Data Mining MTAT.03.183 (6EAP)


Introduction

Jaak Vilo
2011 Fall
Lecturer

• 1986-1991 U Tartu (diploma)
• 1991-1999 U Helsinki (sequence pattern discovery, PhD)
• 1999-2002 EMBL-EBI, UK (bioinformatics)
• 2002- EGeen -> Quretec (Biobank and Data Mgmt)
• 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?
Course

• **Lectures:** Thursdays 10:15-12, Liivi 2-403
• **Seminars:**
  – group 1 - Tuesday, 16:15-18 Liivi 2-315
  – group 2 - Wednesday, 10:15-12 Liivi 2-122
  – group 3 - Thursday, 12:15-14 Liivi 2-315
  – group 4 - To be discussed
• **Mailing list:** [ati.dm@lists.ut.ee](mailto:ati.dm@lists.ut.ee)
• **Lecturer:** Prof. Jaak Vilo [vilo@ut.ee](mailto:vilo@ut.ee) (room 327)
• **Assistant:** Aivi Kaljuvee [Aivi.Kaljuvee@ut.ee](mailto:Aivi.Kaljuvee@ut.ee) (306)
Seminars

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

• Participation is obligatory (>75%)
Homework

• **Tasks/assignments**
  – 5 tasks/week + possibly bonuses

• **Report/mark all completed tasks**
  – *written reports* on tasks – upload system
    • Reasonable size solutions only
    • PDF/results preferred
  – **ready to present** in front of a group
  – TA does not need to read all solutions!

• **Deadline always before class start**
Grading requirements (100 points)

- **Homeworks (40%)**
  - Participation! >75% of seminars
  - Min 40% of assignments completed!
  - 10-12 weeks, 5 tasks = 50-60 tasks total + bonuses
  - 10 tasks – no points. After that each gives 1 point.

- **Projects + report (20%)** - obligatory

- **Exam (40%)** – with 50% threshold!

- **Total: 100% + thresholds**

- **All deadlines are strict.**
HW grading examples

• 10 tasks = 0 points
• 30 tasks = 20 points
• 50 tasks = 40 points
• 60 tasks = 50 points (max + 10)
• Each bonus task is extra!
4AP = 6EAP

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

• 25% in class
• 40% reading, homeworks
• 30% projects, writing, ...
• 5% exam
Drivers for data mining

• Massive data

• Faster computers, lot’s of disk

• Business and science needs

• Competition
Sources of data (growth)

- Devices and monitoring
- Internet/web
- logs
- transactional db
- consumer
- multimedia(!)
- Science: astronomy, biology, physics, ...
- cheaper storage, compute power
- ...

•...
DATA ≠ Information

- **DATA** – just raw DB or files
- **Information** – extracted facts, summaries, etc.
- **Wisdom/knowledge** – conclusions, interpretations and insights by humans
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...
RELATED ARTICLES

Bogus Grass-Roots Politics on Twitter
Data-mining techniques reveal fake Twitter accounts that give the impression of a vast political movement.

Data Mining for Dodgy Machines
A study highlights efforts to take down ISPs that allow malicious activity.

A Smarter Way to Dig Up Experts
Data-mining techniques could make it easier to locate expertise.
Germany vs. Facebook: Like Button Declared Illegal, Sites Threatened With Fine

Updated: German websites based in the state of Schleswig-Holstein have until the end of September to remove Facebook's 'like' button or face a fine of up to 50,000 Euro.

Germany has a long tradition of using laws to protect its citizen’s privacy. Home owners, for example, can ask Google to pixelate their houses in Street View (maybe so that their garden gnomes can stay incognito?). Facebook’s facial recognition feature has also come under fire in recent weeks. The latest target of Germany’s privacy advocates is Facebook’s ‘like’ button (‘Gefällt mir,’ in German). Thilo Weichert, the head of the Independent Centre for Privacy Protection of the northern German state of Schleswig-Holstein, argues that Internet sites based in his state that use the ‘like’ button are illegally sending this data to Facebook, which in turn uses it to illegally create a profile of its users web habits.

Note: the original article didn’t sufficiently stress the fact that Weichert’s jurisdiction is limited to Schleswig-Holstein only. I’ve updated the story to reflect this more clearly.

Weichert argues that data from any user who clicks the ‘like’ button – including those who
KDD conference

• We invite submission of papers describing innovative research on all aspects of knowledge discovery and data mining. Examples of topic of interest include (but are not limited to): classification and regression methods, semi-supervised learning, clustering, feature selection, social networks, mining of graph data, temporal and spatial data analysis, scalability, privacy, visualization, text analysis, Web mining, recommender systems, and so on.

• Papers emphasizing theoretical foundations are particularly encouraged, as are novel modeling and algorithmic approaches to specific data mining problems in scientific, business, medical, and engineering applications. We welcome submissions by authors who are new to the KDD conference, as well as visionary papers on new and emerging topics. Authors are explicitly discouraged from submitting papers that contain only incremental results and that do not provide significant advances over existing approaches.

• Submitted papers will be assessed based on their novelty, technical quality, potential impact, and clarity of writing. For papers that rely heavily on empirical evaluations, the experimental methods and results should be clear, well executed, and repeatable. Authors are strongly encouraged to make data and code publicly available when possible.
Tan, Steinbach, Kumar

- Introduction to Data Mining
- 2006 Pearson/Addison Wesley
- Seems the best single source
• **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)
Textbooks

- Tan, Steinbach, Kumar: Introduction to data mining
- Chakrabarti et al. Data Mining: know it all. Morgan Kaufmann 2008 [Elsevier](http://www.elsevier.com) [Amazon](http://www.amazon.com) [Google](http://www.google.com)
- Bramer: Principles of Data Mining (Springer, 2007) [Amazon](http://www.amazon.com) [Springer](http://www.springer.com) [Google](http://www.google.com)
- Trevor Hastie, Robert Tibshirani, Jerome Friedman: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. (Springer 2009) [Tibshirani](http://web.stanford.edu/~hastie) [Amazon](http://www.amazon.com)
What’s it all about?

Data

BIG DATA

SCIENCE IN THE PETABYTE ERA

DB
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!
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
Why Mine Data? Commercial Viewpoint

- Lots of data is being collected and warehoused
  - Web data, e-commerce
  - Purchases at department/grocery stores
  - Bank/Credit Card transactions

- Computers have become cheaper and more powerful

- Competitive Pressure is Strong
  - Provide better, customized services for an edge (e.g. in Customer Relationship Management)
Why Mine Data? Scientific Viewpoint

- Data collected and stored at enormous speeds (GB/hour)
  - remote sensors on a satellite
  - telescopes scanning the skies
  - microarrays generating gene expression data
  - scientific simulations generating terabytes of data

- Traditional techniques infeasible for raw data
- Data mining may help scientists
  - in classifying and segmenting data
  - in Hypothesis Formation
Mining Large Data Sets - Motivation

- There is often information “hidden” in the data that is not readily evident
- Human analysts may take weeks to discover useful information
- Much of the data is never analyzed at all

The Data Gap

From: R. Grossman, C. Kamath, V. Kumar, “Data Mining for Scientific and Engineering Applications”
What is Data Mining?

Many Definitions

- Non-trivial extraction of implicit, previously unknown and potentially useful information from data

- Exploration & analysis, by automatic or semi-automatic means, of large quantities of data in order to discover meaningful patterns
What is (not) Data Mining?

What is not Data Mining?

- Look up phone number in phone directory
- Query a Web search engine for information about “Amazon”

What is Data Mining?

- Certain names are more prevalent in certain US locations (O’Brien, O’Rurke, O’Reilly… in Boston area)
- Group together similar documents returned by search engine according to their context (e.g. Amazon rainforest, Amazon.com,)}
Origins of Data Mining

- Draws ideas from machine learning/AI, pattern recognition, statistics, and database systems.
- Traditional Techniques may be unsuitable due to:
  - Enormity of data
  - High dimensionality of data
  - Heterogeneous, distributed nature of data

Diagram:

Data Mining
- Statistics/AI
- Machine Learning/Pattern Recognition
- Database systems
Data Mining Tasks

- **Prediction Methods**
  - Use some variables to predict unknown or future values of other variables.

- **Description Methods**
  - Find human-interpretable patterns that describe the data.

From [Fayyad, et.al.] Advances in Knowledge Discovery and Data Mining, 1996
Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]
examples from Machine Learning

• 1950’ies – checkers (Arthur Samuels 1959)
• 1960’ies – NN – perceptron and its 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, ...
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

Data Cleaning → Data Integration → Data Warehouse → Task-relevant Data → Selection → Data Mining → Pattern Evaluation
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
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/

• http://en.wikipedia.org/wiki/Netflix_Prize

480K customers

18K movies

~ 100M ratings

Test on 2.8M withheld ratings
Social network

- Graph of connections
- Social network mining
Improve Healthcare, Win $3,000,000.

The goal of the prize is to develop a predictive algorithm that can identify patients who will be admitted to a hospital within the next year, using historical claims data.

More than 71 million individuals in the United States are admitted to hospitals each year, according to the latest survey from the American Hospital Association. Studies have concluded that in 2005, over $30 billion was spent on unnecessary hospital admissions. Is there a better way? Can we identify earlier those most at risk and ensure they get the treatment they need? The Heritage Provider Network (HPN) believes that the answer is "yes".

To achieve its goal of developing a breakthrough algorithm that uses available patient data to predict and prevent unnecessary hospitalizations, HPN is sponsoring the Heritage Health Prize Competition (the "Competition"). HPN believes that incentivized competition is the best way to achieve the radical breakthroughs necessary to begin fixing America's health care system.
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)
Biomedical data mining

• Analyse:
  – DNA,
  – Genotype information
  – disease histories
  – find associated genes
  – predict and classify diseases and outcomes
  – discover “how biology works”
  – ...

Jaak Vilo and other authors
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 😊