

29.10.19

Einführung in Data Science und maschinelles Lernen mit R

Einführung



- **Vorstellung**
- **Kursinhalte**
- **Was ist Data Science?**
- **Warum R?**
- **Wozu Rstudio?**
- **Datenstrukturen in R**

- **Am Ende der Veranstaltung immer die Anwesenheitsliste unterschreiben**
- **Bitte Bescheid sagen, wenn ihr aussteigt!**

Einzelvorstellung

Name

Interesse

Kurzfragebogen

bit.ly/oc-datascience

Vorstellungsrunde

5 Minuten

jeweils 3 sich unbekannte Personen

Vorstellungsrunde

5 Minuten

jeweils 3 sich unbekannte Personen

29.10. 18:00- 20:00	Einführung Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
05.11. 18:00- 20:00	Grafische Darstellung von Daten Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
12.11. 18:00- 20:00	R-Projekte und Versionierung mit Git und GitHub Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
19.11. 18:00- 20:00	Datenaufbereitung mit Tidyverse Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
07.01. 18:00- 20:00	Einführung in das maschinelle Lernen Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
14.01. 18:00- 20:00	Support Vector Maschinen Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
21.01. 18:00- 20:00	Neuronale Netze und Deep Learning Starterkitchen, Kuhnkestraße 6, Wissenschaftspark
28.01. 18:00- 20:00	Präsentation der Auswertungsprojekte Starterkitchen, Kuhnkestraße 6, Wissenschaftspark

“What's the difference between data science, machine learning, and artificial intelligence?”

(<http://varianceexplained.org/r/ds-ml-ai/>)

Data science produces **insights**.

Machine learning produces **predictions**.

Artificial intelligence produces **actions**.

DATA SCIENCE

Anwendung statistischer Auswertungsverfahren

mit deren Hilfe Ergebnisse für einen Bericht oder Artikel erzeugt werden.

- Deskriptive Statistiken
- Statistische Schlussfolgerungen
- Visualisierung von Daten
- Design von Experimenten

MASCHINELLES LERNEN

Anwendung eines statistisches Auswertungsverfahren

zur Schätzung von Modellparametern (auf Basis einer Kreuzvalidierung)

mit deren Hilfe eine Funktion erstellt wird,

die Prognosen liefert und die Modellparameterschätzung durch Hinzufügen neuer Daten ständig verbessert.

- Prognose zukünftigen Verhaltens
- Prognose unbeobachteten Verhaltens
- Prognose aktueller Zustände (etwa, ob ein Bild einen Vogel enthält)

KÜNSTLICHE INTELLIGENZ

„an autonomous agent executes or recommends actions“

(Poole, Mackworth, & Goebel, 1998)

„Systeme mit einem ‚intelligenten‘ Verhalten, die ihre Umgebung analysieren und mit einem gewissen Grad an Autonomie handeln, um bestimmte Ziele zu erreichen.“

(Europäische Kommission, 2018)

„Unter Künstlicher Intelligenz verstehen wir hochentwickelte Softwaresysteme, welche lernfähig und trainierbar sind, um komplexe Aufgaben bewältigen können.“

(KI-Strategie des Landes Schleswig-Holstein, 2019)

KÜNSTLICHE INTELLIGENZ

- Programme, die Spiele spielen (Deep Blue, AlphaGo)
- Programme zur Steuerung von Robotern
- Programme zur Optimierung (Google Maps)
- Programme zur Verarbeitung natürlicher Sprache (Bots)

EIGENSCHAFTEN VON R

- Von Statistikern entwickelte Auswertungssprache
(Python etwa kommt aus dem IT-/Tech-Umfeld)
- Interpretersprache
(Syntax wird Zeile für Zeile interpretiert, vergleichbar eines Taschenrechners)
- Konsolen-Programm ohne grafische Benutzeroberfläche

```

1  #####
2  # Prepare Environment
3  #####
4  setwd("C:/Users/Steffen/Arbeit/opencampus/99_source-code/elearning-statistics/00_functions")
5  source("02_LimeSurvey-API.R") # import LimeSurvey API
6  setwd("../../auswertung pre-challenge")
7
8  library(inspectdf)
9  library(dplyr)
10 library(readr)
11 library(stringr)
12
13
14 #####
15 # Import Answer Data
16 #####
17 introData <- as_tibble(ls.getResponses(615549, sHeadingType = "code")) # import of the Introductory Questionnaire
18 closingData <- as_tibble(ls.getResponses(722958, sHeadingType = "code")) # import of the Closing Questionnaire
19
20
21 #####
22 # Prepare Table for Scoring
23 #####
24 # Shrink Datasets to Knowledge Items and join them
25 introItems <- select(introData , matches("Q2|Q3"))
26 closingItems <- select(closingData , matches("Q2|Q3"))
27 # shorten name to exclude questionnaire ID
28 names(introItems) <- substr(names(introItems),3, 10)
29 names(closingItems) <- substr(names(closingItems),3, 10)
30 # join items according to question IDs
31 items <- full_join(introItems, closingItems)
32
33 # construct table including all given answers, how often they occurred (percentage and count) and include corresponding score categories if provided
34 scoreTable <- get.scoreTable(items)
35 write.csv2(file="ScoringGuide_new.csv", scoreTable, row.names = F, na="")
36 scoreTable <- read_csv2(file="ScoringGuide_new.csv")
37

```

NACHTEILE VON R

- Man muss programmieren
- Im Bereich des maschinellen Lernens weniger verbreitet als Python
- Weniger Schnittstellen (APIs) als Python zu speziellen Anwendungen im IT-Bereich

VORTEILE VON R

- Open Source
- Sehr große Community
- Im wissenschaftlichen Bereich zunehmend Standard für statistische Auswertungen
- Umfangreichste Funktionssammlung im Bereich Data Science / Statistik
- Leichte Integration von anderen Programmiersprachen

RSTUDIO

(Beliebteste) Umgebung zur Erstellung und Ausführung von R-Programmen

- Komfortable Bearbeitung und Ausführung der Syntax,
- Zugriff auf Hilfe-Dateien,
- Anzeige aktuell im Speicher geladener Daten und Funktionen,
- Anzeige grafischer Ausgaben,
- Installation zusätzlicher Funktionspakete („Packages“)
- und vieles mehr...

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function Addins

Project: (None)

02_prepare_data.R prepare_survey_data.R Untitled1* ScoringGuide 02_LimeSurvey-API.R

```

1 #####
2 # Prepare Environment
3 #####
4 setwd("C:/Users/Steffen/Arbeit/opencampus/99_source-code/elearning-statistics/00_functions")
5 source("02_LimeSurvey-API.R") # import LimeSurvey API
6 setwd("../..../auswertung pre-challenge")
7
8 library(inspectdf)
9 library(dplyr)
10 library(readr)
11 library(stringr)
12
13
14 #####
15 # Import Answer Data
16 #####
17 introData <- as_tibble(ls.getResponses(615549, sHeadingType = "code")) # import of the Introductory Questionnaire
18 closingData <- as_tibble(ls.getResponses(722958, sHeadingType = "code")) # import of the Closing Questionnaire
19
20
21 #####
22 # Prepare Table for Scoring
23 #####
24 # Shrink Datasets to Knowledge Items and join them
25 introItems <- select(introData, matches("Q2|Q3"))
26 closingItems <- select(closingData, matches("Q2|Q3"))
27 # shorten name to exclude questionnaire ID
28 names(introItems) <- substr(names(introItems), 3, 10)
29 names(closingItems) <- substr(names(closingItems), 3, 10)
30 # join items according to question IDs
31 items <- full_join(introItems, closingItems)
32
33 # construct table including all given answers, how often they occurred (percentage and count) and include corresponding
34 scoreTable <- get.scoreTable(items)
35 write_csv2(file="ScoringGuide_new.csv", scoreTable, row.names = F, na="")
36 scoreTable <- read_csv2(file="ScoringGuide_new.csv")
37

```

Environment History Connections

Global Environment

Data

Object	Size
closingData	46 obs. of 82 variables
closingItems	46 obs. of 27 variables
config.ls	List of 3
introData	54 obs. of 70 variables
introItems	54 obs. of 17 variables
items	100 obs. of 27 variables
scoredAnswers	54 obs. of 17 variables
scoreTable	274 obs. of 5 variables
scoreTable_old	248 obs. of 5 variables
ScoringGuide	274 obs. of 5 variables

Values

Object	Size
posScores	int [1:274] NA NA 31 NA NA 31 NA NA 31 NA ...

Functions

Object	Size
base64_to_df2	function (x, sep = ";")
check.scoreTable	function (scoreTable, verbose = TRUE)
get.scoreTable	function (items)
ls.addParticipants	function (surveyID, idset, testUser = TRUE, ...)

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

Name	Size	Modified
.conda		
.keras		
.RData	2.5 KB	Oct 10, 2019, 4:11 PM
.Rhistry	2 KB	Oct 17, 2019, 2:57 PM
190424_Umlaufbeschluss Fenster Schae...	546.5 KB	Apr 24, 2019, 8:21 PM
Benutzerdefinierte Office-Vorlagen		
Citavi 6		
Dummy		
Fax		
Meine Datenquellen		
Outlook-Dateien		
PDF-Archiv		

12:1 (Untitled) R Script

Console Terminal Jobs

C:/Users/Steffen/Arbeit/opencampus/99_source-code/auswertung pre-challenge/

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Help

Go to file/function

Run

Source on Save

Source

02_prepare_data.R x prepare_survey_data.R x Untitled1* x ScoringGuide x 02_LimeSurvey-API.R x

```

1 #####
2 # Prepare Environment
3 #####
4 setwd("c:/Users/steffen/Arbeit/opencampus/99_source-code/elearning-statistics/00_functions")
5 source("02_LimeSurvey-API.R") # import LimeSurvey API
6 setwd("../..../auswertung pre-challenge")
7
8 library(inspectdf)
9 library(dplyr)
10 library(readr)
11 library(stringr)
12
13
14 #####
15 # Import Answer Data
16 #####
17 introData <- as_tibble(lis.getResponses(615549, sHeadingType = "code")) # import of the Introductory Questionnaire
18 closingData <- as_tibble(lis.getResponses(722958, sHeadingType = "code")) # import of the Closing Questionnaire
19
20
21 #####
22 # Prepare Table for Scoring
23 #####
24 # Shrink Datasets to Knowledge Items and join them
25 introItems <- select(introData, matches("Q2|Q3"))
26 closingItems <- select(closingData, matches("Q2|Q3"))
27 # shorten name to exclude questionnaire ID
28 names(introItems) <- substr(names(introItems), 3, 10)
29 names(closingItems) <- substr(names(closingItems), 3, 10)
30 # join items according to question IDs
31 items <- full_join(introItems, closingItems)
32
33 # construct table including all given answers, how often they occurred (percentage and count) and include correspondi
34 scoreTable <- get.scoreTable(items)
35 write.csv2(file="scoringGuide_new.csv", scoreTable, row.names = F, na="")
36 scoreTable <- read_csv2(file="scoringGuide_new.csv")
37

```

37:1 (Untitled) R Script

Console Terminal Jobs

C:/Users/Steffen/Arbeit/opencampus/99_source-code/auswertung pre-challenge/

Environment History Connections

Global Environment

Data

Object	Size
closingData	46 obs. of 82 variables
closingItems	46 obs. of 27 variables
config.ls	List of 3
introData	54 obs. of 70 variables
introItems	54 obs. of 17 variables
items	100 obs. of 27 variables
scoredAnswers	54 obs. of 17 variables
scoreTable	274 obs. of 5 variables
scoreTable_old	248 obs. of 5 variables
ScoringGuide	274 obs. of 5 variables

Values

Object	Size
posScores	int [1:274] NA NA 31 NA NA 31 NA NA 31 NA ...

Functions

Object	Size
base64_to_df2	function (x, sep = ";")
check.scoreTable	function (scoreTable, verbose = TRUE)
get.scoreTable	function (items)
ls.addParticipants	function (surveyID, idset, testuser = TRUE, aPartic

Files Plots Packages Help Viewer

New Folder Delete Rename More

Home

Name	Size	Modified
.conda		
.keras		
.RData	2.5 KB	Oct 10, 2019, 4:11 PM
.Rhistry	2 KB	Oct 17, 2019, 2:57 PM
190424_Umlaufbeschluss Fenster Schauen...	546.5 KB	Apr 24, 2019, 8:21 PM
Benutzerdefinierte Office-Vorlagen		
Citavi 6		
Dummy		
Fax		
Meine Datenquellen		
Outlook-Dateien		
PDF Architect		

AUFGABEN

- Finde heraus wie man ein R-Package installiert und installiere das Package „fortunes“
- Finde heraus, was es macht.
- Berechne mit Hilfe von R den Mittelwert der folgenden beiden Zahlen:
9,87654321 und 1,23456789

AUFGABEN

- **Google!**
Google!

- Stackoverflow

- Cheatsheets

- Dokumentation der Funktionen in R

DATENSTRUKTUREN

Es gibt drei Grundtypen für ein Datum

- Boolean (TRUE / FALSE)
- Numeric (1.1392)
- String ("Text")

und zusätzlich abgeleitete, speziellere Typen (integer, date, ...)

- Integer (Untertyp von Numeric; 12)
- Date (Untertyp von Numeric; "2019-04-11")
- Factor (Untertyp von String; "female"/"male")

VEKTOREN

Alle Elemente eines Vektors haben den gleichen Typ

→ Jeder Vektor hat einen eindeutigen Typ

- `v1 <- c(FALSE, TRUE, TRUE)`
- `ort <- c("kiel", "hamburg", "berlin")`
- `v2 <- c(112, 343, 235)`
- `alter <- 12`

MATRIZEN

Bestehen aus Vektoren gleicher Länge und gleichen Typs

→ Jede Matrix hat einen eindeutigen Typ.

- `m <- matrix(c(1,2,3,4), 2, 2)`

entspricht: $\begin{pmatrix} 1 & 3 \\ 2 & 4 \end{pmatrix}$

LISTEN

Können beliebige Elemente enthalten:

- Vektoren und Matrizen unterschiedlichen Typs
- sonstige R-Objekten (etwa auch Funktionen)

→ Listen haben keinen eindeutigen Typ

- `l1 <- list(v1, alter, c(4,7,9))`
- `l2 <- list(name="Fred", child.ages=c(4,7,9))`

SPEZIELLE LISTEN

Datentabellen

bestehen aus Vektoren gleicher Länge
(aber potentiell unterschiedlichen Typs)

→ Datentabellen haben keinen eindeutigen Typ

- ☐ Data Frame (base Package)
- ☐ Tibble (tidyverse Package)
- ☐ Data Table (data.table Package)

SELEKTION VON DATEN

Am Beispiel des Data Frames zu `mtcars`

Selektion einer Spalte als Vektors über den Namen

- `mtcars$mpg`

Selektion einer Spalte als *Vektor* über die Position

- `mtcars[,1]` (oder `mtcars[[1]]`)

Selektion einer Zeile als *Data Frame* über die Position

- `mtcars[1,]`

SELEKTION VON DATEN

Am Beispiel des Data Frames zu `mtcars`

Selektion einer Spalte als Vektors über den Namen

- `mtcars$mpg`

Selektion einer Spalte als *Vektor* über die Position

- `mtcars[,1]` (oder `mtcars[[1]]`)

Selektion einer Zeile als *Data Frame* über die Position

- `mtcars[1,]`

SELEKTION VON DATEN

Selektion von Daten über einen Vektor vom Typ Boolean

[1] Konstruktion des Vektors

- `mtcars$hp < 100`
- `mtcars$gear == 5`

[2] Selektion der Fälle (Zeilen) mit dem Wert TRUE

- `mtcars[mtcars$hp<100,]`
- `mtcars[mtcars$gear==5,]`

AUFGABEN

- Speichere den Datensatz `airquality` in der Variable `airQuality`.
- Berechne die Gesamtdurchschnittstemperatur.
- Berechne die Durchschnittstemperatur für den Monat Juli.
- Vergleiche, ob die Monate Juli und Mai sich signifikant in ihrer Durchschnittstemperatur unterscheiden.

YouTube DE Suchen ANMELDEN

RStudio File Edit Code View Plots Session Build Debug Tools Window Help

Environment History

Global Environment

Values

x	7
y	3

Files Plots Packages Help Viewer

Name	Description	Version
<input type="checkbox"/> proxy	Distance and similarity measures	0.4-15
<input type="checkbox"/> pryr	Tools for Computing on the Language	0.1.2
<input type="checkbox"/> qpcR	Modelling and analysis of real-time PCR data	1.4-0
<input type="checkbox"/> qqman	Q-Q and manhattan plots for GWAS data	0.1.2
<input type="checkbox"/> quanteda	Quantitative Analysis of Textual Data	0.9.4
<input type="checkbox"/> quantreg	Quantile Regression	5.2.1
<input type="checkbox"/> R6	Classes with Reference Semantics	2.1.2
<input type="checkbox"/> randomForest	Breiman and Cutler's Random Forests for Classification and Regression	4.6-12
<input type="checkbox"/> RBGL	An interface to the BOOST graph library	1.46.0
<input type="checkbox"/> RColorBrewer	ColorBrewer Palettes	1.1-2
<input type="checkbox"/> Rcpp	Seamless R and C++ Integration	0.12.4

Introduction to Rstudio (R Basics #1)

2.278 Aufrufe • 10.05.2016

31 0 TEILEN SPEICHERN

Nächstes Video

AUTOPLAY

Tutorial - Getting Data into RStudio

Steve Grambow

<https://www.youtube.com/watch?v=QvBo1RBptvY>

OPENCAMPUS.sh

`code`cademyCatalogPricingTry Pro For Free🔍🔔🧑‍🎓

🔄 Reset Progress

Learn R

R is a popular language used by data scientists and researchers. If you are working with data, R is a fantastic language to learn.

START

OverviewSyllabus

1 Learn R: Introduction

>

</>

Introduction to R Syntax

Interactive Lesson

50%

PRO

PRO

Calculating Population Change Over Time with R

Freeform Project


PRO



PRO

Introduction to R

Multiple Choice Quiz

<https://www.codecademy.com/learn/learn-r>






[Learn](#)
[Practice](#)
[Projects](#)
[Assessment](#)
[Pricing](#)
[For Business](#)
1,957 XP



INTERACTIVE COURSE

Introduction to R

Continue Course



🕒 4 hours |
 ▶ 0 Videos |
 📄 62 Exercises |
 👤 1,373,478 Participants |
 📊 6,200 XP |
 Download the app:
 


Course Description

In Introduction to R, you will master the basics of this widely used open source language, including factors, lists, and data frames. With the knowledge gained in this course, you will be ready to undertake your first very own data analysis. Oracle estimated over 2 million R users worldwide in 2012, cementing R as a leading programming language in statistics and data science. Every year, the number of R users grows by about 40%, and an increasing number of organizations are using it in their day-to-day activities. Begin your journey to learn R with us today!

1 Intro to basics

13%

Take your first steps with R. In this chapter, you will learn how to use the console as a calculator and how to assign variables. You will also get to know the basic data types in R. Let's get started.

[VIEW CHAPTER DETAILS](#)

Continue Chapter

2 Vectors

0%

We take you on a trip to Vegas, where you will learn how to analyze your gambling results using vectors in R. After completing this chapter, you will be able to create vectors in R, name them, select elements from them, and compare

This course is part of these tracks:

Data Analyst with R
Data Scientist with R
R Programmer
R Programming



Jonathan Cornelissen
Co-founder of DataCamp

Jonathan Cornelissen is one of the co-founders of DataCamp, and is interested in everything related to data

<https://www.datacamp.com/courses/free-introduction-to-r>