Data Mining and Visualization – Overview

August 29, 2019

Motto

In God we trust, all others bring data

attributed to
William Edwards Deming[*] **(1900-1993)**

[*] American statistician, known, in particular, for promotion of statistical methods in industrial production and management

Instructor, Assistant, Webpage, etc.

- Instructor: Krzysztof (Krys) Podgórski
- The easiest way to communicate is through e-mail: **Krzysztof.Podgorski@stat.lu.se**
- Webpage: https://krys.neocities.org
- Assistant: Johan Larsson, e-mail: **Johan.Larsson@stat.lu.se**
- The main source of the information about the course: webpage – visit it frequently as the material that is posted there will change.
- The web address is

  https://krys.neocities.org/Teaching/DataMining/DataMining.html
- Office hours each Monday between 16:00 and 17:00 or by appointment

Course Organization

- Syllabus is available in the form of a the webpage (print it on your own if a hard copy is more convenient for you).
- Three parts of a big comprehensive “examination”:

Assignments, **Projects**, **Presentation**

- Assignments – Individually at home plus discussion in the classroom
- Computer Projects – In groups of two: decide with whom would you like to work, mostly done in the computer lab. If not completed then it can be finished at home but then a printed report showing that the tasks have been completed have to be submitted. The same applies in the case of absence from a lab session.
- Presentation – 30min presentation of the own study based on a chosen data set (this will be discussed in detail when halfway through the course), a written rapport after the presentation needs to be submitted.

Grade

- For each of the three parts score will be assigned on the scale from 0-100
- This score contributes equally to the total score which is computed according to the formula:

$$T = (S_1 + S_2 + S_3)/3,$$

  where S_1, S_2, S_3 represent scores for the corresponding parts.
- The final grade will be assigned according to the table:

Percentage	**Grade**
49 - 0	F
54 - 50	E
64 - 55	D
74 - 65	C
84 - 75	B
100 - 85	A

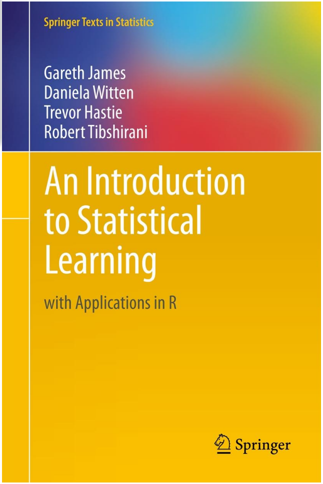
Assignments – cover basics and main topics of the course

- The total of **five assignments**, that are based on the covered course material although they do not cover the material completely.
- They will comprise of a set of simple questions that will help to clarify introduced topics.
- Some questions will also help in preparing to the lab sessions.
- Assignments will be worked out at home and **due the week following the date they are posted in our schedule**.
- They should be submitted in an electronic form (scanning of the text is allowed) through **CANVAS**.
- They will be available at the webpage but there can be last minute changes thus, please, download a copy only on the date at which they are listed in the schedule.

Course contents and the textbook

An Introduction to Statistical Learning with Applications in R,
by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani;
1st edition; Springer; 2013.

Cover 2.jpg

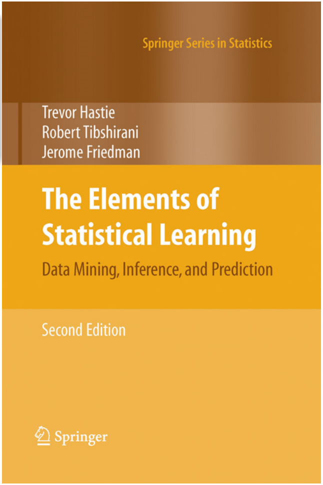


- This is a simplified version of the book that is the most popular and most comprehensive monograph on the statistical data mining (see the next slide).
- According to the authors, and concurred by many, the material of the book could serve as basis for a sequence of courses equivalent to 30hp.
- For us to adopt it within a 7.5hp course frame, significant reduction of the material is needed.
- For the same reason the course will not follow the 'chronology' of the textbook.

The ‘other’ textbook

**The Elements of Statistical Learning,
Data Mining, Inference, and Prediction
by Trevor Hastie, Robert Tibshirani, and Jerome Friedman;
2nd edition; Springer.**

- As mentioned, this is by far the most popular and most comprehensive monograph on the subject.
- Originally the course was designed with this book in mind.
- However, publication of its simplified version which is easier to follow within a regular course frame, shifted this monograph into the background of our presentation.
- Some of the final lectures are using the material that is only available in this textbook thus it contains all the material needed for the course.

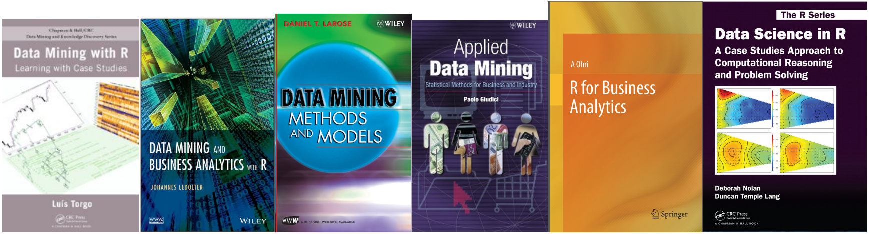

Springer Series in Statistics
Trevor Hastie
Robert Tibshirani
Jerome Friedman
The Elements of Statistical Learning
Data Mining, Inference, and Prediction
Second Edition
Springer

Access to the textbook

- The authors of *‘An Introduction to Statistical Learning’* have created a **webpage** on which the book is accessible in the electronic format.
- Hard copy is also accessible at Springer.
- The mentioned webpage contains also other useful material such as programs, data, and erratum.
- The ‘other’ textbook, *‘The Elements of Statistical Learning’* is also available in the electronic form through Lund University’s **online library** due to the agreement of the university with Springer.

Supplementary books

There is a plethora of other books published recently on **Data Mining, Big Data, Data Analytics** – a consequence of growing interests in the methodology. The following are examples (we may occasionally take a look into some of them, especially when the final projects are discussed):



Depending on your background, some of them can be more accessible than others.

Lab projects – working with data

- Analysis of the data using the methods discussed in the lecture will demonstrated through **five projects** scheduled throughout the course. The projects will require using statistical software. At the moment the most popular data mining softwares are **R** and **Python**, although commercial such as Matlab are also often utilized.
- R-package – free and very popular statistical package, very good for statistical computing although less compelling in handling large matrices and multivariate visualization
- We have opted for R-package to present analyses the data and the methods of data mining.
- The choice has been dictated by a large number of supporting materials in R that are available for illustration of data mining methods.

Downloading R-package and first steps

- Statistical R-package available for free download here
- Available on any PC platform (Mac, Windows, Linux).
- Worry free and fast downloading procedure (a couple of minutes).
- We will be working in the command line window of R (most direct way of accessing R-package).
- No experience is required – all of the code that will be needed will be provided on our webpage!
- There some so-called R front-ends (such *R Commander* or *R-Studio* or *Jupyter*) that ease writing more complex programming in R
- **only a very basic R installation** with the primitive *copy-and-paste-to-the-command-line* approach is truly needed as a method of running the programs. **You can do better if you wish so!**

Example of a very simple R session

- Suppose the following data are in file `Table2_1.txt` located at directory `/Lecture1/Table2_1.txt`

```
0.51   0.51   0.51   0.50   0.51   0.49   0.52   0.53   0.50   0.47
0.51   0.52   0.53   0.48   0.49   0.50   0.52   0.49   0.49   0.50
0.49   0.48   0.46   0.49   0.49   0.48   0.49   0.49   0.51   0.47
0.51   0.51   0.51   0.48   0.50   0.47   0.50   0.51   0.49   0.48
0.51   0.50   0.50   0.53   0.52   0.52   0.50   0.50   0.51   0.51
```

- Then the following code reads the data and computing its mean and standard deviation:

```
#Getting data in a vector
x=scan("/Data/Table2_1.txt")
mean(x)
#[1] 0.4998
sd(x)
#[1] 0.01647385
```

- # at the beginning of the line denotes a commentary (so the following characters are not interpreted by R when the lines are copied to the command line).

The same session through webpage

- We go through steps of running this session using our webpage.

  Step One Create a directory in which you intend to work with your programs.

  Step Two Download data that you intend to work on to this directory.

  Step Three Download *R*-code the same directory and open the file by a convenient text editor of your choice.

  Step Four Open *R* and by copying and pasting run the programs from the code.

  Step Five Interpret the results.

Final project/presentation – a study of a real data set

The final presentation will discuss a study of real data by means of learned data mining methods.

- Data Description – data, with a detailed description will be provided
- Problem Formulation – initial problem will be suggested in the description of the data, precise formulation will be required
- Analysis Methods – the choice of analytical tools and their discussion is expected
- Result Interpretation – analysis of the data has to be performed, results have to be presented and interpreted, final conclusions and suggestions need to be formulated

Concluding thought

We are drowning in information and starving for knowledge

Rutherford D. Roger[*]

[*] American librarian (Yale University)