UNIVERSITEIT TWENTE.

Data Science [201400174]

Course year 2021/2022, Quarter 1B

DATE
November 14, 2021

EXCERPT
Chapter 0: Introduction

TEACHERS
Faizan Ahmed
Nacir Bouali
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Nicola Strisciuglio
Estefania Talavera
Brenda Voorthuis
Shenghui Wang

COURSE COORDINATOR
Maurice van Keulen (quartile 1A)
Karin Groothuis-Oudshoorn (quartile 1B)
Faizan Ahmed (quartile 2A)

PROJECT OWNERS
Faiza Bukhsh
Karin Groothuis-Oudshoorn
Maurice van Keulen
Elena Mocanu
Mannes Poel
Michel van Putten
Mohsen Jafari Songhori
Luc Wismans

Chapter 0

Introduction

Course in a nutshell:

- You work in pairs.
- You need to have 2 topic assignments signed off. They will be checked and are mandatory, but will not contribute to the final grade.
- You need to finish 1 project. This will be graded.
- All submissions should be made through Canvas.
- You need to present your project in a presentation at the end of the course. Each presentation has a 10 minutes slot.

0.1 Global overview of the course

Data Science is an interdisciplinary field that lies at the intersection of computer science, statistics, visualization and the social sciences. Scientific and economic progress is increasingly powered by our capabilities to explore big data sets. Data scientists dig for value in data by analyzing for instance texts, application usage logs, and sensory data. They are the driving force behind the successful innovation of Internet companies like Google, Twitter, and Yahoo. There is an increasing need for data scientists and big data engineers seen in job advertisements. The need for data scientists and big data analysts is apparent in almost every aspect of our society, including computer science, medicine, physics, and the humanities.

The goal of the course Data Science is to teach several data science skills needed in various phases of data analysis projects. The course concept is geared towards *self study* in an assignment & project-driven manner, i.e., it is designed to offer a rich environment for flexible, effective, and efficient self study with ample guidance and supervision. The course is assessed with a project that takes about half of the course. There are several projects offered from which the student can choose. A project is composed of a real-world data set and a *challenge*, i.e., what knowledge can potentially be extracted from the data or what the project owner wants to do with the data. The data science skills are offered as *technical topics* from which the student has to choose two. The projects indicate which technical topics provide the necessary skills for doing the project, so the choice for project and technical topics should be coherent.

Each topic consists of one lecture and a practicum for learning the basic skills. The practicum and project are done in pairs. Supervision is provided during practicum-sessions twice per week shared with all topics and projects. The project is assessed by the project owner and the topic teachers of both topics. The project grade is the grade for the course. The list of projects and topics will be revised every year.

If your study programme allows it, there is a follow-up course, called "Data Science Additional Topics". It consists of doing the course again, but now with two different topics and a different project. you can do this is different quartiles or in one (in the latter case, you would do 4 topics and 2 projects; the two projects will be graded separately on two separate course codes). It is not possible to do (an extension of) the same project for Additional Topics: it needs to be a different project. It is not possible to follow the course a third time.

The following topics are offered (in the first quartile, only DPV and DM are offered):

Data Preparation and Visualization [DPV]
(Topic teachers: M. van Keulen and Faizan Ahmed)
The skills for Data Preparation and Data Visualization taught are in essence drawn from technologies developed for Business Intelligence. They are, however, also effective for data science. The topic teaches (a) data warehousing techniques for extracting and transforming data (ETL), (b) modeling data for analytic purposes using the multidimensional modeling approach of OLAP, and (c) data visualization techniques.

Data Mining [DM]
(Topic teacher: E. Mocanu and K. Groothuis-Oudshoorn)
Data mining is about discovering patterns in large data sets involving methods from artificial intelligence, machine learning, statistics, and database systems. The topic teaches (a) classification, (b) clustering, and (c) regression.
This topic can be done both using the programming language **R** as well as the programming language **Python**.

Information Extraction Using Natural Language Processing [IENLP]
(Topic teacher: N. Bouali)
Most information is available in a form rather unsuitable for processing by computers, namely natural language text. This topic teaches (a) text mining (analyzing text directly), (b) rule-based techniques for information extraction, and (c) statistical techniques for information extraction and natural language processing. This topic is preferably done in combination with "Data Mining".

Feature extraction from Time Series data [TS]
(Topic teacher: F. Ahmed)
Sensors and other measurements increasingly produce massive amounts of data with space and time dimensions. The analysis of spatio-temporal data has many applications. The topic focuses on key techniques for preparing time series data for analysis, such as peak detection, filtering, Fourier analysis (FFT), dynamic time warping (DTW), and prediction models.

Semi-structured data [SEMI]
(Topic teacher: M. van Keulen)
There exist several data exchange and knowledge representation standards. This topic teaches the most important standards and skills to manipulate data in these standards: (a) XML and its associated standards SQL/XML, XPath and XQuery for publishing and manipulation with both relational as well as XML databases, (b) JSON storage and manipulation in relational databases, and (c) Semantic Web standard RDF with its associated standards SPARQL for remote querying, also known as "Linked Open Data".

Data Integration [DINT]
(Topic teacher: S. Wang and M. van Keulen)
An often-needed activity in Data Science is combining the data from two or more independent data sources. Its purpose is usually data enrichment: Given a data set with information about a certain entity (e.g., patients, users, products, locations, etc.), we would like to add additional information *about these same entities* from a different data source. The topic teaches the most important steps in a data integration pipeline: matching, mapping, merging, and evalution as well as dealing with often occurring complications: no or no reliable IDs, attributes having different names, inconsistencies between the sources, large data sources, etc.

Probabilistic Databases and Data Quality [PDBDQ]
(Topic teacher: M. van Keulen)
Much effort in data preparation is devoted to dealing with data quality problems. A prime example is data integration: there is a risk of picking a wrong match between sources or a wrong value in case of inconsistencies. Probabilistic database technology has the potential of representing data quality problems as uncertainty in the data, and storing and querying it. The topic teaches the most important skills for (a) using probabilistic database technology, and (b) how to represent several kinds of data quality problems as uncertainty in the data.

Process Mining [PM]
(Topic teacher: F. Bukhsh)
Process mining aims to improve understanding and efficiency of business processes by analysing event logs with specialized data-mining algorithms. The topic teaches the most important concepts and skills for applying and understanding Process Mining: (a) petri nets: the theoretical foundation of process mining, (b) concepts like event log, causal trace, and the Alpha algorithm, (c) using the ProM tool for process discovery, (d) answering analytical questions for a discovered process, and (e) using the ProM tool for process conformance checking.

In order to cover for deficiency in programming the following optional additional topic is offered. Note that the following topic is meant as an extra help for the students.

Typic 0 - Introduction to R/Python
(Topic teachers: N. Bouali and Karin Groothuis-Oudshoorn)
This topic introduces basic programming concepts in both R and Python. Students can follow this topic in *addition to the two topics* they registered for in the course. The main contents of this topic are with respect to both languages R and Python: (a) The Programming Environment (b) Variables and Control Flow (c) Basic Data Structures (d) Functions (e) Libraries for Data Processing, Visualization and Manipulation.
Note that for the last part, the libraries covered depend on the topic(s) you're registered for.

0.2 What is optional and obligatory?

Presence at sessions There are three types of sessions:

- Lectures: introductory lecture and topic lectures happen in week 1; halfway the quartile a few additional lectures will be organized.
- Practical sessions: In weeks 2–9, there are two practical sessions per week.
- Presentation sessions: In week 10, there are presentation sessions. You only need to attend one, namely the one you are presenting in.

Of all these, the only obligatory presence is your presentation session. The lectures and the practicum sessions are optional; they are offered to help you studying for the topics and to support you in carrying out the project. Note that you are also welcome to attend other presentation sessions, not only the one in which you present.

Deliverables The topics have assignments that you need to complete. Gather all you answers in one PDF and submit it to Canvas. Handwritten solutions are acceptable: just scan them or take a picture and include them in your PDF.

The topic assignments need to be signed off by a teacher or one of the assistants. A signed off assignment is visible under "Grades" on Canvas. You may receive feedback on Canvas. You need not wait for the topic assignments to be signed off before continuing. The requirement is that both chosen topics are signed off before the end of the course.

Optional Some aspects/topics may be especially interesting for you. Or there may be a specific 'side path' you are interested in not necessarily part of the topic or project but within the realm of the "Data

Science" theme. Feel free to approach the main teacher about this to discuss how you may deepen your knowledge and skills within the course.

0.3 Grading, resits, repairs

0.3.1 Grading

The **grade for the course is the grade for the project**. The grade for a project is based on the presentation, the report, and other deliverables specified in the course guide for the project. The project will be assessed by two teachers or the project owner of the project and a teacher.

The deliverables for the project are a report and a presentation. To allow for an efficient evaluation of the report, we require that you follow a predefined structure for the report (a template is available on Canvas).

A passing grade will only be given if the requirements below are fulfilled.

1. Both topic assignments should be sufficient and signed off by one of the teachers or assistants.
2. All specified deliverables for both topics and project have been submitted to Canvas.

Note: We do not assess assignments and/or project reports that are submitted by e-mail or through any channel other than Canvas. This is necessary to ensure proper archiving as well as proper tracking of your progress.

0.3.2 Repairs

Since there is no exam, there is also no resit. Nevertheless, in case you fail for the project, there are possibilities to repair it. The following rules apply:

- You are only offered the opportunity to repair a project, if the topic assignments of the associated topics are sufficient and signed off. The reason behind this rule is that you need to have shown to have done a serious attempt before we are willing to invest in offering you the opportunity for a repair.
- **Repair of a project by means of a new attempt (full repair)**. Analogous to a resit, a repair of a project grade consists of doing another (different) project chosen from the ones offered in the course.
- **Repair of failing grade to obtain a 6 (direct repair).** In case of a repair of a failing grade (i.e., less than 5.5), you can request specific repair assignment(s) (the "direct repair"). The repair assignments focus on those aspects that were too weak in the original attempt. Note that this repair option for failing grades does not require (re-)doing a full project. This repair option allows for a maximal score of 6.

  The general procedure for a direct repair is the following:
  1. The student requests a direct repair from the main teacher within 3 months after reception of a failing grade.
  2. The main teacher assigns a teacher for the repair.
  3. The student submits a repair plan to this teacher. The repair plan should list the steps to be taken in order to address the weak points of the project. A good repair plan lists the feedback point from the assessor and the way to improve this point.
  4. The repair plan is assessed and approved by the teacher.
  5. The student implements the items from the repair plan, and hands in the repair to the teacher, who notifies the main teacher.
  6. Note, that this is not an iterative procedure. Direct repairs of a failing grade are analogues to resits. So the repair is performed by the student and at the end assessed once (with the added benefit, that the repair plan is approved and the student knows that if the plan is properly implemented, the repair will be successful).
- How the report for the repair is submitted is determined by the main teacher (usually per e-mail to the teacher). The teacher may also decide on another presentation to be given.

In case you simply fail to complete the topics and/or project during the quartile you started, it is no problem to continue and present during a subsequent quartile. This is possible in all quartiles with the following considerations: continuing when the course is given in the 2nd and 3rd quartile has no limitations; in the 1st quartile the course is given in a limited manner; in the fourth quartile, there is no supervision but arrangements for a presentation can be made. Please inform the main teacher if you will not finish your topic assignments, do not intend to present your project, or other circumstances that need continuation later. Signed off topics remain valid for a maximum of one year.

0.4 Supervision and guidance

Immediate supervision and guidance with the topics and project is limited to practical sessions. The teaching assistants will be present to assist yu with questions and problems on any topic or project. We strive for having a teacher present at all practical sessions as well.

Outside the practical sessions, you cannot expect immediate guidance or help. We do, however, use a system called Discord (see Canvas), that remains available also outside scheduled hours. The intention is that this environment is used for (1) students helping students with problems, (2) teaching assistants and teachers helping students with problems, (3) discussions about the theory, practical sessions, project, and the technology in general.

Discord can be used from within your web browser as well as with native applications on mobile phones, tablets, and laptops. It can be installed from the appropriate app store. We have our own Discord server for Data Science running (see Canvas for details).

0.5 Project report template

On Canvas you find a template for your project report. The template is based on a template from the ACM (Association for Computing Machinery) which is widely used in scientific publishing. The template contains formatting help and instructions and is available for Microsoft Word and LaTeX.

0.5.1 Structure

Every project is different, but we recommend the following structure for your report:

- *Abstract:* A short overview of your paper.
- *Introduction:* What is the problem setting of the project (and why is this a problem at all)? What is your concrete (research) question you want to solve? What is your overall idea for solving it (briefly). What comes in the remainder of the paper?
- *Background and/or Related Work:* This section contains the required background knowledge for reading the report. For instance, if you use a specific type of feature selection or a (not so standard) visualization it should be explained here. A related work part is necessary, if other authors have done the same (maybe solved the same problem on the same data set with a different approach). Depending on your topic, the section could either be ”Background” (very practical topic), ”Background and Related Work” (practical topic with additional scientific references) or ”Related Work” (very scientific topic). Depending on the topic, the section can also be split into ”Background” and ”Related Work”.
- *Approach:* Here you describe how you solved the problem. A good ”approach” section is not a historical review on what you’ve tried step by step (and where you failed) – so it is not a experience report. But rather you describe the approach that finally worked. Things that failed (and why) go into the discussion section. If you tried multiple paths and all of them are reasonable and lead to different results (with advantages and disadvantages), those should be described (and discussed later). This section might of course contain subsections.

- *Experiments:* This section (empirically) proves what your approach claims. The experiment section contains all the information necessary to judge to which extent the problem you stated in the introduction is solved by the steps you proposed in the approach section. Usual subsections contain a data set description, the general experimental setup and the results.
- *Discussion:* Here you interpret the results you achieved in the experiments, and discuss it in relation to other works (if applicable). It should also contain the limitations (what did not work and an hypothesis why) and observations you made during the experiment.
- *Summary: or Conclusions* This section summarizes the work. In a way it is a different view on the things you wrote in the introduction. Which problem was attacked, was it solved? What would be next steps and/or application areas. Do not repeat the details, but summarize the work.

The page limit for the report is **5 pages**. If you have many additional visualizations for instance, you can add them in the appendix, which does not count towards the page limit. However, everything that is core to your arguments and your results should be in the main part of the document. In other words, the appendix can contain supporting material, but should not contain core material. The reason for the limit is that i) you have a guideline of how long such a project report should be and ii) to teach you to produce concise and yet comprehensive descriptions of a larger project work.