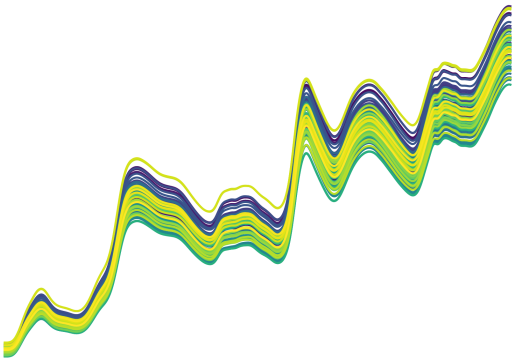
INFOMDA2: Course Syllabus

Erik-Jan van Kesteren



Course Description

The ever-growing influx of data allows us to develop, interpret and apply an increasing set of learning techniques. However, with this increase in data comes a challenge: how to make sense of the data and identify the components that really matter in our modeling efforts. This course gives a detailed and modern overview of statistical learning with a specific focus on high-dimensional data.

In this course we emphasize the tools that are useful in solving and interpreting modern-day analysis problems. Many of these tools are essential building blocks that are often encountered in statistical learning. We also consider the state-of-the-art in handling machine learning problems. We will not only discuss the theoretical underpinnings of supervised learning, but focus also on the skills and experience to rapidly apply these techniques to new problems.

During this course, participants will actively learn how to apply the main statistical methods in data analysis and how to use machine learning algorithms and visualization techniques, especially on high-dimensional data problems. The course has a strongly practical, hands-on focus: rather than focusing on the mathematics and background of the discussed techniques, you will gain hands-on experience in using them on real data during the course and interpreting the results.

Course Objectives

At the end of this course, students are able to apply and interpret the theories, principles, methods and techniques related to contemporary data science and understand and explain different approaches to data analysis:

- apply data visualization and dimension reduction techniques on high dimensional data sets
- apply, implement, understand and explain methods and techniques that are associated with advanced data modeling, including regularized regression, principal components, correspondence analysis, neural networks, clustering, time series, text mining and deep learning
- evaluate the performance of these techniques with appropriate performance measures.
- select appropriate techniques to solve specific data science problems
- motivate and explain the choice for techniques to investigate data problems
- implement and understand generic data science tools, such as model evaluation, visualization and validation techniques
- interpret and evaluate the results of analyses and explain these techniques in simple terminology to a broad audience
- understand and explain the principles of high-dimensional data visualization and the grammar of graphics.
- construct appropriate visualizations for each data analysis technique in R

Required Readings

Sections from the following freely available books:

- **ISLR**: James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning* (2nd ed.). Springer. statlearning.com
- **SLS**: Hastie, T., & Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparsity* CRC Press. web.stanford.edu/~hastie/StatLearnSparsity
- **R4DS**: Wickham, H., & Grolemund, G. (2016). *R for data science: import, tidy, transform, visualize, and model data.* O'Reilly Media, Inc. r4ds.had.co.nz
- **FPP3** : Hyndman, R. J. & Athanasopoulos, G. (2021). *Forecasting: Principles and Practice (3rd ed.)* Otexts. otexts.com/fpp3
- **TTMR**: Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach.* O'Reilly Media, Inc. tidytextmining.com

Several freely available articles.

Required Software

In this course, we will exclusively use R & RStudio for data analysis. First, install the latest version of R for your system (see `https://cran.r-project.org/`). Then, install the latest (desktop open source) version of the RStudio integrated development environment (`link`).

We will make extensive use of the `tidyverse` suite of packages, which can be installed from within `R` using the command `install.packages("tidyverse")`.

Course Policy

Weekly course flow

- There will be a lecture and a Q&A session each week.
- The required readings should be read before the lecture. These are *not* optional.
- Each practical needs to be handed in before the next lecture.
- In the Q&A sessions you can ask for help on the practical assignments.

Grading policy

- To develop the necessary skills for completing the assignments and the exam, 9 `R` practicals must be made and submitted. These exercises are not graded, but students must fulfill them to pass the course.
- There are two pass/fail assignments. Successful and timely completion of these assignments will grant you a bonus point on the exam. You make and submit these as a group.
- **100%** of your grade will be determined by an exam featuring both knowledge questions as well as practical data analysis skills in `R`. Some example questions will be made available to you so you can prepare.

Class Schedule

Key dates and deadlines

Day	Date	Time	Description
Wednesday	11-11-2020	13:15 - 15:00	Lecture 1
Friday	13-11-2020	11:00 - 12:45	Q&A 1
Wednesday	18-11-2020	13:15 - 15:00	Lecture 2
Friday	20-11-2020	11:00 - 12:45	Q&A 2
Wednesday	25-11-2020	13:15 - 15:00	Lecture 3
Friday	27-11-2020	11:00	Deadline assignment 1 EDA
Friday	27-11-2020	11:00 - 12:45	Q&A 3
Wednesday	02-12-2020	13:15 - 15:00	Lecture 4
Friday	04-12-2020	11:00 - 12:45	Q&A 4
Wednesday	09-12-2020	13:15 - 15:00	Lecture 5
Friday	11-12-2020	11:00 - 12:45	Q&A 5
Wednesday	16-12-2020	13:15 - 15:00	Lecture 6
Friday	18-12-2020	11:00 - 12:45	Q&A 6
Wednesday	13-01-2021	13:15 - 15:00	Lecture 7
Friday	15-01-2021	11:00 - 12:45	Q&A 7
Wednesday	20-01-2021	13:15 - 15:00	Lecture 8
Friday	22-01-2021	11:00	Deadline assignment 2 Prediction
Friday	22-01-2021	11:00 - 12:45	Q&A 8
Wednesday	27-01-2021	13:15 - 15:00	Lecture 9
Friday	29-01-2021	11:00 - 12:45	Q&A 9
Friday	05-02-2021	13:30 - 16:30	Exam
Friday	05-03-2021	TBD	Resit

Lecture 1: Introduction & the bet on sparsity

`11-11-2021 | 13:15 - 15:00`

Required reading

- This syllabus
- The course website uudav.nl
- ISLR paragraph 6.2 shrinkage methods
- ISLR paragraph 6.4 considerations in high dimensions
- Review the tidyverse style guide `link`

Practical preparation

Read the prerequisites on the practicals website uudav.nl. Install R and RStudio as per the instructions there.

Lecture 2: Dimension reduction

`18-11-2020 | 13:15 - 15:00`

Required reading

- ISLR paragraph 6.3 dimension reduction methods
-

Lecture 3: Regression with dimension reduction

`25-11-2020 | 13:15 - 15:00`

Required reading

- ISLR:

Pass/fail assignment 1

`Link`. Hand in on blackboard **before** practical 3 (`27-11-2020 | 11:00`).

Lecture 4: Deep learning

`02-12-2020 | 13:15 - 15:00`

Required reading

- ISLR chapter 10 deep learning

Lecture 5: Supervised learning: classification (1)

`09-12-2020 | 13:15 - 15:00`

Required reading

- ISLR:
  - Paragraph 2.2.3
  - Chapter 4 (except 4.4.4 on QDA)
  - Paragraph 5.1.5

Lecture 6: Supervised learning: classification (2)

`16-12-2020 | 13:15 - 15:00`

Required reading

- ISLR:
  - Paragraph 2.2.3
  - Chapter 4 (except 4.4.4 on QDA)
  - Paragraph 5.1.5
  - Paragraph 8.1 & 8.2

Optional reading

ISLR Chapter 9 & Paragraph 4.4.4

Winter break

Lecture 7: Supervised learning: Nonlinear extensions

`13-01-2021 | 13:15 - 15:00`

Required reading

- ISLR:
  - Paragraph 3.5
  - Chapter 7

Lecture 8: Unsupervised learning: principal components and correspondence analysis

`20-01-2021 | 13:15 - 15:00`

Required reading

- ISLR: paragraph 10.1, 10.2 and 10.4
- Greenacre, biplots in practice: Chapter 8, 9 and 10
- Van der Heijden: "An Extended Study into the Relationship between Correspondence Analysis and Latent Class Analysis" (*only the example*, the rest is optional reading)

Pass/fail assignment Prediction

Link. Hand in on blackboard **before** practical 8 (`22-01-2021 | 11:00`).

Lecture 9: Unsupervised learning: cluster analysis

`27-01-2021 | 13:15 - 15:00`

Required reading

- ISLR Chapter 10 paragraphs 10.3 and 10.5
- Kumar Chapter 8 Cluster analysis: all paragraphs through 8.2.1

Optional reading: - Kumar Chapter 6: Association analysis all paragraphs through 6.2.2

Exam

`05-02-2021 | 13:30 - 16:30`

Resit

Target date: `05-03-2021`, to be confirmed.