Data Mining

(*in Knowledge Discovery*)

Contents

1. Course Coordinator

Evgueni Smirnov (coordinator)
DKE, UM
SSK 39,
Tel. 043 388 2023
e-mail: smirnov@maastrichtuniversity.nl

2. Introduction

Data mining is a relatively new scientific field that enables finding interesting knowledge (patterns, models and relationships) in very large databases. It is the most essential part of the knowledge-discovery process and has the potential to predict events or to analyse them in retrospect. Data mining has elements of artificial intelligence, machine learning, and statistics

3. Why Data Mining

A typical database contains data, information or even knowledge if the appropriate queries are submitted and answered. The situation changes if you have to analyse large databases with many variables. Elementary database queries and standard statistical analysis are not sufficient to answer your information need. Your intuition guides you to understand that the database contains more knowledge on a specific topic that you would like to know explicitly. Data mining can assist you in discovering this knowledge. The course shows you within two mounts how this works. You will learn new techniques, new methods, and tools of data mining. Hands-on education is involved.

4. Course Description

The course focuses on techniques with a direct practical use. A step-by-step introduction to a powerful (freeware) data-mining tool will enable you to achieve specific skills, autonomy and hands-on experience. A number of real data sets will be analysed and discussed. In the end of the course you will have your own ability to apply data-mining techniques for research purposes and business purposes.

5. Aim of the Course

The primary aim of the course is to provide an introduction to the fundamental concepts found throughout the field of data mining and machine learning.

6. Course Content

The main topics covered in the course are:

- Knowledge Discovery Process
- Stages in Knowledge Discovery:
    - Data Preparation
    - Data Mining:
        - Decision-Tree Induction
        - Rule Induction
        - Instance-Based Learning
        - Bayesian Learning
        - Ensemble Techniques
        - Clustering
        - Association Rules
    - Interpretation and Evaluation of Data-Mining Results
- Tools for Data Mining

7. Required Background

The course does not require any background in artificial intelligence, machine learning, or statistics. A general background in science is sufficient as is a high degree of enthusiasm for new scientific approaches.

8. Course Material

The main textbook for the course is Mitchell(1997). In addition, supporting literature (online courses, hyperlinks, applets) can be found at the *Data-Mining* site that is accessible through Blackboard (https://eleum.unimaas.nl).

T. M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

9. Software Environments

The main software environment used in the course is the Weka data-mining environment. You can download Weka, and its user manual from:

www.cs.waikato.ac.nz/ml/weka/

10. Course Schedule

Week 1

Lectures: *Course Opening*
Knowledge Discovery Process and Data Mining
Induction, Concept Learning, and Version Spaces

Labs: *Induction with Version Spaces*
Induction with Version Spaces
(applet lab)

Week 2

Lectures: *Decision-Tree Induction*
Decision-Rule Induction

Labs: *Introduction to the Weka Data-Mining Environment*
Induction of Decision Trees and Decision Rules in Weka
(simple datasets)

Week 3

Lectures: *Evaluation of Inductive Models*
Data Preparation

Labs: *Induction of Decision Trees and Rules on UCI datasets*
(UCI datasets: training error, test data error, k-fold error, filtering)

Week 4

Lectures: *Instance-based Induction*
Bayesian Learning

Labs: *A Real Data-Mining Challenge: Analysis of the Caravan Dataset*

Week 5

Lectures: Clustering

Labs: *Clustering Simple Data*

Week 6

Lectures: *Association Rules*

Labs: *Basket Data Analysis*

Week 7

Lectures: *Ensemble Techniques*

Labs: *Ensembles of Decision Trees and Naïve Bayes Classifiers*

Week 8

Exam

11. Course Composition

The course consists of three components: lectures, laboratory sessions, and a final exam. The lectures will present the main theoretical and practical aspects of data mining. Each lecture is followed by a laboratory session. During the laboratory session students will be given tasks that involve a data-mining problem. The solutions of the tasks should be submitted individually after at most one week. The course ends with a final open-book exam.

12. Evaluation

Students are evaluated according to their performances on:

- **Laboratory Tasks** (individual). There will be 7 laboratory tasks. The maximum number of points if all the tasks are solved successfully is 10.
- **Final Exam** (individual). The maximum number of points for the exam is 10.

A student's overall score is based on the evaluations of the lab tasks, homework assignments, and final exam, in the following way:

Overall score = 0.4 × score on lab tasks + 0.6 × score on final exam

Please note that the students, who have missed the group meetings but not more than 30% of them, will be given a provisional overall grade. They will receive credits for the course only when they have successfully completed an additional assignment.