TILBURG UNIVERSITY

Data Mining

SYLLABUS — VERSION 04.09.20

Composed by Chris Emmery
Lecturers: Chris Emmery MSc, dr. Gonzalo Nápoles

Contents

Part I: Course Administration

Structure

The course is divided in three components that deal with three separate parts of Data Mining: theory, practice, and application.

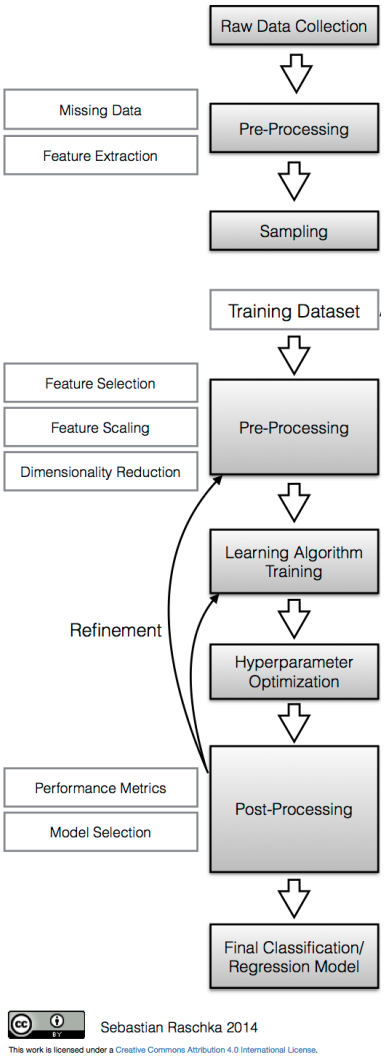
Lectures	all	The lectures serve as an introduction to global concepts discussed within the field of Data Mining. Given that we do not follow the structure of a book, it is **strongly** recommended that you attend the lectures and take notes. Part of the lectures discuss practical sides of the material discussed in the lectures: how do particular algorithms work, how do we interpret them, evaluate them, what are specific cases where they are applied, etc.
Reading Material	theory	The reading material is a mix between further fleshing out the concepts discussed in the lectures, and providing some grounding in the field of Data Mining.
Assignments	application	The assignments deal with the application of Data Mining techniques to actual data sets. Knowing about certain algorithms and best practices is not enough; particular insights and intuitions play an important role — and can only be learned by doing.

Topics & Workflow Relation

The topical structure of this course is best explained by going over a typical workflow as visualized in Figure 1. What it means to process and analyse data, and make predictions based on the patterns in this data is introduced *Introduction to Data Science*. From here on, we deal with three main topics: `data`, `algorithms` and `evaluation`.

Data

Figure 1: A (simplified) overview of the typical Data Mining workflow by Sebastian Raschka (full source).



THINKING about `data` is an involved process, and is therefore spread over several video lectures and practicals. *Introduction to Data Science* sets you off thinking about real world data as vectors of numbers. As working with such a representation is easier explained in context of `algorithms`, this part is put off until the basics of these are explained in Week 2.

First, we take a look at what happens before (`Raw Data Collection` in Figure 1); how data can be structured and accessed in Practical 1 & 2 (*Data*, *Pandas*) and *Representing Data*. Here, Pandas will be the Python tool of choice for data manipulation and visualization. We come back to the mathematical representation of data during Week 2 with a specific example how particular modalities pose additional challenges for such representations (*Working with Text Data I*). Practical 2 involves `Pre-processing` (Figure 1), and we work towards having our data cleaned an prepared for experiments.

After the midterm, we go into more advanced `Pre-processing` topics (see second Pre-processing part in Figure 1), such as informed reduction of the amount of data and feature selection (*Data Reduction & Decomposition*), inherent challenges that language poses (*Working with Text Data II*), and dealing with large collections of data (*Mining Massive Data*).

Algorithms

UNDERSTANDING and working with algorithms requires understanding their (most basic) mathematical workings. We will be covering a varied range of algorithms to make predictions with, or discover patterns in data. We cover the basics of regression (Linear Regression) and classification (Logistic Regression) in (unsurprisingly) *Regression & Classification*. We demonstrate the difference between distance-based algorithms (like k-Nearest Neighbours) and algorithms relying on measures of information (Decision Trees) in *Working with Text Data I* and probabilities (Naive Bayes) in *Working with Text Data II*. After the midterm, we go into more advanced topics such as unsupervised, and semi-supervised algorithms (*Data Reduction & Decomposition*, *Association Rule Mining*). Their workings and application examples are all part of `Learning Algorithm Training` in Figure 1. Tuning their parameters, and understanding how these affect learning is part of `Hyperparameter Optimization`. The latter is covered in *Algorithmic Fitting & Tuning* and *Best Practices, Common Pitfalls*. You will be able to run several algorithms from Practical 3 onward, for which we will

be using Scikit-learn.

Evaluation

MEASURING and critically assessing the performance of your models (the `Sampling` and `Post-Processing` parts in Figure 1) is the most important part of the entire workflow. While the basic tools are quite straight-forward to understand, deeper concepts related to algorithmic predictions such as generalization (error), bias, variance, and generalizability require an understanding of all elements combined. We will be highlighting these concepts in *Algorithmic Fitting & Tuning* and *Best Practices, Common Pitfalls*, but they are generally discussed throughout the entire course.

Schedule

THE SCHEDULE is subject to change. Please always make sure you have the latest version of the syllabus.

wk	Date	Lectures		Practicals
1	01-09	Introduction to Data Science	Representing Data	Pandas
2	08-09	Regression and Classification	Working with Text Data I	Scikit-Learn
3	15-09	Algorithm Fitting & Tuning	Best Practices, Common Pitfalls	DIY
!!	18-09	**Midterm**		
4	22-09	Working with Text Data II		
5	29-09	Data Reduction & Decomposition	Clustering	Cleaning
6	06-10	Association Rule Mining	Mining Massive Data	Research
7	13-10	Deep Learning for Data Mining (guest lecture)		

Whenever possible, slides will be made available on Canvas before the lecture. Video Lectures are provided via Mediasite (links will be posted on Canvas), practicals are included in the syllabus.

Examination

PASSING the course is determined by three examination components:

Assignments	COMPLETE	(Y/N)
Midterm	GRADE	20%
Exam	GRADE	80%

Practical Assignments

Evaluation The assignments are **not** graded, but are **mandatory**. You get a `complete` mark based on handing in an assignment, after it has successfully passed a sanity check[1]. Note that there is no 'right' way to do the assignments; i.e. if you aren't able to figure out a particular part, that does not mean you do not pass the assignment.

[1] Determines if you at least tried the assignment

Deadlines All notebook assignments should be handed in the day *before* the next practical. Consequently, if there's no practical scheduled this means you will have two weeks to finish the work.

Making the Assignments You are welcome to collaborate on the assignments. Do note that they should be handed in individually[2].

[2] To hand in the assignment, please submit the `.ipynb` notebook to Blackboard. See Practical 1 for more detailed instructions

Discussing the Results Every new practical will provide a recap to discuss the take-aways from the previous one. If you feel this is unsatisfactory, feel free to open up a Discussion on the forum.

Midterm

THE MIDTERM will test your theoretical knowledge and practical insights (not skills, so no programming) regarding the material we have discussed at that point, and serves as a) a way to gauge your comprehension of the basic material (first 3 weeks), and b) as preparation for

the question style of the final exam. As such, the way to prepare is to study the associated (video) lectures (and notes), and papers. There are no calculations required for the midterm — this does however **not** imply that we will not ask questions regarding the algorithms or any other metrics. More details in the margin. [3]

Exam

THE EXAM covers all material of the course, **including** math parts; we expect you to understand and be able to apply everything shown in the lectures. Other than that, the examination style is similar to that of the midterm, with a few changes. [4]

During the Exams

THE EXAMINATIONS are conducted via Canvas. Please read all the questions **carefully**. The devil is often in the details. Backtracking (i.e. skipping back to a particular questions) is **not** possible.

A JUSTIFICATION DOCUMENT will have to be prepared for a subset of questions (last 5 for the midterm, last 10 for the exam). It should explain/justify the answers in detail, with 2-3 sentences each. Hence, while answering the questions, you are advised to already take notes into a separate document (remember: **no backtracking**).

After the Exams

THE JUSTIFICATION DOCUMENT can be submitted within limited (20 minutes for the midterm, 30 for the exam) via Canvas. Provide the document (in time!) via Canvas (via the announced module) — **do not forget to include your name and student number**.

GRADES should be expected to be up within two weeks. After, you will be given the opportunity to review the test at allocated timeslots. Sign-up and details for these timeslots will be published on Canvas.

For your final grade, the highest achieved grade between the exam and the resit counts for your average grade. There are no entry requirements for either the exam or the resit. As such, there is no risk involved when skipping one, or attending both.

To pass the entire course, your weighted average needs to be *at least* 5.5 or higher (i.e. not a grade that rounds to a 5.5)[5]. Please note

[3] Midterm configuration:

- 20 % of course grade.
- Pen and paper (no calculator needed).
- Fifteen multiple-choice style questions: four options, only **one** correct.
- Open book (e.g. slides, reading materials, and materials discussed during the practical sessions, are allowed).

[4] Exam configuration:

- 80 % of course grade.
- Pen, paper, and calculator.
- Thirty multiple-choice style questions: four options, only **one** correct.
- **Important**: you can bring ONE A4-sized piece of paper worth of notes; double-sided, in any font size you are still able to read.

[5] Everything between a 5 and 6 is rounded to a whole number. Any grade outside of this range is rounded to halves.

that because of this, you do **not** need a passing grade for the midterm — as long as your average is passing, you pass the course.

Questions & Answers

We realize that the course attracts an incredibly heterogeneous group every block. Adjusting the speed of the lectures to fit everyone's background is sadly an impossible task. We have partly accommodated for this in the video lectures; these you can watch on your own pace If you are still struggling with the material, we offer several ways of helping you out.

Asking Questions

Forum We highly prefer you turn to the **Discussion forum** on Canvas for any questions you have. We monitor it constantly during the course, and are more than willing to go into detail on the material there. The Forum allows us to maintain the FAQ document, and offers a transparent platform where answers to theoretical and practical questions are accessible to everyone.

Q&A Session Questions regarding the material can be asked during the scheduled Q&A Session in the second block of the practical lectures. Details and times will be made available on Canvas.

E-mail If your question is of personal nature, feel free to drop us an e-mail at the addresses above.

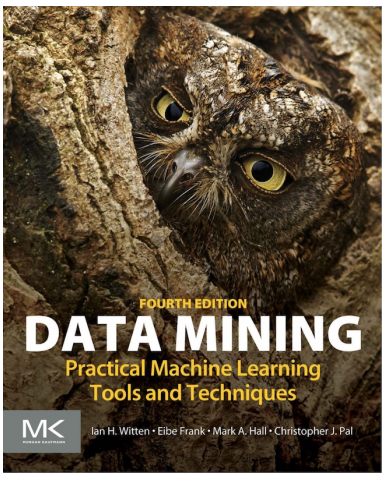
Part II: Lecture Materials

Supplementary Material

While the course is ought to be self-contained, it often helps to have some supplementary material for further reading.

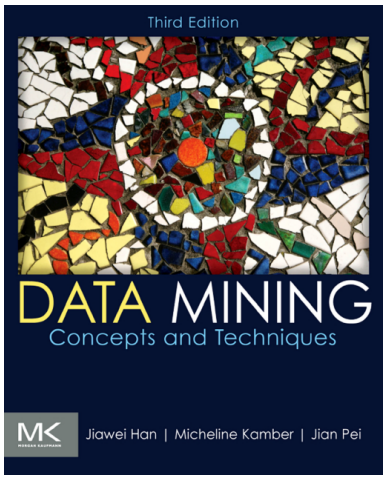
Data Mining Books

There are three books we recommend:



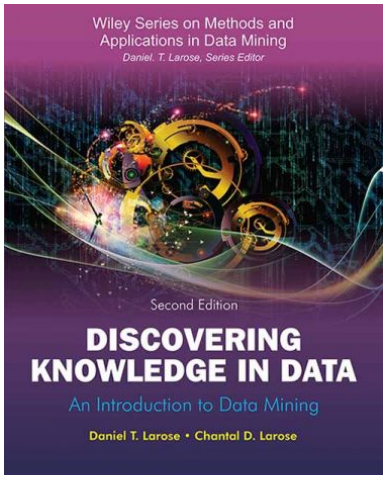
Data Mining: Practical Machine Learning Tools and Techniques

Ian H. Witten, Eibe Frank, Mark A. Hall



Data Mining: Concepts and Techniques

Jiawei Han, Micheline Kamber



Discovering Knowledge in Data: An Introduction to Data Mining

Daniel T. Larose, Chantal D. Larose

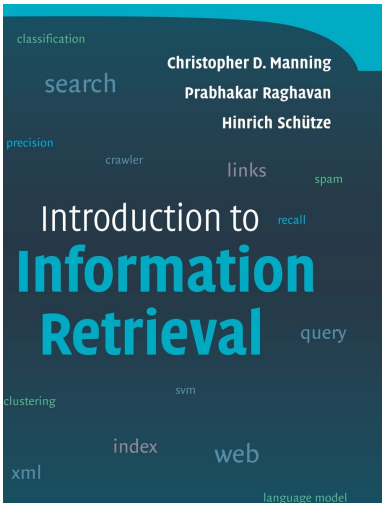
They all cover different ground[6]: Witten et al. is most popular — but assumes you are using WEKA, Han & Kamber has more of a data-oriented take and Larose has a strong focus on use-cases and more detailed explanation of the algorithms. However, all cover quite an extensive amount of material that is **not** covered in this course.

[6] These books officially *not* free. Their `.pdf`'s are known to float somewhere around the Internet, however.

Other Books

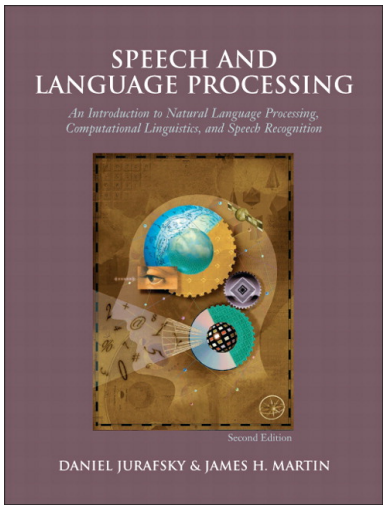
For additional reading to accompany working with text / documents specifically[7], these are **free** and helpful:

[7] Please note that these books deal with Information Retrieval and Natural Language Processing, which could be courses on their own.



Introduction to Information Retrieval

Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze



Speech and Language Processing

Daniel Jurafsky, James H. Martin

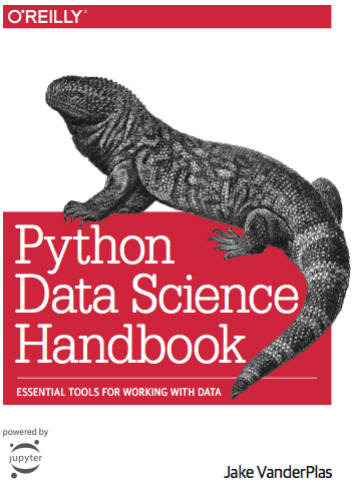
Online Material

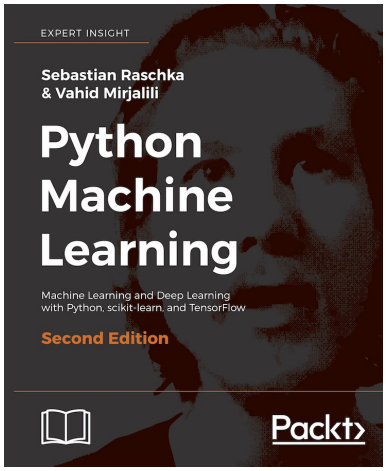
Any other helpful material we'll add during the course below. Please note that some might be full courses or other material that expects some particular background (in e.g. computer science / maths). Get what you can from them. These are links; if any break, try to search for the names.

Andrew Ng's MOOC on Machine Learning	link
MIT's Artifical Intelligence course	link
Brandon Foltz' YouTube channel on stats	link
Jeff Miller's YouTube series on the maths behind Machine Learning	link

Course Follow-up

Some of my (Chris) favourite (open-source) authors:





Practicals: **Python Data Science Handbook** (free!)

Jake Vanderplas

Lectures: **Python Machine Learning**

Sebastian Raschka