Data modeling foundations

Offered by: Department of Mathematics and Computer Science
Language: English
Primarily interesting for: Data Science (BSc) and other BSc students interested in the MSc Data Science in Engineering
Prerequisites: 2IT60 Logic and set theory (or equivalent)
2WBB0 Calculus B (or equivalent)
Contact person: dr. George Fletcher

What is this package about?
This elective package is for students in the bachelor program Data Science and other programs (except the bachelor program Computer Science), who are interested in the master program Data Science in Engineering (DSiE). The courses in this package are required prior knowledge for enrollment in this master program. Students are only admissible in the master program DSiE if they have successfully completed both the packages (1) Data modeling foundations and (2) Algorithms.

<table>
<thead>
<tr>
<th>Courses name</th>
<th>Scheduled</th>
<th>Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Data management for data analytics (JBI050)</td>
<td>2(A)</td>
<td>2</td>
</tr>
<tr>
<td>2. Linear algebra and applications (2DBI00)</td>
<td>4(E)</td>
<td>1</td>
</tr>
<tr>
<td>3. Probability &amp; statistics (2DI90)*</td>
<td>4(C)</td>
<td>2</td>
</tr>
</tbody>
</table>

* Data Science BSc students are not required to take 2DI90 and may replace it with any other allowed elective.

Course descriptions

Data management for data analytics
As we enter an era of big data and data science, core knowledge and skills in data modeling and data management are now recognized as essential in many disciplines. The primary goal of JBI050 is to master the core best-practices of data management systems, applied towards using contemporary tools to support effective data analytics. In particular, this course focuses on preparing students to meet contemporary data modeling and data management challenges which arise in applications in their own fields of study.

Linear algebra and applications
This course offers a wide range of interesting linear algebra techniques and very nice applications, including Big Data applications. These techniques are almost indispensable for the current Big Data era. You will learn how to solve linear systems, arguably the most common and important scientific problem. Least squares methods can be used to determine an approximate line (or polynomial, or spline) through a number of points (used in Computer Graphics). Rotations and reflections are useful for Computer Graphics and Robotics. Angles
between vectors can be used to compare tastes of movies, music, or books, and to predict such tastes for the future. You will learn the idea behind Google PageRank and its connection with eigenvalues, how it is computed, and you can also apply similar techniques to finally determine which of Ajax, Feyenoord, or PSV is the best team. We can use text mining methods to determine the most important tweets or keywords from a set of tweets; useful for effective communication. The course is designed to be interesting for many students, including students from Computer Science, Data Science, and Robotics. The subjects are fascinating on their own, but also form an ideal preparation for subsequent courses such as Data Mining and Machine Learning and Computer Graphics. The course offers a nice combination of theoretical (math) and practical (implementing) work, partly done in groups. You can use a language you already know (for instance Python or Matlab) or learn a new language.

Probability & statistics

Part I - Probability Theory: In this part of the course you will learn the foundations of probability theory. In many situations, and in particular in the study and design of hardware and software systems, it is extremely difficult to predict exactly how a certain system will behave. It is therefore important to be able to model such uncertainty in a sensible way, so that one can assess the system’s performance in an effective way. After successfully completing this part of the course, you should be able to perform elementary probability calculations with stochastic models. Moreover, you should be able to make use of probability distributions for modeling and analysis of situations where randomness occurs (or when it is considered a good modeling tool). You should be able to adequately document calculations that form the bases of such analysis. As part of the course you will learn various integral calculus concepts, such as integration by parts and the substitution method. Finally, you will learn some basic principles behind Monte-Carlo methods, and perform simple computations with the software package R.

Part II – Statistics: This part of the course builds upon what you learned in part I. Using the tools of probability theory we will devise statistical models that will enable you to analyze and describe data, and make meaningful statements about it. After this part of the course, students should be able to perform elementary statistical analyses of data. They should know how to use statistical software in a proper way to obtain both qualitative and quantitative data analysis results. They will learn how to use and construct point estimators and confidence/prediction intervals and how to apply tests of hypotheses one and two samples. Furthermore they should be able to recognize simple regression problems, and do simple statistical analysis for such models, using the software package R.