Introduction to R

Jan-Philipp Kolb

05 Mai, 2019

Introduction round

- Where are you from? What are you studying/working?
- What are your expectations of this course?
- Where do you think you can use Machine Learning in the future?

Preliminaries

- This topic is huge - we concentrate on presenting the application in R
- Usually we have big differences in knowledge and abilities of the participants - please tell, if it is too fast or slow.
- We have many **exercises** because at the end you can only learn on your own
- We have many **examples** - try them out
- If there are questions - always ask
- R is more fun together - ask your neighbor

Why R is a good choice . . .

- . . . because it is an **open source language**
- . . . outstanding graphs - **graphics**, **graphics**, **graphics**
- . . . relates to other languages - **R can be used in combination with other programs** - e.g. **data linking**
- . . . R can be used **for automation**
- . . . Vast Community - **you can use the intelligence of other people ;-)**
- . . .
- Because of the large comunity
- New statistical methodologies are implemented quite fast
- Because R can be combined with other programs like Postgresql or Python

Constraints

Newer modules in Python

- Machine learning is a field that changes rapidly.
- Some new tools are first developed in Python.
- The package reticulate offers the possibility to use these modules from an R environment.
- Good news - Python is also Open Source

Big Data

- Especially if you work with web data, you quickly have to deal with large amounts of data.
- Therefore one must fall back on databases, which can be used in combination with R.

Content of this part

- Introduction to programming in R

what is relevant for this course.

- How to import data?
- What to do with missing values?
- Parallelization

Import data

Using a path to import data

The titanic dataset

X	pclass	survived	name	sex	age	sibsp	parch	ticket	fare
1	1	1	Allen, Miss. Elisabeth Walton						
2	1	1	Allison, Master. Hudson Trevor						
3	1	0	Allison, Miss. Helen Loraine						
4	1	0	Allison, Mr. Hudson Joshua Creighton						
5	1	0	Allison, Mrs. Hudson J C (Bessie Waldo Daniels)						
6	1	1	Anderson, Mr. Harry						

The function `scan` to import data

The R-package data.table

Get an overview

```
##   Ozone Solar.R Wind Temp Month Day
## 1    41     190  7.4   67     5   1
## 2    36     118  8.0   72     5   2
## 3    12     149 12.6   74     5   3
## 4    18     313 11.5   62     5   4
## 5    NA      NA 14.3   56     5   5
## 6    28      NA 14.9   66     5   6
```

Overview with data.table

```
## Warning: package 'data.table' was built under R version
##      Ozone Solar.R Wind Temp Month Day
##    1:    41     190  7.4   67     5   1
##    2:    36     118  8.0   72     5   2
##    3:    12     149 12.6   74     5   3
##    4:    18     313 11.5   62     5   4
##    5:    NA      NA 14.3   56     5   5
```

Exercise

- Compute the logarithm of `x`, return suitably lagged and iterated differences,
- compute the exponential function and round the result

```
## [1]  3.3  1.8  1.6  0.5  0.3  0.1 48.8  1.1
```

The pipe operator

```
## [1]  3.3  1.8  1.6  0.5  0.3  0.1 48.8  1.1
```

How to deal with missing values

```
##      Ozone Solar.R Wind Temp Month Day
##   1:    41     190  7.4   67     5   1
##   2:    36     118  8.0   72     5   2
##   3:    12     149 12.6   74     5   3
##   4:    18     313 11.5   62     5   4
##   5:    NA      NA 14.3   56     5   5
##  ---
## 149:    30     193  6.9   70     9  26
## 150:    NA     145 13.2   77     9  27
## 151:    14     191 14.3   75     9  28
## 152:    18     131  8.0   76     9  29
## 153:    20     223 11.5   68     9  30

##      Ozone Solar.R Wind Temp Month Day
##   1:    41     190  7.4   67     5   1
##   2:    36     118  8.0   72     5   2
##   3:    12     149 12.6   74     5   3
```

Clean the titanic data set

```
## Warning: package 'dplyr' was built under R version 3.5.

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:data.tab
##
##     between, first, last

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union

## Warning: package 'bindrcpp' was built under R version 3
```

Get an overview of the data

```
## Observations: 1,045
## Variables: 13
## $ X         <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12,
## $ pclass    <fct> Upper, Upper, Upper, Upper, Upper, Upp
## $ survived  <fct> Yes, Yes, No, No, No, Yes, Yes, No, Ye
## $ name      <fct> Allen, Miss. Elisabeth Walton, Allison
## $ sex       <fct> female, male, female, male, female, ma
## $ age       <dbl> 29.0000, 0.9167, 2.0000, 30.0000, 25.0
## $ sibsp     <int> 0, 1, 1, 1, 1, 0, 1, 0, 2, 0, 1, 1, 0,
## $ parch     <int> 0, 2, 2, 2, 2, 0, 0, 0, 0, 0, 0, 0, 0,
## $ ticket    <fct> 24160, 113781, 113781, 113781, 113781,
## $ fare      <dbl> 211.3375, 151.5500, 151.5500, 151.5500
## $ cabin     <fct> B5, C22 C26, C22 C26, C22 C26, C22 C26
## $ embarked  <fct> S, S, S, S, S, S, S, S, S, C, C, C, C,
## $ home.dest <fct> St Louis, MO, Montreal, PQ / Chestervi
```

Example Data - Housing Values in Suburbs of Boston

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	bla
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296		
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242		
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242		
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222		
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222		
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222		

Normalize your data

Set a seed

Time measurement

```
## Time difference of 0.790045 secs
```

How many cores are available

```
## Warning: package 'doParallel' was built under R version

## Loading required package: foreach

## Warning: package 'foreach' was built under R version 3.5

## Loading required package: iterators

## Loading required package: parallel

## [1] 4
```

Make cluster

```
## Time difference of 0.7690439 secs
```

The swirl package

Resources

- Course materials for the Data Science Specialization