UNIVERSITÄT
HEIDELBERG
ZUKUNFT
SEIT 1386

Knowledge Discovery in Databases

Erich Schubert, Michael Gertz

Heidelberg University
Institute of Computer Science
Database Systems Research Group

Winter Semester 2017/18

Part

Organisational Issues

Who, When, Where

- Instructor: Dr. Erich Schubert
- Audience: Students with Computer Science in major or minor course of study; students from CS related disciplines
- Time and Place: Mondays 2-4pm, Thursdays 9-11am, INF 205, HS
- Course material and further information in Moodle
  `https://elearning2.uni-heidelberg.de/course/view.php?id=16572`
  Enrollment key: kDd-1817
- Exercises:
  - Time and Place: Thursdays 2-4pm, INF 327 SR 1
    First exercise on Thursday, October 22
  - Discussion of current and past assignments, practical exercises using software

Prerequisites and Exam

- Lectures: Algorithms and Data Structures
- Basic skills: Statistics and probability theory (stochastics)
- Participation
  - Lectures
  - Exercises covering theoretical, conceptual, and practical aspects of the material covered in class
- Final exam: Monday, February 5th (last week of lectures); 90 minutes; open book and notes. In order to participate in the final exam, students have to reach at least 50% of the points from homework assignments
- Homework assignments: Weekly, groups of up to three students (fixed groups)

Objectives

- Get a deep understanding of concepts, models, and techniques underlying the KDD process and different data mining techniques, including
  - data preprocessing
  - clustering
  - classification
  - regression
  - frequent pattern analysis
  - outlier detection
  - graph mining
- Be able to apply different data mining techniques to analyze and explore real-world datasets.
- Particular focus: how to employ **database techniques** in the context of KDD
  - large-scale datasets
  - database support (index structures, secondary storage based approaches, special operators, ...)

Course Content

1. Introduction
2. Foundations: data, statistics, probability theory
3. Clustering
4. Classification
5. Frequent Pattern Mining
6. Outlier Detection
7. [Graph Mining]
8. [Mining High-dimensional Data]
9. [Regression]

Literature

- Pang-Ning Tan, Michael Steinbach, and Vipin Kumar
  Introduction to Data Mining
  Addison-Wesley, 2005. ISBN: 0-321-32136-7
- Jiawei Han, Micheline Kamber, and Jian Pei
  Data Mining: Concepts and Techniques, 3rd edition
  Morgan Kaufmann, 2011. ISBN: 978-0123814791. URL: http://hanj.cs.illinois.edu/bk3/
- David J. Hand, Heikki Mannila, and Padhraic Smyth
  Principles of Data Mining
  MIT Press, 2001. ISBN: 9780262082907. URL:
  https://mitpress.mit.edu/books/principles-data-mining
- Martin Ester and Jörg Sander
  Knowledge Discovery in Databases - Techniken und Anwendungen
  Springer, 2000. ISBN: 3-540-67328-8 (in German)

Literature /2

- Trevor Hastie, Robert Tibshirani, and Jerome Friedman
  The Elements of Statistical Learning: Data Mining, Inference and Prediction
  Springer, 2009. ISBN: 978-0-387-84858-7. URL:
  https://statweb.stanford.edu/~tibs/ElemStatLearn/
- Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman
  Mining of Massive Datasets, 2nd Ed
  Cambridge University Press, 2014. ISBN: 978-1107077232. URL: http://www.mmds.org/#book
- Ian H. Witten, Frank Eibe, and Mark A. Hall
  Data mining: practical machine learning tools and techniques, 3rd Edition
  Morgan Kaufmann, Elsevier, 2011. ISBN: 9780123748560

Literature /3

Slides include material from:

- W. Lehner (TU Dresden)
- K.-U. Sattler (TU Ilmenau)
- M. Ester (Simon Fraser U), J. Sander (U Alberta)
- P. Tan, M. Steinbach, V. Kumar (Michigan State Univ., Univ. of Minnesota)
  http://www-users.cs.umn.edu/~kumar/dmbook/index.php
- J. Han, M. Kamber, J. Pei
  http://www.cs.uiuc.edu/~hanj/bk3/