Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie

School of Information Technology
Skool vir Inligtingtegnologie

Department of Computer Science
Departement Rekenaarwetenskap

COS 781: Data Mining
COS 781: Data-Ontginning

Lecturers: Mr. W. S. van Heerden, Mr. M. Riekert
Last Revision: 26 July 2015

©Copyright reserved
# Contents

1 Overview 1  
1.1 Description .................................................. 1  
1.2 Prior Knowledge .............................................. 1  
1.3 Related modules .............................................. 1  
1.4 Study units .................................................. 1  
1.5 Specific Topics ............................................... 2  

2 Outcomes 2  
2.1 Career ..................................................... 2  
2.2 Course ..................................................... 3  

3 Plagiarism policy 4  

4 Instructors 4  
4.1 Course coordinator ......................................... 4  
4.2 Lecturers ................................................... 4  
4.3 Administrative support ..................................... 4  

5 Organisation 5  
5.1 Class attendance ............................................. 5  
5.2 Module website ............................................... 5  
5.3 Mark Administration ......................................... 5  
5.4 Announcements ............................................... 5  
5.5 Lectures ..................................................... 6  
5.6 Interaction with the Instructors ............................. 6  

6 Study Material 6  
6.1 Prescribed textbooks ....................................... 6  
6.2 Additional references ....................................... 6  
6.3 Software ................................................... 7  

7 Assessment 7  
7.1 Semester mark ............................................... 7  
7.2 Assignments ................................................ 7  
7.3 Presentation ................................................ 7  
7.4 Examination .................................................. 8  
7.5 Final Mark Calculation ...................................... 8  

8 Lecture Schedule 8
1 Overview

1.1 Description
Data mining is the automated extraction of knowledge, or hidden patterns from very large data bases. The focus of this course is on how the computational intelligence techniques covered in the COS 710 and COS 711 courses (i.e. evolutionary algorithms, swarm intelligence and neural networks) can be used for knowledge extraction. In addition, traditional machine learning techniques such as decision trees and rule induction will be covered. The pre-processing of data, in preparation for data mining algorithms, will be treated, as well as ways in which very large databases can be processed. Exploratory data analysis is also investigated. Finally, some attention will be given to the extraction of hidden knowledge from unstructured data, such as text and images.

1.2 Prior Knowledge
There are no official prerequisites, beyond admittance to the Honours degree program. Some prior knowledge of the various CI approaches, which are covered in COS 710 and COS 711, is advisable. In particular, a working knowledge of evolutionary computation, swarm intelligence and neural networks is recommended. Students who apply themselves to the work will not have a problem, even if they have not completed COS 710 or COS 711. Some of the assignments will possibly also require programming, which students may implement in any programming language on any platform of their choosing.

1.3 Related modules
The following modules touch on topics that are covered in COS 781:

<table>
<thead>
<tr>
<th>Code</th>
<th>Description</th>
<th>Related Units</th>
</tr>
</thead>
<tbody>
<tr>
<td>COS 710</td>
<td>Artificial Intelligence I</td>
<td>Advanced algorithmic analysis</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advanced search</td>
</tr>
<tr>
<td>COS 711</td>
<td>Artificial Intelligence II</td>
<td>Machine learning and neural networks</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Advanced knowledge representation and reasoning</td>
</tr>
</tbody>
</table>

1.4 Study units
This module covers the main themes of data mining through the following study units, taken from the ABET Computing Accreditation Commission guidelines:

<table>
<thead>
<tr>
<th>Knowledge Area</th>
<th>Study Unit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Algorithms and Complexity</td>
<td>Advanced algorithmic analysis</td>
</tr>
<tr>
<td>Information Management</td>
<td>data mining</td>
</tr>
<tr>
<td>Intelligent Systems</td>
<td>Machine learning and neural networks</td>
</tr>
<tr>
<td></td>
<td>Advanced knowledge representation and reasoning</td>
</tr>
<tr>
<td></td>
<td>Advanced search</td>
</tr>
</tbody>
</table>
1.5 Specific Topics

This course covers the following specific topics:

- Introductory data analysis concepts
- Exploratory data analysis
- Data preparation
- Data clustering
- Rule induction and decision tree building
- Evolutionary computation for data mining
- Neural network theory and application for data mining
- Unstructured data mining (text, image and audio)
- Mining big data

2 Outcomes

2.1 Career

The aim of the course is to impart a broad knowledge of data mining to students, with the focus on the application of computational intelligence paradigms to data mining. This will equip students for any ICT career in which more advanced data analysis and visualisation is required.
2.2 Course

The outcomes of the study units covered by this course, taken from the ABET Computing Accreditation Commission guidelines, are as follows:

**Advanced knowledge representation and reasoning**
- Compare and contrast the most common models used for structured knowledge representation, highlighting their strengths and weaknesses.

**Data mining**
- Explain the role of finding associations in commercial market basket data.
- Compare and contrast different conceptions of data mining as evidenced in both research and application.
- Characterize the kinds of patterns that can be discovered by association rule mining.
- Describe how to extend a relational system to find patterns using association rules.
- Evaluate methodological issues underlying the effective application of data mining.
- Identify and characterize sources of noise, redundancy, and outliers in presented data.
- Identify mechanisms (on-line aggregation, anytime behavior, interactive visualization) to close the loop in the data mining process.
- Describe how the various close-the-loop processes improve the effectiveness of data mining.

**Advanced algorithmic analysis**
- Explain the use of randomization in the design of a data mining algorithm for a problem where a deterministic algorithm is unknown or much more difficult.

**Machine learning and neural networks**
- Determine which of two learning styles (supervised learning, and unsupervised learning) is appropriate within a data mining domain.
- Compare and contrast each of the following techniques, providing examples of when each strategy is superior for data mining: decision trees, and neural networks.
- Implement a simple learning system for data mining, using decision trees, and neural networks, as appropriate.

**Advanced search**
- Explain what genetic algorithms are and contrast their effectiveness with the classic data mining techniques.
3 Plagiarism policy

The Department of Computer Science considers plagiarism to be a serious offence. Disciplinary action will be taken against students who commit plagiarism. Plagiarism includes copying someone else’s work without consent, copying a friend’s work (even with consent) and copying material (such as text or program code) from the Internet. Copying will not be tolerated in this module.

For a formal definition of plagiarism, the student is referred to http://www.ais.up.ac.za/plagiarism/index.htm (from the main page of the University of Pretoria site, follow the Library quick link, and then click the Plagiarism link). If you have any form of question regarding this, please the lecturer, to avoid any misunderstanding. Also note that the principle of code re-use does not mean that you should copy and adapt code to suit your solution.

Note that all assignments submitted for this module implicitly agree to this plagiarism policy, and declare that the submitted work is the student’s own work. Assignments will be submitted to a variety of plagiarism checks. Any typed assignment may be checked using the Turnitin system. After plagiarism checking, assignments will not be permanently stored on the Turnitin database.

4 Instructors

4.1 Course coordinator

The module coordinator takes primary responsibility for the module and its content:

<table>
<thead>
<tr>
<th>Name</th>
<th>Office</th>
<th>Telephone</th>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr Willem van Heerden</td>
<td>IT 5-38</td>
<td>(012) 420-5468</td>
<td><a href="mailto:wvheerden@cs.up.ac.za">wvheerden@cs.up.ac.za</a></td>
</tr>
</tbody>
</table>

4.2 Lecturers

There are two (2) lecturers for this course, who will present lectures and assess evaluation opportunities:

<table>
<thead>
<tr>
<th>Name</th>
<th>Office</th>
<th>Telephone</th>
<th>e-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mr Willem van Heerden</td>
<td>IT 5-38</td>
<td>(012) 420-5468</td>
<td><a href="mailto:wvheerden@cs.up.ac.za">wvheerden@cs.up.ac.za</a></td>
</tr>
<tr>
<td>Mr Marius Riekert</td>
<td>IT 5-43</td>
<td>(012) 420 3561</td>
<td><a href="mailto:mriekert@cs.up.ac.za">mriekert@cs.up.ac.za</a></td>
</tr>
</tbody>
</table>

4.3 Administrative support

General degree and examination-related administrative queries should be directed to:

<table>
<thead>
<tr>
<th>Name</th>
<th>Office</th>
<th>Telephone</th>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms Elmarie Willemse</td>
<td>IT 4-18</td>
<td>(012) 420-2504</td>
<td><a href="mailto:ewillemse@cs.up.ac.za">ewillemse@cs.up.ac.za</a></td>
</tr>
</tbody>
</table>

The head of the Computer Science Department must be contacted via his secretary:

<table>
<thead>
<tr>
<th>Name</th>
<th>Office</th>
<th>Telephone</th>
<th>E-mail</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ms Angela Bekker</td>
<td>IT 4-17</td>
<td>(012) 420-2361</td>
<td><a href="mailto:abekker@cs.up.ac.za">abekker@cs.up.ac.za</a></td>
</tr>
</tbody>
</table>
5 Organisation

5.1 Class attendance

First and foremost, class attendance is vital to maintain a good academic record. Material not covered in the slides or textbook may also be discussed during lectures. Please ensure that you attend these forums so that you are aware of important announcements, additional discussions and material not covered in this study guide.

5.2 Module website

The CS module page for COS 333 is maintained on the web page of the Department of Computer Science. You can reach this page by visiting the department’s website at http://www.cs.up.ac.za/, clicking the Undergraduate Portal link, and finally clicking the COS333 link in the list. You can also reach the CS module page directly, by visiting http://www.cs.up.ac.za/courses/COS333/. The CS module page will be used to host most of the study material for this module, and is thus the primary point of contact for electronic information and resources. The CS module page will host the following content:

- All announcements.
- File downloads, links and notes.
- All practical specifications.
- Practical submission upload slots.
- The discussion forum.

You are advised to regularly monitor the COS 333 website for any updates and new announcements. Please note that this is the 26 July 2015 version of the course study guide, and refer to the website for the latest version.

5.3 Mark Administration

Marks for assessment opportunities will be posted electronically as soon as they are available. A mark query procedure will be published on the module site when marks are published. Follow these instructions carefully. Failure to do so may result in your query not being attended to. Mark queries must be sent within five (5) working days of their publication, and will be attended to as soon as possible after they have been received. Retain physical proof of all assessment opportunity marks (i.e., test scripts, and marked practicals) until after the module’s examination.

5.4 Announcements

Announcements will be made during class time, and via the module website. While every effort will be made to repeat announcements both in class and via the website, it is not guaranteed that this will be the case. It is therefore expected that students attend all classes, and check the module website on a daily basis.
5.5 Lectures

There is one (1) lecture per week for COS 781, as follows:

<table>
<thead>
<tr>
<th>Day</th>
<th>Time</th>
<th>Venue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wednesday</td>
<td>18:30–20:30</td>
<td>IT 4-3</td>
</tr>
</tbody>
</table>

The venue may be changed should the group be of such a size as to allow everyone to fit into a smaller venue. An announcement will be published if a new venue is chosen.

5.6 Interaction with the Instructors

You can speak with the lecturers after classes. If you require a longer appointment, please send an e-mail to request one. When communicating via email, please be courteous and use clear language during all interactions. Always make appointments via e-mail, and do so well in advance (not the day before an important deadline). It is not advisable to visit the lecturers without an appointment, since they have a busy schedule and may be unavailable. Ask questions to ensure your understanding of the work is up-to-date.

Please also note that lecturers will be unavailable during the weeks that they do not present a class. These arrangements are due to the fact that these weeks are reserved for the lecturers’ research work, on the instruction of the Head of Department. On these days, they will not be available for meetings and it is very likely that they will not respond to e-mails. Therefore, please always contact the lecturer presenting classes at the time of your query.

6 Study Material

You are advised to regularly monitor the COS 781 website (at http://www.cs.up.ac.za/courses/COS781) for any updates and new announcements. Please note that this is the 26 July 2015 version of the course study guide, and refer to the website for the latest version. Class attendance is vital to maintain a good academic record. Please ensure that you attend these forums so that you are aware of important announcements, additional discussions and material not covered in this study guide.

6.1 Prescribed textbooks

There is no prescribed textbook for this course. Summarising slides may be provided for certain of the lecture themes. In the absence of slides, it is the students’ responsibility to take adequate notes during the lecture.

6.2 Additional references

During the course, the lecturers may provide a number of articles. In some cases, the article itself may be provided. In other cases, references may be given, and it is the students’ responsibility to ensure that they obtain a copy. Guest lectures may also be presented during class times, and all material related to these lectures will also be examinable, unless stated otherwise.
6.3 Software

Links to certain data mining software packages may be provided during the course of the semester. It is the student's responsibility to decide which software to use for assignments, and obtain copies of such software if references have not been provided. It may be necessary for students to write and use their own data mining software, should adequate software not be available.

7 Assessment

The semester mark consists of the following assessment opportunities: Four (4) assignments, and a presentation. Each opportunity is discussed separately, below:

7.1 Semester mark

The semester mark for the module will be calculated as follows:

\[
\text{Semester Mark} = \text{Assignments (85%)} + \text{Presentation (15%)}
\]

7.2 Assignments

Four (4) assignments will be given throughout the semester. These assignments will count equally towards an 85% contribution towards the semester mark. Plagiarism is not acceptable, and will result in disciplinary action. Assignments may be analysed for plagiarism via the Turnitin service. In such an event, the assignments will not be permanently uploaded to the Turnitin database. If you have any questions about the plagiarism checking we will use, please ask one of the lecturers.

Two of the assignments will take the form of written assignments on the theory behind one or more particular aspect of data mining. The questions may deal with work covered in class, or related topics that were not explicitly discussed. In the latter case, further research based on scientific publications may be necessary.

The remaining two assignments will require the submission of a written research report, summarising the findings of a data mining exercise. Note that these assignment must be written in a scientific manner. Language must be clear and scientific. Adequate background information must be provided with the assumption that the reader does not know anything about the research field of data mining (but does have a basic background in computer science). Properly cited references must be given. All reported results must be experimentally and statistically verified.

In all cases, make sure that you follow the submission instructions published with each assignment to the letter. Failure to do so (for example, submitting files in the incorrect format, or failing to submit an assignment on time) will result in a zero mark.

7.3 Presentation

Towards the end of the semester, students will have to deliver a presentation on a relevant and interesting topic within the field of data mining. Presentations will be individual, or group-based (depending on the size of the class — specific information will be released closer to the date of the presentations). Further details, related to the selection of topics, will be made
available closer to the presentation date. This presentation will count as a 15% contribution towards the semester mark.

7.4 Examination
The examination will be a three (3) day take-home written paper. The examination must be completed individually, and submitted before the stated deadline. Plagiarism is a serious offence, and will result in disciplinary action. All the work covered during the semester will be examinable, as well as additional material provided through notes and guest lectures. Furthermore, some additional research will also be required to answer certain material. Further examination details will be released at a later date, when they become available.

7.5 Final Mark Calculation
Each student's final mark for COS 781 will be calculated as follows:

\[ \text{Final Mark} = \text{Examination mark (50%)} + \text{Semester Mark (50%)} \]

In addition, the following subject-related regulations hold for this course:

- All students automatically receive examination entrance.
- To pass the course, a student must obtain a final mark of at least 50%.
- A student will pass the course with distinction if he/she obtains at least 75% for the final mark.

8 Lecture Schedule
The following lecture schedule is subject to change, and depends on demand. Any changes to the timetable will be discussed with the students registered for this course, and will be announced on the course website:

<table>
<thead>
<tr>
<th>Week</th>
<th>Date</th>
<th>Topic</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Wed 22 Jul</td>
<td>Introduction</td>
</tr>
<tr>
<td>2</td>
<td>Wed 29 Jul</td>
<td>Lecture 1</td>
</tr>
<tr>
<td>3</td>
<td>Wed 5 Aug</td>
<td>Lecture 2</td>
</tr>
<tr>
<td>4</td>
<td>Wed 12 Aug</td>
<td>Lecture 3</td>
</tr>
<tr>
<td>5</td>
<td>Wed 19 Aug</td>
<td>Lecture 4</td>
</tr>
<tr>
<td>6</td>
<td>Wed 26 Aug</td>
<td>Lecture 5</td>
</tr>
<tr>
<td>7</td>
<td>Wed 2 Sep</td>
<td>Lecture 6</td>
</tr>
<tr>
<td>8</td>
<td>Wed 9 Sep</td>
<td>Lecture 7</td>
</tr>
<tr>
<td>9</td>
<td>Wed 16 Sep</td>
<td>Lecture 8</td>
</tr>
<tr>
<td>10</td>
<td>Wed 23 Sep</td>
<td>No lecture (Spring Day)</td>
</tr>
<tr>
<td>11</td>
<td>Wed 30 Sep</td>
<td>Lecture 9</td>
</tr>
<tr>
<td>12</td>
<td>Wed 7 Oct</td>
<td>No lecture (Recess)</td>
</tr>
<tr>
<td>13</td>
<td>Wed 14 Oct</td>
<td>Lecture 10</td>
</tr>
<tr>
<td>14</td>
<td>Wed 21 Oct</td>
<td>Lecture 11</td>
</tr>
<tr>
<td>15</td>
<td>Wed 28 Oct</td>
<td>Lecture 12</td>
</tr>
<tr>
<td>16</td>
<td>Wed 4 Nov</td>
<td>Presentations</td>
</tr>
</tbody>
</table>