argumente bei der kritischen Reflexion heraus. In den sehr regen Diskussionen zeigte sich, dass die Lernenden ein großes Interesse an der Problemstellung haben.

Literatur

Feuerverger, A., He, Y., Khatri, S. (2012). Statistical Significance of the Netflix Challenge. Statistical Science, 27 (2), 202–231.

Koren, Y., Bell, R., Volinsky, C. (2009). Matrix Factorization Techniques for Recommender Systems. IEEE Computer Society Press, 42 (8), 30–37.

Opel, S., Schlichtig, M., Schulte, C., Biehler, R., Frischemeier, D., Podworny, S., Wassong, T. (2019). Entwicklung und Reflexion einer Unterrichtssequenz zum Maschinellen Lernen als Aspekt von Data Science in der Sekundarstufe II. In: A. Pasternak (Hrsg.), Proceedings zur 18. GI-Fachtagung Informatik und Schule, 285–294.

Sarwar, B., Karypis, G., Konstan, J. and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In: Proc. 10th Int. Conf. on the WWW, 285–295. ACM, NY

Schönbrodt, S. (2019). Maschinelle Lernmethoden für Klassifizierungsprobleme – Perspektiven für die mathematische Modellierung mit Schülerinnen und Schülern. Springer Spektrum, Wiesbaden

Sube, M. (2019). Entwicklung und Evaluation von Unterrichtsmaterial zu Data Science und mathematischer Modellierung mit Schülerinnen und Schülern, Dissertation, RWTH Aachen