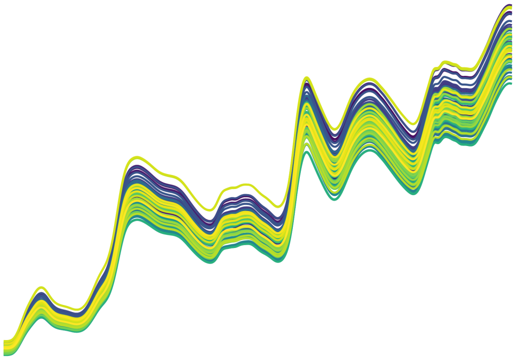
INFOMDA2: Course Syllabus

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 different techniques, but focus also on the skills and experience needed 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.

Prerequisites

The course INFOMDA1 (or equivalent) serves as a sufficient entry requirement for this course. For information about the contents of the INFOMDA1 course, refer to its course website.

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 to understand and explain different approaches to data analysis:

- apply data visualization and dimension reduction techniques on high dimensional data sets
- 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.
- interpret and evaluate the results of (high-dimensional) data analyses and explain these techniques in simple terminology to a broad audience.
- understand and explain the principles of high-dimensional data analysis and visualization.
- construct appropriate visualizations for each data analysis technique in R.

Required Readings

Freely available sections from the following 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
- **ESL**: Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction.* New York: Springer Verlag. web.stanford.edu/~hastie/ElemStatLearn.
- **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
- **MBCC**: Bouveyron, C., Celeux, G., Murphy, T., & Raftery, A. (2019). *Model-based Clustering and Classification for Data Science: With Applications in R* Cambridge University Press. cambridge.org
- **TTMR**: Silge, J., & Robinson, D. (2021). *Text mining with R: A tidy approach.* O'Reilly Media, Inc. tidytextmining.com
- **SLP3**: Jurafsky, D., Martin, J.H. (2021) *Speech and language processing.* (3rd ed.) https://web.stanford.edu/~jurafsky/slp3/
- Some freely available articles & chapters.

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 lab session each week. Both are in-person.
- The required readings should be read before the lecture. These are *not* optional.
- There are some take-home exercises to be done before each lab session; these will be discussed during the lab session.
- There are some additional exercises to be done during the lab session; the answers to these will be made available after the session.

- Hand-in of practicals and assignments is done on blackboard

In-person course policy

- INFOMDA2 is a fully offline course, with in-person lectures and lab sessions.
- We find it important for interactive and collaborative learning that the course is offline, hence there is no teams environment for this course.
- If you miss a session, e.g., due to sickness, you should catch up in the regular way:
  - Read the readings
  - Go through the lecture slides
  - Do the practicals
  - Ask your peers if you have questions
  - (after the above) ask the lecturer for further explanation
- We realize that events may occur during the course for which we will have to adapt (e.g., a pandemic)
- If (and only if) this leads to considerable problems, we will reconsider the in-person course policy and adjust to a hybrid / online setting

Grading policy

- To develop the necessary skills for completing the assignments and the exam, 8 `R` practicals must be made and handed in. These exercises are not graded, but students must fulfill them to pass the course.
- **25%** of your grade will be determined by a group assignment resulting in a report, which includes an intermediate peer feedback moment.
- **75%** of your grade will be determined by a final exam featuring both knowledge questions as well as practical data analysis skills in `R`. Some example questions will be made available in due time to you so you can prepare.

Class Schedule

You can find the up-to-date class schedule with locations on mytimetable.uu.nl.

Key dates and deadlines

Day	Date	Time	Location	Description
Wednesday	13-11-2024	13:15 - 15:00	DLT500 6.27	Lecture 1
Friday	15-11-2024	11:00 - 12:45	BBG 219	Lab 1
Wednesday	20-11-2024	13:15 - 15:00	DLT500 6.27	Lecture 2
Friday	22-11-2024	11:00 - 12:45	BBG 219	Lab 2
Wednesday	27-11-2024	13:15 - 15:00	DLT500 6.27	Lecture 3
Friday	29-11-2024	11:00 - 12:45	BBG 219	Lab 3
Wednesday	04-12-2024	13:15 - 15:00	DLT500 6.27	Lecture 4
Friday	06-12-2024	11:00 - 12:45	BBG 219	Lab 4
Wednesday	11-12-2024	09:00 - 10:45	DLT500 6.27	Lecture 5
Friday	12-12-2024	11:00 - 12:45	BBG 219	Lab 5
Wednesday	18-12-2024	13:15 - 15:00	DLT500 6.27	NO LECTURE
Friday	20-12-2024	11:00 - 12:45	BBG 219	NO LAB
Break				
Wednesday	08-01-2025	13:15 - 15:00	DLT500 6.27	Lecture 6
Friday	10-01-2025	11:00 - 12:45	BBG 219	Lab 6
Wednesday	15-01-2024	13:15 - 15:00	DLT500 6.27	Lecture 7
Friday	17-01-2024	11:00 - 12:45	BBG 219	Lab 7
Wednesday	22-01-2024	13:15 - 15:00	DLT500 6.27	Lecture 8
Friday	24-01-2024	11:00 - 12:45	BBG 219	Lab 8
Wednesday	29-01-2024	13:30 - 16:30	TBD	Exam
Some day	TBD	13:30 - 16:30	TBD	Resit

Lecture 1: Introduction & betting on sparsity with the LASSO

Required reading

- This syllabus
- ISLR section 6.2 shrinkage methods
- ISLR section 6.4 considerations in high dimensions
- SLS chapter 1

- SLS sections 2.1 - 2.4.2
- SLS chapter 4 until (not including) example 4.1.

Optional reading

- Review the tidyverse style guide

Lecture 2: Dimension reduction 1

Required reading

- ISLR section 12.1-12.2 (pp. 497-510)
- ISLR section 6.3
- ESL section 3.4 (pp. 66-67; skip subsections 3.4.2-3.4.4)
- ESL section 14.5 (pp. 534-536; skip subsections 14.5.2-14.5.5)
- ESL section 14.6 (pp. 553-554; skip subsection 14.6.1)
- ESL section 14.7 (pp. 557-570; skip subsection 14.7.3)

Lecture 3: Dimension reduction 2

Required reading

- ESL section 14.7 (pp. 557-570); subsection 14.7.3 can be skipped
- ESL section 14.6 (553-554); subsection 14.6.1 can be skipped
- Article: Hyvarinen, A.; Oja, E. (2000): “Independent Component Analysis: Algorithms and Application”, Neural Networks, 13(4-5):411-430. (Technical but pedagogical introduction; available on internet).
- Hofmann, T. (2001). Unsupervised Learning by Probabilistic Latent Semantic Analysis. Machine Learning, 42, 177-196.

Lecture 4: Clustering

Required reading

- ISLR section 12.4
- Tan, Steinbach, Karpatne, & Kumar (2019) Introduction to Data Mining section (second edition) section 7.5 cluster evaluation, available here

- SLS sections 8.5.1 and 8.5.2.

Optional reading

- The remainder of Introduction to Data Mining Chapter 7 Cluster analysis

Lecture 5: Model-based clustering

Required reading

- Mixture models: latent profile and latent class analysis by Daniel Oberski (2016) link
- MBCC sections 2.1 and 2.2

Optional reading

- MBCC sections 2.3, 2.4, 2.8
- MBCC chapter 8 (not freely available)

Lecture 6: Deep learning

Required reading

- ISLR chapter 10, up to and including section 10.3.3.

Winter break

Lecture 7: Text mining 1

Required reading

- TTMR chapter 6
- SLP3 sections 6.2, 6.3, 6.5.

Lecture 8: Text mining 2

Required reading

- SLP3 sections 6.8, 9.2, 9.3