INFOMDWR: Course Syllabus 2024

Erik-Jan van Kesteren Hakim Qahtan Ayoub Bagheri

2024-09-05

1 Introduction

Data do not fall from heaven, but are created, manipulated, transformed, and cleaned - in any data analysis, therefore, the treatment of the data itself is just as important as the modeling techniques applied to them. In this course, you will get acquainted with and implement a variety of techniques to go from raw data to analyses, visualizations and insights for science and business applications. This is an overview course designed to give you the tools and skills to use and evaluate data science methods.

The course consists of two parts, data wrangling and data analysis, which are intertwined.

Prerequisites

We assume that students who will join the course will have knowledge of statistics up to regression and analysis of variance, as well as some experience in programming in languages such as R and Python.

Objectives

At the end of this course, students have attained the following objectives:

- Know, explain, and apply data retrieval from existing relational and nonrelational databases, including text, using queries built from primitives such as select, subset, and join both directly in, e.g., SQL and through the python & R programming languages.
- Know, explain, and apply common data clean-up procedures, including missing data and the appropriate imputation methods and feature selection.
- Know, explain, and apply methodology to properly set-up data analysis experiments, such as train, validate, and test and the bias/variance trade-off.
- Know, explain, and apply supervised machine learning algorithms, both for classification and regression purposes as well as their related quality measures, such as AUC and Brier scores.

- Know, explain, and apply non-supervised learning algorithms, such as clustering and (other) matrix factorization techniques that may or may not result in lower-dimensional data representations.
- Be able to choose between the different techniques learned in the course and be able to explain why the chosen technique fits both the data and the research question best.

2 Course Policy

This course is worth 14 ECTS, which means it is designed to give a full-time workload.

Weekly course flow

A regular week in this course consists of three lectures (Monday-Wednesday morning) and three lab sessions (Monday-Wednesday afternoon). The material is introduced on a theoretical level in the lectures and then put into practice in the lab sessions. The practical work done in these labs is drawn from real life situations that allow the students to experience how to solve data science problems.

In addition, students will spend time during each week on bi-weekly take-home group assignments.

- The **lectures** are in-person. The required readings should be read before the lecture. These are *not* optional.
- The **lab sessions** are in-person interactive sessions in which you apply the methods you learn about in the lectures. The answers to the exercises in the labs are discussed at the end of each session.
- The skills acquired in the lectures and the labs provide the basis for doing the bi-weekly **take-home assignment**. This assignment is made in groups of 3-5 students and handed in via Brightspace.

Synchronous course policy

- INFOMDWR is an offline-first course, with mostly in-person lectures and lab sessions.
- We find it important for interactive and collaborative learning that the course is offline-first.
- 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 lab teacher for further explanation

Who to ask what

There are many teachers in this course. If you have questions, first **ensure the answer isn't in this syllabus** and then follow the table below:

Question type	How to ask
Course proceedings	Email course coordinators
Content - general	Email / ask lab teachers
Practical content	Email / ask lab teachers
Assignment content	Email / ask lab teachers
Lecture content	Email Lecturer (Hakim Qahtan / Daniel Oberski)

Grading policy

Your final grade in the course consists of the following grading components:

- Biweekly assignments (20%): every two weeks, there is a group assignment. Each assignment is graded and worth 5% of the final grade.
- Midterm exam (40%): Halfway through the course, there is a midterm exam. The content of this exam pertains to the first half of the course.
- Final exam (40%): At the end of the course, there is a final exam. The content of this exam emphasizes non-exclusively the teaching material of the second half of the course.

To pass the course:

- The weighted final grade of all grading components should be greater than or equal to 5.5.

Resit:

- If you obtain a final grade between 4.0 and 5.4, you are eligible for the resit.
- You can only retake one of the two exams. By default, this will be the one with the lowest grade. If you obtained the same grade for both, you will redo the final exam.
- The grade attained in the resit will replace the grade from the selected exam.

3 Course materials

Required Software

In this course, we will use a variety of software, but mainly SQLite, Python and R. Try to install both on your computer by the start of the course; we will also have a set-up computer lab on the first day to help you with this process.

Installing DB Browser for SQLite For the SQL parts, we recommend installing DB Browser for SQLite. Installation instructions for mac, windows, and linux can be found here.

Installing Python & Jupyter For the python parts of the course, we will use Google Colab, which is an interactive online notebook environment; this means no installation is necessary! However, you do need a google account, so make sure you have one (or make one specifically for the course).

Installing R & RStudio 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")`.

Required readings

Freely available sections from the following books:

Book	Title (Authors)	URL
`DBSC`	Database System Concepts (Silberschatz, Korth, Sudarshan)	db-book.com
`MMDS`	Mining Massive Datasets (Leskovec, Rajaraman, Ullman)	mmds.org
`PDA`	Python for Data Analysis, 3E (Wes McKinney)	wesmckinney.com/book/
`DMCT`	Data Mining: Concepts and Techniques (Han, Kamber, Pei)	Find it online for free.
`R4DS`	R for Data Science (Grolemund & Wickham)	r4ds.hadley.nz
`ISLR`	Introduction to Statistical Learning (James et al.)	statlearning.com
`DLBK`	Deep Learning (Goodfellow, Bengio, Courville)	deeplearningbook.org
`FIMD`	Flexible Imputation of Missing Data (van Buuren)	stefvanbuuren.name/fimd
`SLP3`	Speech and language processing (Jurafsky & Martin)	stanford.edu/~jurafsky/slp3
`TTMR`	Text mining with R: A tidy approach (Silge & Robinson)	tidytextmining.com/
`OMNG`	Operations Management 4th ed. (Reid & Sanders)	Find it online for free.

And (parts of) the following references:

- Oberski, D.L. (2016). Mixture models: Latent profile and latent class analysis. Modern statistical methods for HCI, 275-287. URL
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM computing surveys (CSUR), 54(6), 1-35. URL
- Hennig, C. (2015). Clustering strategy and method selection. arXiv preprint arXiv:1503.02059. URL
- Bouveyron, C., Celeux, G., Murphy, T. B., & Raftery, A. E. (2019). Model-based clustering and classification for data science: with applications in R (Vol. 50). Cambridge University Press.
- Some other freely available articles & chapters

4 Class Schedule

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

Week	Date	Topic	Type	Reading
1	2024-09-05	Introduction to the course	Lecture	Syllabus
1	2024-09-05	Introduction lab: setting up your computer	Lab	R4DS 3, 5, 7
1	2024-09-06	Data models	Lecture	DBSC 1, 3.2, 3.9
1	2024-09-06	Data models	Lab	
2	2024-09-09	Data extraction with SQL	Lecture	DBSC 2.6, 3.3 - 3.8
2	2024-09-09	Data extraction with SQL	Lab	
2	2024-09-10	Integrity constraints in databases	Lecture	DBSC 3.2.1, 4.4, 6.1, 6.2
2	2024-09-10	Integrity constraints in databases	Lab	
2	2024-09-11	Functional dependency	Lecture	DBSC 7.1-7.4.1
2	2024-09-11	Functional dependency	Lab	
3	2024-09-16	Indexing & data integration	Lecture	DBSC 14.1 - 14.7
3	2024-09-16	Indexing & data integration	Lab	
3	2024-09-17	Hetero. data analysis & string similarity	Lecture	MMDS 3.1-3.5

Week	Date	Topic	Type	Reading
3	2024-09-17	Hetero. data analysis & string similarity	Lab	
	2024-09-17	Social Event (De vagant)	Social	
3	2024-09-18	Data extraction in Python	Lecture	PDA 5, 6.1
3	2024-09-18	Data extraction in Python	Lab	
4	2024-09-23	Data preparation 1	Lecture	DMCT 3.1, 3.2, 12.1 - 12.4 , PDA 7.1
4	2024-09-23	Data preparation 1	Lab	
4	2024-09-23	10:00AM assignment 1	Deadline	
4	2024-09-24	Data preparation 2	Lecture	DMCT 3.3-3.5, PDA 7.2
4	2024-09-24	Data preparation 2	Lab	
4	2024-09-25	Data visualization	Lecture	R4DS 3, 5, 7 (optionally 2-8)
4	2024-09-25	Data visualization using ggplot	Lab	
5	2024-09-30	Exploratory data analysis	Lecture	R4DS 3, 5, 7 (optionally 2-8)
5	2024-09-30	Exploratory data analysis in R	Lab	
5	2024-10-01	Supervised learning: Regression	Lecture	ISLR 1, 2.1, 2.2, 3.1, 3.2, 3.3, 3.5
5	2024-10-01	Supervised learning: Regression models in R	Lab	
5	2024-10-02	Q&A	Lecture	
5	2024-10-02	No lab, time to study	Lab	
5	2024-10-04	Midterm exam	Exam	
6	2024-10-07	10:00AM assignment 2	Deadline	
6	2024-10-07	Supervised learning: model evaluation	Lecture	ISLR 5.1, 8.1, 8.2.1, 8.2.2, 8.2.3
6	2024-10-07	Supervised learning: model evaluation	Lab	
6	2024-10-08	Supervised learning: classification	Lecture	ISLR 4.1, 4.2, 4.3, 4.4.1, 4.4.2

Week	Date	Topic	Type	Reading
6	2024-10-08	Supervised learning: classification	Lab	
6	2024-10-09	Deep learning	Lecture	ISLR 10, DLBK 11, (optionally 6)
6	2024-10-09	Deep learning	Lab	
7	2024-10-14	Missing data 1: Mechanisms	Lecture	FIMD 1.1, 1.2, 1.3, 1.4
7	2024-10-14	Missing data mechanisms	Lab	
7	2024-10-15	Missing data 2: Solutions	Lecture	FIMD 1.1, 1.2, 1.3, 1.4
7	2024-10-15	Imputation methods	Lab	
7	2024-10-16	Clustering	Lecture	ISLR 12.1, 12.4
7	2024-10-16	Clustering	Lab	
8	2024-10-21	10:00AM assignment 3	Deadline	
8	2024-10-21	Model-based clustering	Lecture	Oberski (2016), optional Hennig (2016), Bouveyron et al. (2019)
8	2024-10-21	Model-based clustering using MClust	Lab	
8	2024-10-22	Text mining 1	Lecture	SLP3 2.1, 2.4, 6.2, 6.3, 6.5, 6.8; TTMR 3
8	2024-10-22	Text mining 1	Lab	
8	2024-10-23	Text mining 2	Lecture	SLP3 2.1, 2.4, 6.2, 6.3, 6.5, 6.8; TTMR 3
8	2024-10-23	Text mining 2	Lab	
	2024-10-23	Inspecting the mid-term exam	Exam Inspection	
9	2024-10-28	Time series	Lecture	OMNG pages 265-294 (forecasting)
9	2024-10-28	Time series	Lab	
9	2024-10-29	Data streams	Lecture	MMDS 4.1 – 4.4
9	2024-10-29	Data streams	Lab	
9	2024-10-30	Algorithmic fairness	Lecture	Mehrabi et al. (2021)

Week	Date	Topic	Type	Reading
9	2024-10-30	Algorithmic fairness	Lab	
10	2024-11-04	10:00AM assignment 4	Deadline	
10	2024-11-05	No lecture: study time	Lecture	
10	2024-11-05	No lab: study time	Lab	
10	2024-11-06	Q&A	Lecture	
10	2024-11-06	No lab: study time	Lab	
10	2024-11-08	Final exam	Exam	
10	2025-01-06	Resit exam	Exam	