Technische Universität Braunschweig | Institut für Finanzwirtschaft
Abt-Jerusalem-Str. 7 | 38106 Braunschweig | Deutschland

Univ.-Prof. Dr. Marc Gürtler
Tel. +49 (0) 531 391-2896
Fax +49 (0) 531 391-2899
marc.guertler@tu-bs.de
https://www.tu-bs.de/fiwi

Maschinelles und Statistisches Lernen in der Finanzwirtschaft WS 2023/2024

Di., 16:45-18:15 Uhr, PK 4.3
Beginn der Vorlesung: 24. Oktober 2023

Termin	Präsenzveranstaltung
24.10.2023	1. Grundlagen des maschinellen und statistischen Lernens (Teil 1)
31.10.	*Reformationstag*
07.11.	1. Grundlagen des maschinellen und statistischen Lernens (Teil 2)
14.11.	1. Grundlagen des maschinellen und statistischen Lernens (Teil 3)
21.11.	2. Lineare Regression (Teil 1)
28.11.	2. Lineare Regression (Teil 2)
05.12.	2. Lineare Regression (Teil 3)
12.12.	3. Klassifikation
19.12.	4. Kreuzvalidierung
23.12.-05.01.	*Weihnachtsferien*
09.01.2024	5. Regularisierung
16.01	6. Grundlagen von Entscheidungsbäumen
23.01.	7. Bagging und Random Forests
30.01.	8. Durchführung eines maschinellen Lernprojekts
06.02.	9. Projekt „Identifikation von Kreditkartenbetrug“

Die Veranstaltung wird in Präsenz als zweistündige Vorlesung abgehalten. Darüber hinaus werden auf stud.ip „E-Lectures“ zur Verfügung gestellt, in denen begleitend die relevanten Inhalte der Statistik-Software R erläutert werden. Einmal wöchentlich findet eine Chat-Sprechstunde über stud.ip statt, in der Raum für Fragen und Diskussionen ist. Ferner finden Sie in stud.ip ein Diskussionsforum, in dem Fragen untereinander diskutiert werden können. Am Ende der Veranstaltung sollen die Studierenden in der Lage sein, ein eigenes Projekt zum maschinellen Lernen umzusetzen. Bei nicht zu großer Studierendenzahl wird eine mündliche Prüfung in Form einer Projektpräsentation (die noch genauer erläutert wird) stattfinden.

Auf den folgenden Seiten finden Sie eine Detailbeschreibung des Vorlesungsstoffs der einzelnen Sitzungen und Literaturangaben. Das Buch von James et al. (2023) ist als Download unter dem unten angegebenen Link verfügbar. Ferner werden alle Themen durch eine Vielzahl an finanzwirtschaftlichen Beispielen begleitet. Die beispielhaften Umsetzungen werden mit dem Softwarepaket R realisiert, für das Lernvideos auf stud.ip zur Verfügung stehen.

Kurzbeschreibung der Themen

Thema 1: Grundlagen des maschinellen und statistischen Lernens

Es werden grundlegende Begrifflichkeiten zum maschinellen und statistischen Lernen eingeführt. Es wird das Ziel hoher Prognosegenauigkeit formuliert und im Zusammenhang mit der Prognosegüte auf den nicht reduzierbaren Prognosefehler eingegangen. Es wird auf die Notwendigkeit der Transparenz in den Verfahren hingewiesen, um aus identifizierten Zusammenhängen Schlussfolgerungen und Handlungsempfehlungen ziehen zu können. Es wird dargelegt, wie ein Modell trainiert wird, indem es optimal an die Daten eines Datensatzes angepasst wird. Dabei wird das Problem der Überanpassung thematisiert und erläutert, wie mit dem sogenannten Bias-Variance Trade-off umzugehen ist. Schließlich wird für Klassifikationsprobleme der k-nearest neighbors Klassifikator eingeführt und anhand eines einfachen Beispiels zur Kreditausfallprognose umgesetzt. Schließlich wird auf Schätzprobleme dieses Klassifikators eingegangen.

James et al. (2023) [15-42]

Thema 2: Lineare Regression

Es wird das univariate lineare Regressionsmodell vorgestellt und für dieses das Optimierungsproblem mit der Methode der kleinsten Quadrate („OLS, ordinary least squares“) entwickelt und gelöst. Anhand eines Kreditbeispiels werden die OLS-geschätzten Steigungskoeffizienten ermittelt und ihre Bedeutung erörtert. Weiterhin wird das Bestimmtheitsmaß als Gütemaß zur Beurteilung des Regressionsmodells diskutiert und es wird dargelegt, wie sich aus Stichprobenergebnissen durch Hypothesentests Schlüsse auf die Grundgesamtheit ziehen lassen. Anschließend wird das allgemeine multivariate lineare Regressionsmodell behandelt und die optimale Anzahl der zu wählenden erklärenden Variablen diskutiert, die sich im Spannungsfeld zwischen „Omitted variable bias“ und „overfitting“ bewegt. Schließlich wird noch die Gruppe der qualitativen Prädiktoren vorgestellt, die gerade für Fragen der Kreditentscheidung Relevanz besitzen und deren Einfluss durch Interaktionseffekte beschrieben werden kann.

James et al. (2023) [59-110]

Thema 3: Klassifikation

Es wird das allgemeine Binär-Klassifikationsproblem dargestellt und am Beispiel der Kreditausfallprognose erörtert. Zu diesem Zweck wird konkret das Logit-Modell verwendet und die Maximum-Likelihood-Schätzung der Parameter vorgestellt. Es wird die Prognosegüte beim Logit-Modell untersucht. Zu diesem Zweck wird neben der Fehlerrate die Konfusionsmatrix vorgestellt, um auf dieser Basis die Maßzahlen Sensitivität und Spezifität, aber auch die ROC-Kurve vorzustellen. Aus letzterer lässt sich schließlich noch die für „Area under the curve (AUC)“ ableiten, die für die Klassifikation im Kreditbereich hohe Relevanz besitzt.

James et al. (2023) [129-152]

Thema 4: Kreuzvalidierung

Um von einer Prognose auf Basis einer Stichprobe auf Zusammenhänge in der Grundgesamtheit schließen zu können, ohne über einen zusätzlichen großen Testdatensatz zu verfügen, wird innerhalb der Trainingsdaten ein Validierungsdatensatz von Trainingsdaten getrennt. Da die Prognosegüte je nach Aufteilung der Daten stark variiert, wird das Verfahren der Kreuzvalidierung eingeführt, um Testfehler mit einer geringeren Varianz schätzen zu können. Neben der Kreuzvalidierung für Regressionsprobleme wird aufgrund der Relevanz für die Kreditausfallprognose zusätzlich die Kreuzvalidierung für Klassifikationsprobleme vorgestellt.

James et al. (2023) [197-219]

Thema 5: Regularisierung

Im Rahmen einer linearen Regression werden häufig für die abhängige Variable irrelevante unabhängige Variablen berücksichtigt, was zu einer unnötigen Komplexität des Modells und zu einer hohen Varianz der Regressionskoeffizienten führt. Insofern führt die gezielte Begrenzung der Prädiktoren zu einer reduzierten Varianz und oft zu einer erhöhten Prognosegenauigkeit. Es werden verschiedene Verfahren zur Variablenselektion vorgestellt. Vertieft werden Regularisierungsansätze betrachtet, die die Regressionskoeffizienten begrenzen. Konkret werden die Ridge-Regression und die Lasso-Regression behandelt und mit der klassischen linearen Regression verglichen. Die geschieht anhand eines Kreditkartendatensatzes, für den das Kreditaufnahmeverhalten prognostiziert werden soll.

James et al. (2023) [225-251]

Thema 6: Grundlagen von Entscheidungsbäumen

Es wird dargelegt, wie über Entscheidungsbäume die Menge aller Prädiktorenkombinationen in Teilregionen segmentiert werden und auf Basis der Teilsegmente Prognosen durchgeführt werden. Beispielhaft werden Entscheidungsbäume zur Prognose von Ausgaben einer Krankenversicherung genutzt. Dabei wird auch darauf eingegangen, dass Pruning zu einer verbesserten Out-of-sample-Prognose führen kann. Neben Regressionsproblemen werden auch Klassifikationsprobleme diskutiert und am Beispiel der Prognose von Kreditausfällen aufgezeigt. Schließlich wird die Prognosegüte von Entscheidungsbäumen diskutiert und der Vorteil hoher Transparenz hervorgehoben.

James et al. (2023) [327-340]

Thema 7: Bagging und Random Forests

Es wird dargelegt, dass die Verwendung mehrerer Bäume mit kombinierter Prognose zu einer erheblichen Steigerung der Prognosefähigkeit führt. Es wird das Bagging-Verfahren erläutert, das durch bootstrapping für jeden Baum eine neue zufällige Trainingsmenge für die Baum-Kalibrierung nutzt. Zur Beurteilung werden sowohl der normale Test-Fehler als auch der Out-of-Bag-Fehler vorgestellt. Da beim Bagging eine hohe Korrelation zwischen den Bäumen entstehen kann, wird als Weiterentwicklung das Verfahren des Random Forests vorgestellt, bei dem die Menge der verwendeten Prädiktoren zufällig variiert und auf diese Weise eine „Dekorrelierung“ der Bäume resultiert. Beide Verfahren werden zur Kreditausfallprognose genutzt und hinsichtlich ihrer Prognosegüte verglichen.

James et al. (2023) [340-345]

Thema 8: Durchführung eines maschinellen Lernprojekts

Bevor im letzten Thema ein ausgewähltes Projekt aus dem Bereich des statistischen und maschinellen Lernens vorgestellt wird, wird in diesem Thema erläutert, wie ein solches Projekt durchgeführt werden sollte. Von der Identifikation der Forschungsfrage über die Literaturrecherche und Datenvorverarbeitung bis hin zur Modellierung und Auswertung der empirischen Ergebnisse werden nützliche Tipps und Hinweise gegeben.

Heesen, B. (2023)

Thema 9: Projekt „Identifikation von Kreditkartenbetrug“

Zahlungsbetrug verursacht weltweit Verluste in Milliardenhöhe. Die Erkennung von Auffälligkeiten bei Kreditkartentransaktionen wird jedoch aufgrund der ständig wachsenden Datenmengen immer schwieriger. Aus tausenden von potentiellen Transaktionsmerkmalen die relevanten Betrugsmuster zu identifizieren, ist eine hochkomplexe Aufgabe. In diesem Projekt werden die in den vorangegangenen Themen vorgestellten Lernverfahren auf simulierte Kreditkartentransaktionen angewendet, um betrügerische Transaktionen erfolgreich vorherzusagen. Dabei stellt sich die Frage nach der optimalen Methode für diese Klassifikationsaufgabe.

Literatur

James, Gareth/Witten, Daniela/Hastie, Trevor/Tibshirani, Robert (2023): An Introduction to Statistical Learning with Applications in R, 2. Auflage, korrigierte Fassung, Springer.

Download:

https://hastie.su.domains/ISLR2/ISLRv2_corrected_June_2023.pdf

Heesen, B. (2023): Künstliche Intelligenz und Machine Learning mit R, Springer Fachmedien Wiesbaden.

Download:

https://link.springer.com/content/pdf/10.1007/978-3-658-41576-1.pdf?pdf=button