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 Mgmnt)
- U Tartu, professor (Bioinformatics) 2007
  - EXCS – Center of Excellence
  - STACC – Software Technologies and Applications Competence Center (Tarkvara TAK)
  - research projects
Short CV

EMBL-EBI

EGeen

Estonian Genome Center
University of Tartu

QureTEC

BIOT

STACC
Course times and contacts

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

<table>
<thead>
<tr>
<th>Activity</th>
<th>Hours</th>
<th>Study Hours</th>
<th>Total Hours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lectures</td>
<td>13</td>
<td>1.5</td>
<td>19.5</td>
</tr>
<tr>
<td>Seminars</td>
<td>12</td>
<td>1.5</td>
<td>18</td>
</tr>
<tr>
<td>Weekly study</td>
<td>12</td>
<td>5</td>
<td>60</td>
</tr>
<tr>
<td>Project</td>
<td></td>
<td></td>
<td>40</td>
</tr>
<tr>
<td>Exam study</td>
<td></td>
<td></td>
<td>14.5</td>
</tr>
<tr>
<td>Exam</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td><strong>Total:</strong></td>
<td></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.
Original map made by John Snow in 1854. Cholera cases are highlighted in black.
2+ Million oceanic sensors and growing

Thousands of different data sources in a variety of locations and formats.

#1 footprint of public ocean data on the Web (marinexplore.org)
40,000+ data streams with 1 trillion data values by 31 institutions
100+ organizations and 6,000+ ocean data professionals
Fast insight without expensive software licenses

- Visualize devices and trajectories
- Heat map overlays (temp, sal, etc.)
- Quiver plots (winds, waves, currents)
- High Frequency Radar surface currents
- Graphs - track measurement values
- Interactive ADCP profiles
A Platform for Spatio-Temporal Data
Operational on 40,000+ data streams with 1 trillion data values by 31 institutions

Built on cloud technologies
- **StreamFlow** hybrid dataflow and workflow
- **StreamStorage** machine generated data and metadata storage
- **StreamWork** user experience for exploration, processing, management, collaboration, APIs

Key Features
- Scalable Stream-based Plugin Architecture
- Repeatable and Revertible Processing
- Automated and Manual Data Collection
- Integrated Data Model: Observations & Models
- All Major Ocean Data Types and Formats
- Private and Public Users
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

• [http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-1.html](http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-1.html)
Knowledge

Experience

Creativity
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)
<table>
<thead>
<tr>
<th>Active Competitions</th>
<th>Active Competitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All Competitions</strong></td>
<td><strong>Flight Quest 2: Flight Optimization</strong>&lt;br&gt;Final Phase of Flight Quest 2</td>
</tr>
<tr>
<td><strong>Masters</strong></td>
<td><strong>March Machine Learning Mania</strong>&lt;br&gt;Tip off college basketball by predicting the 2014 NCAA Tournament</td>
</tr>
<tr>
<td></td>
<td><strong>Flu Forecasting</strong> &lt;br&gt;Predict when, where and how strong the flu will be</td>
</tr>
<tr>
<td></td>
<td><strong>Galaxy Zoo - The Galaxy Challenge</strong>&lt;br&gt;Classify the morphologies of distant galaxies in our Universe</td>
</tr>
<tr>
<td></td>
<td><strong>Loan Default Prediction - Imperial College Lon...</strong>&lt;br&gt;Constructing an optimal portfolio of loans</td>
</tr>
<tr>
<td></td>
<td><strong>PAKDD 2014 - ASUS Malfunctional Component...</strong>&lt;br&gt;Predict malfunctional components of ASUS notebooks</td>
</tr>
<tr>
<td></td>
<td><strong>CONNECTOMICS</strong>&lt;br&gt;Reconstruct the wiring between neurons from fluorescence imaging of neural activity</td>
</tr>
</tbody>
</table>
“It’s free, but they sell your information.”
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

Written by: Frederic Lardinois
Big data and DNA: What business can learn from junk genes

The effort to dig out the data in our genomes has led to a rash of discoveries announced Wednesday, but amid the scientific insights are cultural ones that speak to how companies will have to learn to collaborate around big data and manage it.

The science world was rocked Wednesday by the discovery that 80 percent of the human gene that scientists thought was “junk” actually contains genetic regulators that can lead to diseases and certain genetic traits. It’s the scientific equivalent of discovering that ugly old dresser is actually a Louis XIV original, except that in this case, that dresser would also be filled with priceless books that might provide even more discoveries.
Nortal is an international high-end software development solutions provider in public and private sector. Our success is built on long-lasting relationships and delivering the best quality mission-critical solutions to our clients in selected industries across Europe, Middle East and Africa. Nortal employs over 650 people and the growth is continuing.

YOUR FUTURE JOB WILL BE EASIER IF...

• you know what J48 and C4.5 have in common.
• you know the name of the following formula and why you want to use it

\[ s_N = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \bar{x})^2} \]

• you know why data visualization is important
• you know a thing or two about data quality improvement steps within ETL
• you have good domain knowledge of public administration in general
• you yield excellent communication skills and can explain your ideas in „layman“ terms and make people believe you.
• you have an excellent command of English, readiness to travel and do “on-site” missions for various clients in different countries.
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.
Big Data Gets Bigger: Now Google Trends Can Predict The Market

Yesterday three economists, (Tobias Preis of Warwick Business School in the U.K., Helen Susannah Moat of University College London, and H. Eugene Stanley of Boston University) published an eye-opening paper that said Google Trends data was useful in predicting daily price moves in the Dow Jones industrial average, which consists of 30 stocks. Their research result:

"An uptick in Google searches on finance terms reliably predicted a fall in stock prices. “Debt” was the most reliable term for predicting market ups and downs, the researchers found. By going long when “debt” searches dropped and shorting the market when “debt” searches rose, the researchers were able to increase their hypothetical portfolio by 326 percent. (In comparison, a constant buy-and-hold strategy yielded just a 16 percent return.)

This was a 180-degree turnaround from earlier research, by Prof. Preiss, published back in 2010. Back in 2010, he used Google Trends data and found the opposite conclusion:

"The Google data could not predict the weekly fluctuations in stock prices. However, the team found a
Welcome to KDD 2014, an interdisciplinary conference that brings together researchers and practitioners from all aspects of data science, data mining, knowledge discovery, large-scale data analytics, and big data.

Submit Your Work to KDD 2014

Keynote Speakers

Dr. Oren Etzioni
Executive Director
Allen Institute for Artificial Intelligence

Dr. Eric Horvitz
Director
Microsoft Research

Dr. Eric Schadt
Director
Icahn Institute for Genomics and Multiscale Biology

Prof. Sendhil Mullainathan
Professor of Economics
Harvard University
**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, MPI), 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
What’s it all about?

Data

DB

Science in the Petabyte Era
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
Blog: Wayne Eckerson
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

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
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

• 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
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

1. **Data Sources**
   - Paper, Files, Web documents, Scientific experiments, Database Systems

2. **Data Preprocessing/Integration, Data Warehouses**

3. **Data Exploration**
   - Statistical Summary, Querying, and Reporting

4. **Data Mining**
   - Information Discovery

5. **Data Presentation**
   - Visualization Techniques

6. **Decision Making**

Roles:
- End User
- Business Analyst
- Data Analyst
- DBA

Jiawei Han, Micheline Kamber, and Jian Pei
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/


Jaak Vilo and other authors

18K movies

480K customers

≈ 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 2006 well over $33 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”
  – ...
The Data Science Workflow

Acquire data
  Reformat and clean data
  Explore alternatives

Analysis
  Edit analysis scripts
  Debug
  Inspect outputs
  Execute scripts

Dissemination
  Write reports
  Deploy online
  Archive experiment
  Share experiment

Reflection
  Make comparisons
  Take notes
  Hold meetings

Preparation
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


• 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 😊
The Software Technology and Applications Competence Center (STACC) is a research and development organization established in 2009 to conduct high-priority applied research in Estonia. STACC is a joint initiative between the University of Tartu and Tallinn Technical University as the centers of science in Estonia, and the leading IT companies and users of Estonian software and knowledge-based technology. The mission of STACC is to enhance the skills and competitiveness of organizations that are closely involved in software technology in Estonia, while increasing their export potential and gains on software solutions.

Expressions of interest are sought from MASTER STUDENTS in Computer Science doing their thesis in the following areas:

- Data mining in social networks and network data
- Dalvik JVM (Android) performance monitoring tools, mapping the state of the art
- Developing the classifier repository’s data model based on the Neuchatel model

The successful candidate will work on their MSc thesis in a STACC research project.

The position will be for 8 months with an option for extension. The salary range will be 400-600 eur brutto per month for 50% workload. The starting date is 1.10.2012.

Inquiries and expressions of interest (including research statement and CV) should be sent by e-mail to Prof. Jaak Vilo (jaak.vilo [at] ut.ee) and Indrek Vainu (indrek.vainu [at] stacc.ee) by 21. September. More information on STACC is available on www.stacc.ee