Data Mining MTAT.03.183
(4AP = 6EAP)
Introduction

Jaak Vilo
2009 Fall
Lecturer

• 1986-1991 U Tartu
• 1991-1999 U Helsinki (sequence pattern discovery)
• 1999-2002 EMBL-EBI, UK (bioinformatics)
• 2002- EGeen -> Quretec (Biobank and Data Mgmnt)
• U Tartu, professor (Bioinformatics) 2007
  – EXCS – Center of Excellence
  – STACC – Software Technologies and Applications Competence Center (Tarkvara TAK)
  – research projects
Students

• >80 registered
• Estonian vs Foreign
• MSc 1st y / 2nd y ?
• BSc , PhD ?
• Non IT/CS ?
• Why this class? Expectations? (ESSCaSS’08,09...)
Course

• List: [ati.dm@lists.ut.ee](mailto:ati.dm@lists.ut.ee)
• Lectures: 10:15, Liivi 2-403
• Seminars: 12:15, Liivi 2-403
• Prof. Jaak Vilo  vilo@ut.ee
• [http://www.quicktopic.com/43/H/eWqhydvFpUN](http://www.quicktopic.com/43/H/eWqhydvFpUN)
• Other? Skype ?
Seminars

- Three types:
  1. Homework: presentations/discussions
  2. Guest lectures, visitors
  3. Practical labs/training (no concrete plans yet)

- Participation is obligatory (>75%)
Grading requirements

- **Participation!** >75% of seminars
- **Homeworks (30%)** (min 50% of assignments)
- Projects/essays (30%)
- Exam (40%)
- Total: 100% + thresholds
- All deadlines are stringent.
Homework

• Tasks/assignments
  – 5 tasks/week + possibly bonuses
  – About in every 2 weeks (irregular)

• Report/mark all completed tasks
  – written reports on tasks
  – ready to present fully to class
  – there will be some uploading system
  – and/or paper sheets in class

• Deadline always before class start (Thu, 12:15)
4AP = 6EAP

• 4 weeks (4x40h=160h) of **intensive work**
  – assuming basic knowledge of BSc material

• 1/3 in class
• 1/3 reading, homeworks
• 1/3 projects, writing, ...
What is Data Mining?

• Data -> Information, Knowledge, Insight
  – new, interesting, nontrivial, useful ...

• Data size -> Algorithmic challenge
• Predictive, useful -> theoretical and economical challenge

• Why? By practical demand and need...
Textbooks

• Han, Kamber: Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems) Google Books web
• Chakrabarti et al. Data Mining: know it all. Morgan Kaufmann 2008 @Elsevier @AMazon @Google
• Bramer: Principles of Data Mining (Springer, 2007) @Amazon @Springer @Google
• David J. Hand, Heikki Mannila and Padhraic Smyth: Principles of Data Mining (MIT Press, 2001) @MIT Press @Google
• Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (Springer 2009) @Tibshirani @Amazon
• **Han, Kamber**: Data Mining: Concepts and Techniques, Second Edition (The Morgan Kaufmann Series in Data Management Systems)

• **TOC**: [http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf](http://www.cs.uiuc.edu/homes/hanj/bk2/toc.pdf)
What’s it all about?
• Statistics
• Patterns in data
• Learning
• Classification
• Knowledge / Information /
• Algorithms
• Prediction
• …
Sources of data (growth)

• devices
• net/web
• logs
• transactional db
• consumer
• multimedia(!)
• science
• cheaper storage, compute power
• …
Why Data Mining?

- The Explosive Growth of Data: from terabytes to petabytes
  - Data collection and data availability
    - Automated data collection tools, database systems, Web, computerized society
  - Major sources of abundant data
    - Business: Web, e-commerce, transactions, stocks, ...
    - Science: Remote sensing, bioinformatics, scientific simulation, ...
    - Society and everyone: news, digital cameras, YouTube
  - We are drowning in data, but starving for knowledge!
  - “Necessity is the mother of invention”—Data mining—Automated analysis of massive data sets
Evolution of Sciences

- Before 1600, **empirical science**
- 1600-1950s, **theoretical science**
  - Each discipline has grown a *theoretical* component. Theoretical models often motivate experiments and generalize our understanding.
- 1950s-1990s, **computational science**
  - Over the last 50 years, most disciplines have grown a third, *computational* branch (e.g. empirical, theoretical, and computational ecology, or physics, or linguistics.)
  - Computational Science traditionally meant simulation. It grew out of our inability to find closed-form solutions for complex mathematical models.
- 1990-now, **data science**
  - The flood of data from new scientific instruments and simulations
  - The ability to economically store and manage petabytes of data online
  - The Internet and computing Grid that makes all these archives universally accessible
  - Scientific info. management, acquisition, organization, query, and visualization tasks scale almost linearly with data volumes. Data mining is a major new challenge!


Jiawei Han, Micheline Kamber, and Jian Pei, Data Mining: Concepts and Techniques
Evolution of Database Technology

- 1960s:
  - Data collection, database creation, IMS and network DBMS

- 1970s:
  - Relational data model, relational DBMS implementation

- 1980s:
  - RDBMS, advanced data models (extended-relational, OO, deductive, etc.)
  - Application-oriented DBMS (spatial, scientific, engineering, etc.)

- 1990s:
  - Data mining, data warehousing, multimedia databases, and Web databases

- 2000s:
  - Stream data management and mining
  - Data mining and its applications
  - Web technology (XML, data integration) and global information systems
examples from Machine Learning

- 1950’ies – checkers (Arthur Samuels 1959)
- 1960’ies – NN – perceptron and it’s limitations
- 1970’ies – expert systems, decision trees (ID3), …
- 1980’ies – Neural Networks, PAC learning, …
- 1990’ies – Data mining, ILP, Ensembles
- 2000’ – SVM, Kernels, Graphical Models, …
Chapter 1. Introduction

- Why Data Mining?
- What Is Data Mining?
- A Multi-Dimensional View of Data Mining
- Data Mining Functionalities: What Kinds of Patterns Can Be Mined?
- Data Mining: On What Kind of data?
- Time and Ordering: Sequential Pattern, Trend and Evolution Analysis
- Structure and Network Analysis
- Evaluation of Knowledge
- Applications of Data Mining
- Major Challenges in Data Mining
- A Brief History of Data Mining and Data Mining Society
- Summary
What Is Data Mining?

- Data mining (knowledge discovery from data)
  - Extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data
  - Data mining: a misnomer?

- Alternative names
  - Knowledge discovery (mining) in databases (KDD), knowledge extraction, data/pattern analysis, data archeology, data dredging, information harvesting, business intelligence, etc.

- Watch out: Is everything “data mining”?
  - Simple search and query processing
  - (Deductive) expert systems
Knowledge Discovery (KDD) Process

- This is a view from typical database systems and data warehousing communities.
- Data mining plays an essential role in the knowledge discovery process.
Example: A Web Mining Framework

- Web mining usually involves
  - Data cleaning
  - Data integration from multiple sources
  - Warehousing the data
  - Data cube construction
  - Data selection for data mining
  - Data mining
  - Presentation of the mining results
  - Patterns and knowledge to be used or stored into knowledge-base
Data Mining in Business Intelligence

Increasing potential to support business decisions

Decision Making

Data Presentation
Visualization Techniques

Data Mining
Information Discovery

Data Exploration
Statistical Summary, Querying, and Reporting

Data Preprocessing/Integration, Data Warehouses

Data Sources
Paper, Files, Web documents, Scientific experiments, Database Systems

End User
Business Analyst
Data Analyst
DBA

Jiawei Han, Micheline Kamber, and Jian Pei

Data Mining: Concepts and Techniques 24
Collaborative filtering

- Amazon, Netflicks

- Collaborative filtering systems usually take two steps:
  - Look for users who share the same rating patterns with the active user (the user whom the prediction is for).
  - Use the ratings from those like-minded users found in step 1 to calculate a prediction for the active user.
Netflix prize

http://www.netflixprize.com/


18K movies

480K customers

~ 100M ratings

Test on 2.8M withheld ratings
Social network

- Graph of connections
- Social network mining
Web

• Interlinked web sites and pages

• Directed Graph of links

• Information Retrieval, PageRank

• Web mining
Web usage mining

• Software and web usage logs

• Typical use patterns

• User groups, their preferences, behavior

• Can you predict their goals and help to achieve them?
  – distributed online transactions, queries, ... (Google, etc)

Jaak Vilo and other authors

UT: Data Mining 2009
Biomedical data mining

• Analyse:
  – DNA,
  – Genotype information
  – disease histories
  – find associated genes
  – predict and classify diseases and outcomes
  – discover “how biology works”
  – ...
Combinatorial Data Mining Algorithms
(research seminar, Sven Laur, PhD)

Basics ideas and techniques
  – How to find frequent sets in databases
  – How to find frequent motifs in sequences

Algorithmic problems
  – Depth-first vs breath first search
  – How to avoid combinatorial explosion

Interpretation of results
  – Which patterns are important enough?
Combinatorial Data Mining Algorithms
(research seminar, Sven Laur, PhD.)

Other important aspects
   – How to handle noisy data
   – Random sampling vs linear scan

Applications and extensions
   – Association rules in practice
   – Log analysis. Episode rules and usability
   – Graph mining and biochemistry
Combinatorial Data Mining Algorithms
(research seminar, Sven Laur, PhD)

Administrative details

• Combinatorial Data Mining Algorithms
• Gives 3 EAP (2 old AP)
• Takes place on Wednesdays in L122
• First seminar is on 16th of September
• Each participant has to give a presentation
• Project work is combined with DM course
Research at U Tartu


• STACC – Software Technologies and Applications Competence Center
  – companies and universities
  – Skype, Regio, Delfi, Quretec, ...

  – Research problems, topics, scholarships
Research topics

• Publications => Projects, funding

• Relevant to STACC, companies

• Can lead to job offers 😊
UT CS department

• Job offers:

• courses.cs.ut.ee - web site development
  – UT CS department courses web development
  – Other sysdamin and Department development tasks
  – ...