Friedrich-Alexander-Universität
Technische Fakultät

1. Prologue

Knowledge Discovery in Databases

Dominik Probst, Dominik.probst@fau.de
Chair of Computer Science 6 (Data Management), Friedrich-Alexander-University Erlangen-Nürnberg
Summer semester 2024

Who? - Lecturer

Dominik Probst, M.Sc.

- Ph.D. candidate
- Computer Science 6 (Data Management)
- E-Mail: dominik.probst@fau.de
- Website: https://www.cs6.tf.fau.eu/dp

For whom? - Courses of Study

- **Knowledge Discovery in Databases with Exercise (KDDmUe)** - 5 ECTS
  - B.Sc./M.Sc. Data Science
  - B.Sc./M.Sc. Computer Science
  - M.Sc. International Information Systems
  - M.Sc. Medical Engineering
  - M.Sc. Information and Communication Technology
  - Possibly other courses of study (clarify with your examination office)

Important: Module KDD is no longer offered!

Last summer semester, we offered the module KDD (2.5 ECTS - only lecture) for some degree programmes. This module is no longer offered!

For whom? - Prerequisites

- **Mandatory requirements:**
  - Successful completion of the module „Konzeptionelle Modellierung“ (KonzMod)
    - Or a similar course teaching the basics of databases and SQL
- **Useful prerequisites:**
  - Successful completion of the module „Implementierung von Datenbanksystemen“ (IDB)
  - Experience with:
    - Python
    - Jupyter Notebooks
    - Numpy
    - Pandas
    - Algorithms
    - Data structures

What? - Goal and Topics

- **Goal of the module:**
  - Introduce you to the principles of data mining.
    $\Rightarrow$ This is the core of knowledge discovery in databases
- **Topics in the lecture:**
  1. Introduction
  2. Data
  3. Preprocessing
  4. Data Warehousing
  5. Mining Frequent Patterns
  6. Classification
  7. Cluster Analysis
  8. Outlier Analysis
  9. Guest lecture
     (Not part of the exam)
- **Topics in the exercise:**
  1. Introduction to python and pandas optional
  2. Data Analysis and Data Preprocessing
  3. Frequent Patterns
  4. Classification
  5. Clustering
  6. Outlier

What? - Exercises

- **Exercise sessions (in presence):**
  - Working together on exercise sheets
  - Either practical data science tasks or theoretical exercises
    (varies depending on the exercise sheet)
  - Required tools:
    - Laptop capable of running Jupyter Notebooks
      (preferably one per person, one per small group is also possible)
  - Expected preparation:
    - Good understanding of the lecture content
    - Completed "Preparation" section (see exercise sheets)

What? - Submissions

- **Programming tasks (to be done at home):**
  - Implementation of individual algorithms known from the lecture for a deeper understanding
    - Programming language: Python
    - Topics: Frequent Patterns, Classification, and Clustering
  - Have to be submitted to a GitHub classroom
    - Calculation of points performed automatically after each push
    - Improvements possible at any time until the submission deadline
    - Access to our mock exam unlocked upon achieving at least 75% of the available points
  - Work in small groups (up to three persons) is permitted
    - We conduct random checks for plagiarism across groups
      $\Rightarrow$ In cases where plagiarism is detected, all groups involved will receive zero points.

What? - Exam

- **Knowledge Discovery in Databases with Exercise (KDDmUe)** - Written Exam
  - Duration: 90 minutes
  - Questions about both lecture and exercise content
  - Language: English

Important: Do Not Forget to Register!

Without exception, we can only examine participants who have also registered for this exam at the examination office. Please note the information of the examination office as to when registration takes place in this semester.

What? - Literature

- **This lecture is based on the book by Han, Kamber, and Pei:**
  - J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*, 3rd. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2011, ISBN: 0123814790
  - Copies are available at the Science and Technology Branch Library (TNZB).
  - Lecture slides are based on slides provided by Jiawei Han with modifications by Prof. Dr.-Ing. Klaus Meyer-Wegener and Luciano Melodia.
  - Lecture slides have been modified and extended since then.
- **Further books on this topic include, but not limited to:**
  - A. Geŕon, *Hands-on machine learning with Scikit-Learn and TensorFlow : concepts, tools, and techniques to build intelligent systems*, 2nd ed. O'Reilly Media, 2017, ISBN: 978-1491962299
  - H. Du, *Data Mining Techniques and Applications: An Introduction*. Cengage Learning EMEA, May 2010, p. 336, ISBN: 978-1844808915
  - I. H. Witten, E. Frank, M. A. Hall, *et al.*, *Data Mining, Fourth Edition: Practical Machine Learning Tools and Techniques*, 4th. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2016, ISBN: 0128042915

When? - Dates

- **Lecture** - Start in Calendar Week 16 (Today)
  - Monday, 10:15 - 11:45 (H20)
    Lecturer: Dominik Probst

- **Exercises** - Start in Calendar Week 17
  - Group 1: Tuesday, 12:15 - 13:45 (H18)
    Tutor: Lucas Weber
  - Group 2: Monday, 16:15 - 17:45 (H4)
    Tutor: Dominik Probst
  - Group 3: Wednesday, 08:15 - 09:45 (H4)
    Tutor: Anugya Sahu
  - Group 4: Wednesday, 12:15 - 13:45 (H5)
    Tutor: Karan Pahlajani

Registration for Exercises

Registration for exercises is mandatory to ensure an appropriate support to questions regarding setting of exercises should they arise. Registration opens at April 15th, 12:00 o'clock via StudOn.

When? - Preliminary Schedule

Calendar Week	Lecture	Exercise	Submission
16	Prologue + Introduction		
17	Data	Introduction to Python & pandas (optional)	
18	Preprocessing	Group 1 & 2	
19		Data Analysis & Data Preprocessing	
20	Guest lecture + Data Warehousing		
21		Group 3 & 4	
22	Frequent Pattern		
23		Frequent Pattern	
24	Classification		Frequent Pattern
25		Classification	
26	Cluster Analysis		Classification
27			
28	Outlier Analysis + Exam Q&A	Clustering	Clustering
29			

Where? - StudOn

- Register at:
  `https://www.studon.fau.de/crs5533883_join.html`
- Main source for resources. E.g.:
  - Lecture slides
  - Exercise sheets
  - Forum
- Membership required to receive important updates on KDD by mail
- Questions should be asked here (StudOn Forum)

Where? - GitHub

- Public repository at: `https://github.com/FAU-CS6/KDD`
- Version control of our resources including:
  - Lecture slides
  - Exercise sheets

Help Appreciated: Error Corrections

Even though we strive for error-free lecture slides and practice sheets, there is still the possibility that errors have slipped in. You can help us mitigate these inaccuracies: Mail us, or better yet, in the case you have a GitHub account, open up a GitHub issue or create a pull request. Any pointers to errors are very much appreciated.

Any questions about this chapter?

Ask them now or ask them later in our forum:

StudOn Forum
https://www.studon.fau.de/frm5699567.html