Data Mining MTAT.03.183 (6EAP)

http://courses.cs.ut.ee/2012/DM/

Introduction

Jaak Vilo
2012 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

Short CV

EMBL-EBI

- 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

- >60 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 2-405
- Seminars:
  - group 1 - Thursday, 12:15-14 Liivi 2-403
  - group 2 - Thursday, 10:15-12 Liivi 2-404
- Mailing list: att.dm@lists.ut.ee
- Lecturer: Prof. Jaak Vilo vilo@ut.ee (room 327)
- Assistant: Tauno Metsalu Tauno.Metsalu@ut.ee (314)

Seminars

- Homework: presentations/discussions
- Performing homeworks is obligatory (>50%)
- 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 50% of assignments completed!
  - 10-12 weeks, 5 tasks = 50-60 tasks total + bonuses
  - Points: nr of tasks completed – 10.
- 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 (6*26=156) of intensive work
  - assuming basic knowledge of BSc material
- 25% in class
- 40% reading, homeworks
- 30% projects, writing, ...
- 5% exam

Course

- ~13 Lectures 26h
- Self-study 52h
- ~10 Practicals 20h
- Project work 40h
- Exam 14h + 4h
- Total 6*26 h (6EAP)

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
• ...

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, useful -> theoretical and economical challenge

• Why? By practical demand and need...

http://www.cs.ccsu.edu/~markov/ccsu_courses/datamining-1.html

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?

Education online

• http://www.kdnuggets.com/education/online.html
KDD conference

- We invite submission of papers describing innovative research on all aspects of knowledge discovery and data mining. Examples of topics 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

Textbooks


What’s it all about?

- Introduction to data mining
- Google Books
- 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
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]

Examples from Machine Learning

- 1950’s – checkers (Arthur Samuel
- 1960’s – NN – perceptron and its limitations
- 1970’s – expert systems, decision trees (ID3), ...
- 1980’s – Neural Networks, PAC learning, ...
- 1990’s – Data mining, ILP, Ensembles
- 2000’s – 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
Collaborative filtering

– Amazon, Netflxks

• 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


Soal 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”
    - ...

Contents

- 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
- 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 😊