Data Science 1

Sebastian Sauer

2022-02-21 22:34:24

Contents

1 Überblick **5**

1.1 Was Sie hier lernen und wozu das gut ist 7
1.2 Lernziele 7
1.3 Voraussetzungen 7
1.4 Hinweise zu diesem Projekt 7
1.5 Lernhilfen 8
1.6 Modulzeitplan 8
1.7 Literatur 9

2 Modulüberblick **11**

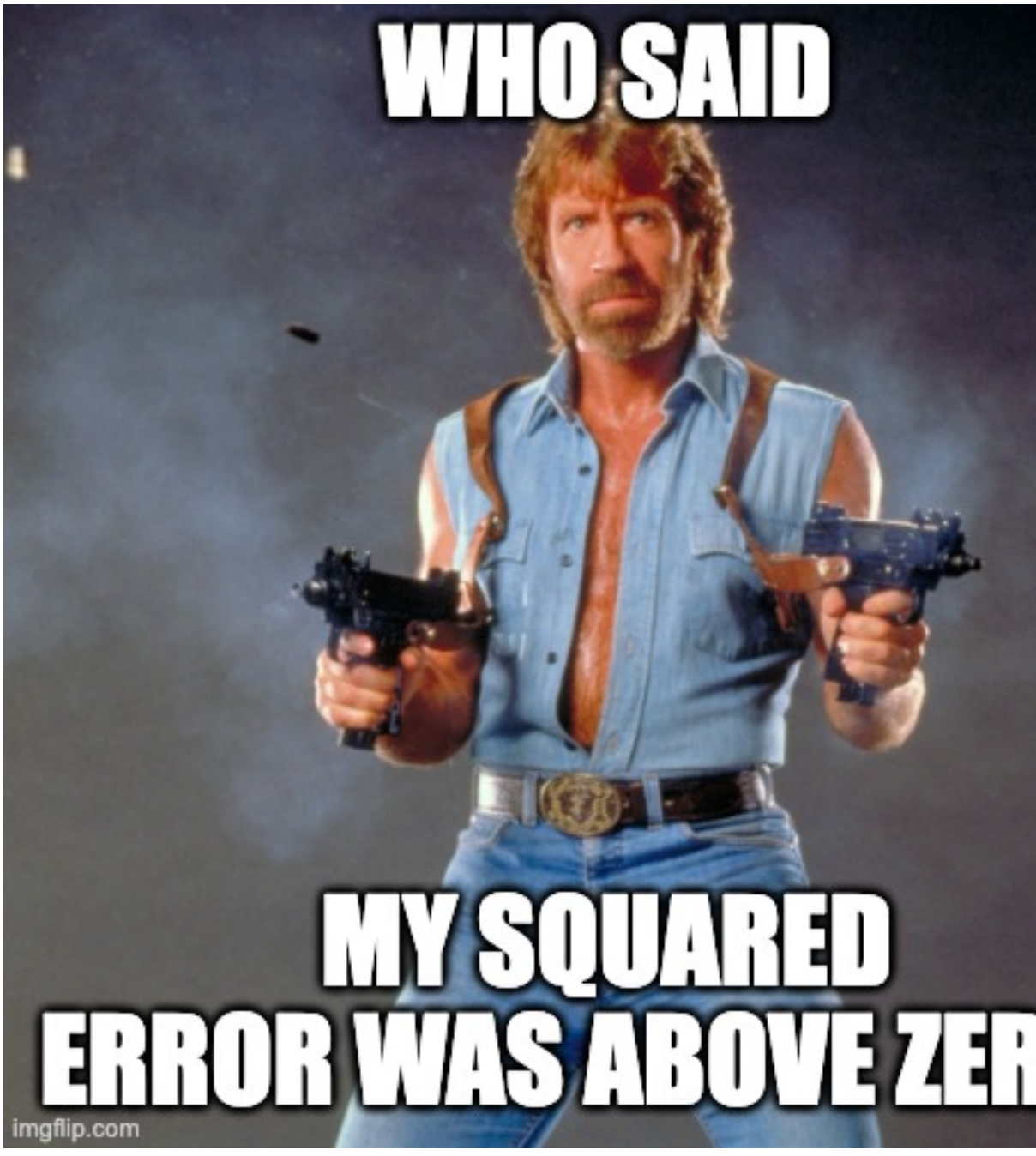
2.1 Grundkonzepte 11
2.2 tidyverse, 2. Blick 12
2.3 tidymodels 12
2.4 kNN 13
2.5 Statistisches Lernen 13
2.6 Wiederholung 13
2.7 Logistische Regression 14
2.8 Naive Bayes 14
2.9 Entscheidungsbäume 14
2.10 Zufallswälder 14
2.11 Fallstudie 15
2.12 Wiederholung 15
2.13 GAM 15

2.14 Lasso und Co 16
2.15 Vertiefung 16

Chapter 1

Überblick

```
knitr::include_graphics("img/662upp.jpg")
```



Imgflip Meme Generator

1.1 Was Sie hier lernen und wozu das gut ist

Alle Welt spricht von Big Data, aber ohne die Analyse sind die großen Daten nur großes Rauschen. Was letztlich interessiert, sind die Erkenntnisse, die Einblicke, nicht die Daten an sich. Dabei ist es egal, ob die Daten groß oder klein sind. Natürlich erlauben die heutigen Datenmengen im Verbund mit leistungsfähigen Rechnern und neuen Analysemethoden ein Verständnis, das vor Kurzem noch nicht möglich war. Und wir stehen erst am Anfang dieser Entwicklung. Vielleicht handelt es sich bei diesem Feld um eines der dynamischsten Fachgebiete der heutigen Zeit. Sie sind dabei: Sie lernen einiges Handwerkszeugs des "Datenwissenschaftlers". Wir konzentrieren uns auf das vielleicht bekannteste Teilgebiet: Ereignisse vorhersagen auf Basis von hoch strukturierten Daten und geeigneter Algorithmen und Verfahren. Nach diesem Kurs sollten Sie in der Lage sein, typisches Gebabbel des Fachgebiet mit Lässigkeit mitzumachen. Ach ja, und mit einigem Erfolg Vorhersagemodelle entwickeln.

1.2 Lernziele

Nach diesem Kurs sollten Sie

- grundlegende Konzepte des statistischen Lernens verstehen und mit R anwenden können
- gängige Prognose-Algorithmen kennen, in Grundzügen verstehen und mit R anwenden können
- die Güte und Grenze von Prognosemodellen einschätzen können

1.3 Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie folgendes Wissen mitbringen:

- grundlegende Kenntnisse im Umgang mit R, möglichst auch mit dem tidyverse
- grundlegende Kenntnisse der deskriptiven Statistik
- grundlegende Kenntnis der Regressionsanalyse

1.4 Hinweise zu diesem Projekt

- Die URL zu diesem Projekt lautet <test.io>.
- Lesen Sie sich die folgenden Informationen bitte gut durch: Hinweise

- Den Quellcode finden Sie in diesem Github-Repo.
- Sie haben Feedback, Fehlerhinweise oder Wünsche zur Weiterentwicklung? Am besten stellen Sie hier einen *Issue* ein.
- Dieses Projekt steht unter der MIT-Lizenz).

1.5 Lernhilfen

1.5.1 Software

- Installieren Sie R und seine Freunde.
- Installieren Sie die folgende R-Pakete:
  - tidyverse
  - tidymodels
  - weitere Pakete werden im Unterricht bekannt gegeben (es schadet aber nichts, jetzt schon Pakete nach eigenem Ermessen zu installieren)
- R Syntax aus dem Unterricht findet sich im Github-Repo bzw. Ordner zum jeweiligen Semester.

1.5.2 Online-Zusammenarbeit

- Frag-Jetzt-Raum zum anonymen Fragen stellen während des Unterrichts. Der Keycode wird Ihnen vom Dozenten bereitgestellt.
- Padlet zum einfachen (und anonymen) Hochladen von Arbeitsergebnissen der Studentis im Unterricht. Wir nutzen es als eine Art Pinwand zum Sammeln von Arbeitsbeiträgen. Die Zugangsdaten stellt Ihnen der Dozent bereit.

1.6 Modulzeitplan

Nr.	Kalenderwoche	Datum	Thema
1	11	14.-18.3.22	Grundkonzepte
2	12	21.3.-25.3.	tidyverse, 2. Blick
3	13	28.3.-1.4.	tidymodels
4	14	4.4.-8.4.	kNN
5	15	11.4.-15.4.	Statistisches Lernen
6	16	18.4.-22.4	Wiederholung
7	17	25.4.-29.4.	Logistische Regression

Nr.	Kalenderwoche	Datum	Thema
8	18	2.4.-6.5.	Naive Bayes
9	19	9.5.-13.5.	Entscheidungsbäume
10	20	16.5.-20.5.	Zufallswälder
11	21	23.5.-27.5.	Fallstudie
12	23	6.6.-10.6.	Wiederholung
13	24	13.6.-17.6.	GAM
14	25	20.6.-24.6.	Lasso und Co
15	26	27.6.-1.7.	Vertiefung

1.7 Literatur

Zentrale Kursliteratur für die theoretischen Konzepte ist (Rhys, 2020). Die praktische Umsetzung in R basiert auf (Silge and Kuhn, 2022).

Chapter 2

Modulüberblick

2.1 Grundkonzepte

2.1.1 Datum

- 14.-18.3.22

2.1.2 Lernziele

- Sie können erläutern, was man unter statistischem Lernen versteht.
- Sie wissen, war Overfitting ist, wie es entsteht, und wie es vermieden werden kann.
- Sie kennen verschiedenen Arten von statistischem Lernen und können Algorithmen zu diesen Arten zuordnen.

2.1.3 Vorbereitung

- Lesen Sie die Hinweise zum Modul.
- Installieren (oder Updaten) Sie die für dieses Modul angegeben Software.
- Lesen Sie die Literatur.

2.1.4 Literatur

- Rhys, Kap. 1

2.1.5 Skript

- Kap. 1

2.1.6 Hinweise

- Bitte beachten Sie die Einteilung in die Züge für den Präsenzunterricht.

2.2 tidyverse, 2. Blick

2.2.1 Datum

- 21.3.-25.3.

2.2.2 Lernziele

- Sie können Funktionen, auch anonyme, in R schreiben.
- Sie können Datensätze vom Lang- und Breit-Format wechseln.
- Sie können Mapping-Funktionen anwenden.
- Sie können eine dplyr-Funktion auf mehrere Spalten gleichzeitig anwenden.

2.2.3 Vorbereitung

- Lesen Sie die Literatur.

2.2.4 Literatur

- Rhys, Kap. 2

2.3 tidymodels

2.3.1 Datum

- 28.3.-1.4.

2.3.2 Literatur

- TMWR

2.4 kNN

2.4.1 Datum

- 4.4.-8.4.

2.4.2 Literatur

- Rhys, Kap. 3

2.5 Statistisches Lernen

2.5.1 Datum

- 11.4.-15.4.

2.5.2 Literatur

- Rhys, Kap. 3

2.5.3 Hinweise

- In dieser Woche fällt die Übung aus.

2.6 Wiederholung

2.6.1 Datum

- 18.4.-22.4

2.6.2 Hinweise

- In dieser Woche fällt die Vorlesung aus.

2.7 Logistische Regression

2.7.1 Datum

- 25.4.-29.4.

2.7.2 Literatur

- Rhys, Kap. 4

2.8 Naive Bayes

2.8.1 Datum

- 2.4.-6.5.

2.8.2 Literatur

- Rhys, Kap. 6

2.9 Entscheidungsbäume

2.9.1 Datum

- 9.5.-13.5.

2.9.2 Literatur

- Rhys, Kap. 7

2.10 Zufallswälder

2.10.1 Datum

- 16.5.-20.5.

2.10.2 Literatur

- Rhys, Kap. 8

2.11 Fallstudie

2.11.1 Datum

- 23.5.-27.5.

2.11.2 Literatur

- Rhys, Kap.9

2.11.3 Hinweise

- Nächste Woche ist BlocKalenderwocheoche; es findet kein regulärer Unterricht statt.
- Diese Woche fällt die Übung aus.

2.12 Wiederholung

2.12.1 Datum

- 6.6.-10.6.

2.12.2 Hinweise

- In dieser Woche fällt die Vorlesung aus.

2.13 GAM

2.13.1 Datum

- 13.6.-17.6.

2.13.2 Literatur

- Rhys, Kap. 10

2.14 Lasso und Co

2.14.1 Datum

- 20.6.-24.6.

2.14.2 Literatur

- Rhys, Kap. 11

2.15 Vertiefung

2.15.1 Datum

- 27.6.-1.7.

2.15.2 Literatur

- Rhys, Kap. 12

2.15.3 Hinweise

- Nach dieser Woche endet der Unterricht.

Bibliography

Rhys, H. (2020). *Machine Learning with R, the tidyverse, and mlr*. Manning publications, Shelter Island, NY. OCLC: on1121083327.

Silge, J. and Kuhn, M. (2022). *Tidy Modeling with R*.