Data Mining MTAT.03.183 (6EAP)
https://courses.cs.ut.ee/2014/dm/
Introduction

Jaak Vilo
2014 Spring

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

Course times and contacts

• Lectures: Thursdays 10:15-12, Liivi 2-405
• Lecturer: Jaak Vilo  vilo@ut.ee (room 327)
• Practice – homework solutions seminars:
  – Wednesday 12:15 - 14:00 (402), Anna Leontjeva, anna.leontjeva@ut.ee
  – Thursday 12:15 - 14:00 (403) Dmytro Fishman
  – Friday 8:15 - 10:00 (402), Dmytro Fishman  dmytrofishman@gmail.com
• Mailing list: ati.dm@lists.ut.ee
  – sympa@lists.ut.ee - subscribe ati.dm
  – http://lists.ut.ee/

Seminars

• Homework: presentations/discussions
• Performing homeworks is obligatory (>50%)
• Participation is obligatory (>70%)

TA-s are here to help you – help them!

• HW to be submitted in PDF + script in
  language of preference (preferably ZIP file)
• Homeworks submitted before 23:59 the day before practice session e.g. for those who
  attend Wednesday practice sessions it is important to submit homework before 23:59
  on Tuesday and so on.
• Those who will use LaTeX will get +1000
  Karma points and our personal respect but it is not obligatory :)

Short CV

EMBL-EBI
EGeen

Quretec
STACC
BI2IT

UT: Data Mining 2011
Homework

• Tasks/assignments
  – 5 tasks/week + bonuses (optional, but helpful)
• 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!
  – Previous night 23:59 – to be able to provide feedback

Grading requirements (100 points)

• Homeworks (40%)
  – Participation! >70% of weeks/seminars
  – Min 50% of assignments completed!
  – 12 weeks x 5 tasks = 60 tasks in total + bonuses
  – Points: nr of tasks completed – 10. (50 tasks == max points)
• 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!
• 10 points == 1 grade!  C -> B,  B->A
• 6 points – 50% chance of improving grade!

6 ECTS

• 6*26=156h of intensive work
  assuming basic knowledge of BSc material
  Algorithms, data structures, probability, stats

• 25% in class
• 40% reading, homeworks
• 30% projects, writing, …
• 5% exam

Course working hours

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectures</td>
<td>13</td>
<td>1.5</td>
</tr>
<tr>
<td>Seminars</td>
<td>12</td>
<td>1.5</td>
</tr>
<tr>
<td>Weekly study</td>
<td>12</td>
<td>5</td>
</tr>
<tr>
<td>Project</td>
<td>40</td>
<td></td>
</tr>
<tr>
<td>Exam study</td>
<td></td>
<td>14.5</td>
</tr>
<tr>
<td>Exam</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></td>
<td><strong>156</strong></td>
</tr>
</tbody>
</table>

Students

• >85 registered
• Estonian vs Foreign
• MSc 1st y / 2nd y ?
• BSc, PhD ?
• Non IT/CS ?
• Why this class? Expectations?
Drivers for data mining

• Massive data
• Faster computers, lot’s of disk
• Business and science needs
• Competition

Sources of data (growth)

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

Original map made by John Snow in 1854. Cholera cases are highlighted in black.
Plumbr

- What has Java VM memory leak detection to do with Data Mining / Machine Learning?

DATA ≠ Information

- DATA – just raw DB or files
- Information – extracted facts, summaries, etc.
- Wisdom & knowledge – enable conclusions, interpretations and insights by humans

What is Data Mining?

- Data -> Information, Knowledge, Insight – new, interesting, nontrivial, useful ...
- Data size -> Algorithmic challenge
- Predictive -> Theoretical Challenge (ML)
- Useful -> Economical challenge
- Why? Practical demand and need...

quotes

- "Drowning in Data yet Starving for Knowledge" ???
- "Computers have promised us a fountain of wisdom but delivered a flood of data"
  William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus
- "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"
  T. S. Eliot

A Platform for Spatio-Temporal Data

- Built on cloud technologies
- Key Features
  - Scalable Stream based Plug-in Architecture
  - Repeatable and Reversible Processing
  - Automated and Manual Data Collection
  - Integrated Data Model: Observations & Models
  - All Major Ocean Data Types and Formats
  - Private and Public Users

DATA ≠ Information (UT: Data Mining 2011)

- "Drowning in Data yet Starving for Knowledge"

quotes

- "Computers have promised us a fountain of wisdom but delivered a flood of data"
  William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus
- "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"
  T. S. Eliot

Plumbr

- What has Java VM memory leak detection to do with Data Mining / Machine Learning?

DATA ≠ Information

- DATA – just raw DB or files
- Information – extracted facts, summaries, etc.
- Wisdom & knowledge – enable conclusions, interpretations and insights by humans

What is Data Mining?

- Data -> Information, Knowledge, Insight – new, interesting, nontrivial, useful ...
- Data size -> Algorithmic challenge
- Predictive -> Theoretical Challenge (ML)
- Useful -> Economical challenge
- Why? Practical demand and need...

quotes

- "Drowning in Data yet Starving for Knowledge" ???
- "Computers have promised us a fountain of wisdom but delivered a flood of data"
  William J. Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus
- "Where is the wisdom we have lost in knowledge? Where is the knowledge we have lost in information?"
  T. S. Eliot
Data Mining - an interdisciplinary field

- Databases
- Statistics
- High Performance Computing
- Machine Learning
- Visualization
- Mathematics
- ...
- Application areas!!

Q:

- What is the difference between data mining, statistics, machine learning and AI?
- Would it be accurate to say that they are 4 fields attempting to solve very similar problems but with different approaches? What exactly do they have in common and where do they differ? If there is some kind of hierarchy between them, what would it be?

Online starting points

  -- e.g. online courses:
  -- [http://www.kdnuggets.com/education/online.html](http://www.kdnuggets.com/education/online.html)
- [http://www.kaggle.com/competitions](http://www.kaggle.com/competitions)

"It's free, but they sell your information."

New Yorker

Jaak Vilo and other authors

UT: Data Mining 2011

Big data and DNA: What business can learn from junk genes

Nortol is an international high-end software development solutions provider in public and private sectors. The solution is built on keeping relationships and delivering the best quality recursive critical solutions to our clients in selected industries across Europe, Middle East and Africa. Nortol employs over 90,000 people and the growth is continuing.

YOUR FUTURE JOB WILL BE EASIER...

1. You should tell your visualization effort.
2. You should be a visualization expert in your work.
3. You should analyze your visualization effort.
4. You should be a visualization expert in your work.
5. You should analyze your visualization effort.

Gigaom

Big Data Gets Bigger: Now Google Trends Can Predict the Market

Big data and DNA: What business can learn from junk genes

A data scientist is a statistician who lives in San Francisco.

Data Science is statistics on a Mac.

A data scientist is someone who is better at statistics than any software engineer and better at software engineering than any statistician.
KDD: Papers submitted to the Research track are solicited in all areas of data mining, knowledge discovery, and large-scale data analytics, including, but not limited to:

- **Algorithms**: Graph and link mining, rule and pattern mining, web mining, dimensionality reduction and manifold learning, combinatorial optimization, relational and structured learning, matrix and tensor methods, classification and regression methods, semi-supervised learning, and unsupervised learning and clustering.

- **Applications**: Innovative applications that use data mining, including systems for social network analysis, recommender systems, mining sequences, time series analysis, online advertising, bioinformatics, systems biology, text/web analysis, mining temporal and spatial data, and multimedia processing.

- **Big Data**: Efficient and distributed data mining platforms and algorithms, systems for large-scale data analytics of textual and graph data, large-scale machine learning systems, distributed computing (cloud, map-reduce, Hadoop), large-scale optimization, and novel statistical techniques for big data.

- **Data mining for social good**: Novel algorithms and applications of data mining to societal problems is especially encouraged. (For deployment of existing algorithms consider the Industry/Govt. track.) Topics include: public policy, sustainability, climate change, medicine and health, education, transportation, biodiversity and energy.

- **Foundations of data mining**: Data mining methodology, data mining model selection, visualization, asymptotic analysis, information theory, and security and privacy.

Tan, Steinbach, Kumar

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


- By Jiawei Han, Micheline Kamber
- $115.95
- ISBN-10: 0123814706
- $265.22 - $34.25

FREE TWO-DAY SHIPPING FOR COLLEGE STUDENTS
What’s it all about?

Data

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

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’Rourke, 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

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.

Data Mining Tasks...

- Classification [Predictive]
- Clustering [Descriptive]
- Association Rule Discovery [Descriptive]
- Sequential Pattern Discovery [Descriptive]
- Regression [Predictive]
- Deviation Detection [Predictive]

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

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

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)

Biomedical data mining
- Analyse:
  - DNA,
  - Genotype information
  - disease histories
  - find associated genes
  - predict and classify diseases and outcomes
  - discover "how biology works"
  - ...

The Data Science Workflow

Philip Guo:
Contents of the course

• Frequent itemsets
• Data preprocessing and cleaning
  – Density estimation, outliers, normalisation, ...
• Clustering and Seriation
• Statistical correlations (conditional probability)
• Machine Learning (prediction) (decision trees, NN, SVM, ...)
• Visualisation
• OLAP and Data Cubes, Business Intelligence
• Text Mining
• Locality Sensitive Hashing
• Stream mining (?)

Research at U Tartu

• BIIT – [http://biit.cs.ut.ee/]
• 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 😊