Data Mining
Analiza i eksploracja danych

Lecturer: JERZY STEFANOWSKI
Institute of Computing Science
Poznań University of Technology
Poznań, Poland

Software Engineering – Master Course
Computer Science, PUT, 2008
Course mission

• Data mining?
  • Extraction of useful information patterns from data.
  • Huge or complex data challenge (DB community).
    • Trend to data warehouses but also flat table files.
  • More than typical data analysis, machine learning or classical decision support!
  • Still „young” field, although quite „fashionable”.
  • Teaching materials and course book …
Course mission - 2

• Thus, our aims:
  • To give you insight in this field, typical methods, examples of applications.
  • Rather focus on algorithms and methodology aspects, not so much on dealing with massive data.
  • Choice of methods.
  • Case study and applications.
  • Comments to available software (WEKA, etc.)
  • Data Sets – UCI Repository and others
Course information

• The planned schedule of lectures (15 weeks):

  • Introduction
  • Data Preprocessing
  • Prediction models (multivariate regression and ANN)
  • Classification (evaluation)
    • Symbolic methods (Decision Trees and Rules)
    • Other on-symbolic methods
  • Association rules and sequence patterns
  • Clustering
  • Mining complex data
  • KDD Process and summary
Background literature

• Han Jiawei and Kamber M. Data mining: Concepts and techniques, Morgan Kaufmann, 2001 (1 ed.), there is 2d
Background literature [Polish translations]

- Larose D., Metody i modele eksploracji danych, PWN, 2008.

Polskie książki
- Dobry podręcznik jeszcze nie istnieje …
Background literature
Acknowledgements:

- Many of the slides are based on my earlier courses:
  - Data mining and advanced data analysis; Knowledge discovery (PUT CS; M.Sc. Course) more at http://www.cs.put.poznan.pl/jstefanowski
  - Some slides are based on ideas „borrowed” from:
    - WEKA teaching materials (Witten & Frank Waikato University; Morgan Kaufmann)
    - Gregory Piatetsky – Shapiro: Data mining course.
    - Jiawei Han: Knowledge discovery in databases.
    - T.Mitchell and P.Flach courses on ML (see their WWW).
- Other course books – see the previous slides
Lecture 1 a.

Data Mining:
Introduction
L1. Introduction to Data Mining: Outline

- Motivation: Why data mining?
- What is data mining?
- Data Mining: On what kind of data and what kind of knowledge representation?
- Basic tasks and methods.
- Examples of applications.
- WEKA and Statistica – software frameworks for this course.
Data mining: what is it?

• Data mining is
  • Extraction of useful patterns from data sources, e.g., databases, texts, web, images.

• Patterns (knowledge representation) must be:
  • Valid, novel, potentially useful, understandable to the users.
What is data mining? More ...

- Data mining is the analysis of data for relationships that have not previously been discovered or known.
- A term coined for a new discipline lying at the interface of database technology, machine learning, pattern recognition, statistics and visualization.
- The key element in much more elaborate process called "Knowledge Discovery in Databases".
- The efficient extraction of previously unknown patterns in very large data bases.
- Data mining is the analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Hand, Mannila, Smyth).
Motivations - data explosion problem

- Automated data collection tools and mature database technology lead to tremendous amounts of data stored in databases, data warehouses and other information repositories.

- More data is generated:
  - Bank, telecom, other business transactions ...
  - Scientific data: astronomy, biology, etc
  - Web, text, and e-commerce

- Very little data will ever be looked at by a human!

- We are drowning in data, but starving for knowledge!

- Knowledge Discovery is **NEEDED** to make sense and use of data.
Big Data Examples

• Europe's Very Long Baseline Interferometry (VLBI) has 16 telescopes, each of which produces 1 Gigabit/second of astronomical data over a 25-day observation session
  • storage and analysis a big problem

• AT&T handles billions of calls per day
  • so much data, it cannot be all stored -- analysis has to be done “on the fly”, on streaming data
Largest databases in 2003

- Commercial databases:
  - Winter Corp. 2003 Survey: France Telecom has largest decision-support DB, ~30TB; AT&T ~ 26 TB

- Web
  - Alexa internet archive: 7 years of data, 500 TB
  - Google searches 4+ Billion pages, many hundreds TB
  - Internet Archive (www.archive.org), ~ 300 TB

- UC Berkeley 2003 estimate: 5 exabytes (5 million terabytes) of new data was created in 2002.
  

- US produces ~40% of new stored data worldwide
We are Data Rich but Information Poor

Databases are too big

Data Mining can help discover knowledge nuggets

„Terrorbytes“
Knowledge Discovery Definition

- Knowledge Discovery in Data is the *non-trivial* process of identifying
  - *valid*
  - *novel*
  - potentially *useful*
  - and ultimately *understandable patterns* in data.

from *Advances in Knowledge Discovery and Data Mining*,

The name first used by AI, Machine Learning Community in 1989 Workshop at AAAI Conference.
Data Mining as a step in A KDD Process

- Data mining: the core step of knowledge discovery process.
Steps of a KDD Process

- Learning the application domain:
  - relevant prior knowledge and goals of application
- Creating a target data set: data selection
- Data cleaning and preprocessing: (may take 60% of effort!)
- Data reduction and projection:
  - Find useful features, dimensionality/variable reduction, invariant representation.
- Choosing a task of a data mining step
  - summarization, classification, regression, association, clustering.
- Choosing proper mining algorithm(s)
- Data mining: searching for patterns of interest
- Interpretation: analysis of results.
  - visualization, transformation, removing redundant patterns, etc.
- Use of discovered knowledge
SKICAT – an example of KDD process

- **Sky Image Cataloging and Analysis Tool (SKICAT)**
- Developed by NASA's Jet Propulsion Laboratory and the California Institute of Technology in the 90's (Fayyad, Djorgowski, Weir et al.).
- **Aim**: A software system to catalog and analyze the estimated half billion sky objects in the second Palomar Observatory sky survey
  - Task 1 – general classes (galaxies, stars, quasars, etc.)
  - Task 2 – find some interesting clusters of objects (quasars with redshift..)
- The survey of the northern sky includes more than 3,000 digitized photographic plates produced by Palomar, located in San Diego.
  - Over 3 terrabytes of images $13000 \times 13000$ pixels
- The SKICAT system will produce a comprehensive survey catalog database containing about one-half billion entries by automatically processing about three terabytes (24 trillion bits, 8-bits to a byte) of image data.
- SKICAT is based on state-of-the-art machine learning, high performance database and image processing techniques.
- SKICAT has a correct sky object classification rate of about 94 percent, which exceeds the performance requirement of 90 percent needed for accurate scientific analysis of the data.
SKICAT – KDD steps

- **Dostępne dane i wiedza początkowa**: kilkadziesiąt tysięcy fotografii o różnych rozdzielczości ręcznie skatalogowanych przez ekspertów.

- **Wybór docelowych danych** – identyfikacja właściwych atrybutów charakteryzujących poszczególne klasy - konieczność uwzględnienia dodatkowej wiedzy astronomicznej and image analysis FOCAS.

- **Krok czyszczenia** – identyfikacja różnych obserwacji odstających i błędów w danych.

- **Krok redukcji** – wybór tylko części z dostępnych danych.

- **Wybór zadania i algorytmu**: klasyfikowanie – drzewa decyzyjne (modyfikacja C4.5, ale także alternatywne klasyfikatory).

- **Ocena rezultatów** (trafność klasyfikacji – tutaj ponad 94%).

- **Zastosowanie** – wspomaganie tworzenia elektronicznego katalogu gwiazd i galaktyk wraz z ich opisami.
Related Fields

Machine Learning

Visualization

Data Mining and Knowledge Discovery

Statistics

Databases
Related fields

- Statistics – model driven approach based on a strong mathematical background; model-hypothesis testing or estimation methods.
  - Modern statistical learning.
- Machine learning – improving performance of intelligent systems; knowing concepts to be learned; strongly related to AI paradigms.
- Databases systems – efficient data management and extractions; data warehouses.
- Knowledge based system – representation and AI methods.
Data mining and related disciplines

What is not data mining?

• Another statistical approach shifted into a new context!
• It is not only machine learning!
• Moreover, it is not
  • (Database) query processing,
  • Expert systems software.
Data Mining: On What Kind of Data?

- Attribute-value tables (standard form / data table)
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, …)
- Sequence data bases
- Other more complex data repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - WWW resources
  - …
What can be discovered?

- Symbolic knowledge representations
  - Decision trees
  - Rules
  - Relations / Logic forms (ILP)
  - Attribute generalizations
  - Probability distributions
  - Conceptual clusters and taxonomies
- Sub-symbolic
  - Artificial neural networks
  - Instance based learners
  - Functions / equations
- Others, ...

Decision trees

- Typical approach to the classification task.

<table>
<thead>
<tr>
<th>Age</th>
<th>Car Type</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Combi</td>
<td>High</td>
</tr>
<tr>
<td>18</td>
<td>Sports</td>
<td>High</td>
</tr>
<tr>
<td>40</td>
<td>Sports</td>
<td>High</td>
</tr>
<tr>
<td>50</td>
<td>Family</td>
<td>Low</td>
</tr>
<tr>
<td>35</td>
<td>Minivan</td>
<td>Low</td>
</tr>
<tr>
<td>30</td>
<td>Combi</td>
<td>High</td>
</tr>
<tr>
<td>32</td>
<td>Family</td>
<td>Low</td>
</tr>
<tr>
<td>40</td>
<td>Combi</td>
<td>Low</td>
</tr>
</tbody>
</table>
Decision rules

<table>
<thead>
<tr>
<th>Age</th>
<th>Car Type</th>
<th>Risk</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Combi</td>
<td>High</td>
</tr>
<tr>
<td>18</td>
<td>Sports</td>
<td>High</td>
</tr>
<tr>
<td>40</td>
<td>Sports</td>
<td>High</td>
</tr>
<tr>
<td>50</td>
<td>Family</td>
<td>Low</td>
</tr>
<tr>
<td>35</td>
<td>Minivan</td>
<td>Low</td>
</tr>
<tr>
<td>30</td>
<td>Combi</td>
<td>High</td>
</tr>
<tr>
<td>32</td>
<td>Family</td>
<td>Low</td>
</tr>
<tr>
<td>40</td>
<td>Combi</td>
<td>Low</td>
</tr>
</tbody>
</table>

Classification / Learning algorithm

if age < 31
or Car Type = Sports
then Risk = High
Rules with variables - ILP

• Using variables and multiple relations:

If height_and_width_of(x,h,w) and h > w
then standing(x)

• The top of a tower of blocks is standing:

If height_and_width_of(x,h,w) and h > w
and is_top_of(x,y)
then standing(x)

• The whole tower is standing:

If height_and_width_of(z,h,w) and h > w
and is_top_of(x,z) and standing(y)
and is_rest_of(x,y)
then standing(x)
If empty(x) then standing(x)
 Numeric prediction – regression function

- Example: 209 different computer configurations

<table>
<thead>
<tr>
<th>Cycle time (ns)</th>
<th>Main memory (Kb)</th>
<th>Cache (Kb)</th>
<th>Channels</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>MYCT</td>
<td>MMIN</td>
<td>MMAX</td>
<td>CACH</td>
<td>CHMIN</td>
</tr>
<tr>
<td>1</td>
<td>125</td>
<td>256</td>
<td>6000</td>
<td>256</td>
</tr>
<tr>
<td>2</td>
<td>29</td>
<td>8000</td>
<td>32000</td>
<td>32</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>208</td>
<td>480</td>
<td>512</td>
<td>8000</td>
<td>32</td>
</tr>
<tr>
<td>209</td>
<td>480</td>
<td>1000</td>
<td>4000</td>
<td>0</td>
</tr>
</tbody>
</table>

- Linear regression function

\[ PRP = -55.9 + 0.0489 \text{ MYCT } + 0.0153 \text{ MMIN } + 0.0056 \text{ MMAX } + 0.6410 \text{ CACH } - 0.2700 \text{ CHMIN } + 1.480 \text{ CHMAX} \]
Association rules

• Transaction data
• Market basket analysis

{Cheese, Milk} $\rightarrow$ Bread [sup=5%, conf=80%]

Association rule:
"80% of customers who buy cheese and milk also buy bread and 5% of customers buy all these products together"
Sequential Patterns

• Sequential pattern mining is the extracting of frequently occurring patterns related to time or other sequences.

• A sequential rule: $A \rightarrow B \rightarrow C$, says that event A will be followed by event B and this by event C with a certain confidence.

• An example:

  “A customer who bought a TV three months ago is likely to order a new DVD player within one month”
Cluster

• A group of examples similar in the sense of some distance measure.

• Unsupervised data

Figure 5. A Simple Clustering of the Loan Data Set into Three Clusters. Note that original labels are replaced by a +.
Non-symbolic? Neural networks

- Single Neuron Model
  - Linear Threshold Unit (LTU)
    - inputs to unit: defined as linear combination
  - Output of unit: threshold (activation) function on net input (threshold $\theta = w_0$)
- Neural Networks
  - Neuron is modeled using a unit connected by weighted links $w_i$ to other units
  - Multi-Layer Perceptron (MLP): future lecture

Vector notation: $o(x) = \text{sgn}(\bar{x}, \bar{w}) = \begin{cases} 
1 & \text{if } \bar{w} \cdot \bar{x} > 0 \\
-1 & \text{otherwise}
\end{cases}$
Major Data Mining Tasks

- **Classification**: predicting an instance class on the basis of its description.
- **Prediction**: predicting a continuous value.
- **Clustering**: finding similarity groups in data.
- **Associations**: e.g. A & B & C occur frequently.
  - Sequence analysis
- **Summarization**: describing a group.
- **Visualization**: graphical methods to show patterns in data.
- **Deviation or Anomaly Detection**: finding important changes or anomalies
- ...
Supervised classification

- Discover a method / function that maps a data element / instance into one of predefined classes
- Learning how to predict the instance class from pre-labeled (classified) examples - classifiers

Many approaches:
- Discriminant Analysis,
- Decision Trees or Rules,
- Bayesian Approaches,
- SVM

... 

Given a set of points from classes ● ○
what is the class of new point ○?
Numeric prediction

- Supervised task, but “output” is numeric,
  - Scheme is being provided with target value
- Evaluation - good statistical background.

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temp.</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play-time</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sunny</td>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>5</td>
</tr>
<tr>
<td>Hot</td>
<td>High</td>
<td>True</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Hot</td>
<td>High</td>
<td>False</td>
<td>55</td>
<td></td>
</tr>
<tr>
<td>Mild</td>
<td>Normal</td>
<td>False</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4. A Simple Linear Regression for the Loan Data Set.
Time series – prediction of time stamped variable

- Predictive data mining – time stamped variable; typical example – stocks, financial data, production planning.
Clustering

Find “natural” grouping of un-labeled examples – similar in some well defined sense.
Clustering

- Examples: customer grouping
- Finding groups of items that are similar
- Clustering is *unsupervised*
  - The class of an example is not known
- Success often measured subjectively

<table>
<thead>
<tr>
<th>Type</th>
<th>Sepal length</th>
<th>Sepal width</th>
<th>Petal length</th>
<th>Petal width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iris virginica</td>
<td>5.8</td>
<td>2.7</td>
<td>5.1</td>
<td>1.9</td>
</tr>
</tbody>
</table>
| Iris versicolor| 6.3          | 3.3         | 6.0          | 2.5         | Iris virginica
| Iris versicolor| 6.4          | 3.2         | 4.5          | 1.5         |
| Iris versicolor| 7.0          | 3.2         | 4.7          | 1.4         |
| Iris setosa    | 4.9          | 3.0         | 1.4          | 0.2         |
| Iris setosa    | 5.1          | 3.5         | 1.4          | 0.2         |
| ...            |              |             |              |             |
| Iris virginica | 101          | 3.3         | 6.0          | 2.5         | Iris virginica
| Iris virginica | 102          | 2.7         | 5.1          | 1.9         |
Descriptive clustering – text mining

• Discovering diverse groups of semantically related documents described with meaningful, comprehensible and compact text labels → LINGO available in Carrot2 project.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Title</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Iraqi Interim Government</td>
<td>official interim government site for Iraq. Includes profiles of officials and information on issues such as elections, security, reconstruction, law, and the Iraqi constitution.</td>
<td><a href="http://www.iraqigovernment.org/">http://www.iraqigovernment.org/</a></td>
</tr>
<tr>
<td>2</td>
<td>BBC News: The Struggle for Iraq</td>
<td>full coverage of the issues and opinions surrounding the Iraq crisis, including the Iraqi constitution, elections, ongoing conflicts and attacks in the region, the trial against Saddam Hussein, daily life in Iraq, and investigations into the reports of prisoner abuse.</td>
<td>[<a href="http://news.bbc.co.uk/1/hi/in_depth/middle_east/2002/conflict_with">http://news.bbc.co.uk/1/hi/in_depth/middle_east/2002/conflict_with</a> IRAQ/default.stm](<a href="http://news.bbc.co.uk/1/hi/in_depth/middle_east/2002/conflict_with">http://news.bbc.co.uk/1/hi/in_depth/middle_east/2002/conflict_with</a> IRAQ/default.stm)</td>
</tr>
<tr>
<td>4</td>
<td>Yahoo! News Full Coverage: Iraq</td>
<td>ongoing, in-depth coverage of the conflict in Iraq, including photos, video, news, and analysis, as well as reports on the Iraqi parliament and attacks in the region.</td>
<td><a href="http://news.yahoo.com/ic/World/Iraq">http://news.yahoo.com/ic/World/Iraq</a></td>
</tr>
<tr>
<td>6</td>
<td>CNN.com: Transition of Power</td>
<td>ongoing coverage of the reconstruction of Iraq and the transition of power.</td>
<td><a href="http://www.cnn.com">http://www.cnn.com</a></td>
</tr>
</tbody>
</table>
Another problem: Link Analysis

Can we find patterns in the network/graph structure or
Can we identify communities of Web users?
Goals of Data Mining – S. Weiss and U. Fayyad views

- Two primary goals of data mining
  - **Prediction**: using some variables or fields in database to predict unknown of future value of other variables.
    - strongest and practical goal of data mining and other fields as pattern recognition, machine or statistical learning
  - **Knowledge description**: finding human-interpretable patterns describing data
    - Becoming more and more important for KDD
  - Technical methods for solving these problems often overlap.

<table>
<thead>
<tr>
<th>Prediction</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Deviation Detection</td>
</tr>
<tr>
<td>Regression</td>
<td>Clustering</td>
</tr>
<tr>
<td>Time series</td>
<td>Association Rules</td>
</tr>
<tr>
<td></td>
<td>Summarization</td>
</tr>
<tr>
<td></td>
<td>Visualization</td>
</tr>
<tr>
<td></td>
<td>Text mining</td>
</tr>
</tbody>
</table>

Classification, Regression, Time series for Prediction

Deviation Detection, Clustering, Association Rules, Summarization, Visualization, Text mining for Knowledge description
Components of Data Mining Algorithms

- **Model or pattern structure → Knowledge representation;** language determining the underlining structure forms that we seek from the data.

- **Score function →** estimates how well a particular patterns (or a model with its parameters) meet the criteria of KDD process.

- **Optimization of parameters and search method:**
  - Searching for the parameters which optimize the score function (quality of the model) → optimal estimation vs. heuristic or greedy iterative..
  - Searching over different model or pattern spaces

- **Data management strategy → handling data access efficiently during search/optimization**

Why Data Mining? -- Potential Applications

- Database analysis and decision support
  - Market and customer analysis; analytical CRM
    - Target marketing and advertising, customer relation management, market basket analysis, cross selling, market segmentation.
  - Risk analysis and management
    - Forecasting, customer changes, quality control, loan approval.
  - Diagnostics (e.g. technical conditions of objects)
  - Fraud detection
- Other Applications:
  - Text mining (news group, email, documents) and Web analysis; Search engines; Adaptive Web servers and e-commerce systems
Customer Changes: Case Study

• Situation: Attrition rate at for mobile phone customers is around 25-30% a year (US)!

Task:

• Given customer information for the past $N$ months, predict who is likely to attrite next month.

• Also, estimate customer value and what is the cost-effective offer to be made to this customer.
Customer Attrition Results

- Verizon Wireless (US) built a customer data warehouse
- Identified potential attriters
- Developed multiple, regional models
- Targeted customers with high propensity to accept the offer
- Reduced attrition rate from over 2%/month to under 1.5%/month (huge impact, with >30 M subscribers)

(Reported in 2003)
Biology: Molecular Diagnostics

- Leukemia: Acute Lymphoblastic (ALL) vs Acute Myeloid (AML)
  - 72 samples, about 7,000 genes

Results: 33 correct (97% accuracy), 1 error (sample suspected mislabelled)

Outcome predictions?
Problems Suitable for Data-Mining

• Require knowledge-based decisions.
• Have a changing environment.
• Have accessible, sufficient, and relevant data.
• Data are difficult or complex.
• Problems are not trivial, cannot be solved "manually" by humans.
• Provides high payoff for the right decisions!

Privacy considerations important if personal data is involved!
Warnings!

• Historically data mining was used in a pejorative sense by statisticians for the idea that, if you search long enough, you can always find some model to fit your data arbitrarily well.

Example:

• David Rhine, a "parapsychologist" at Duke in the 1950’s tested students for "extrasensory perception", by asking them to guess 10 cards—red or black. He found about 1/1000 of them guessed all 10, and instead of realizing that that is what you would expect from random guessing, declared them to have ESP. When he retested them, he found they did no better than average. His conclusion: telling people they have ESP causes them to lose it! Quote from Jeffrey Ullman, Stanford
Weka – software for data mining

• Waikato Environment for Knowledge Analysis (WEKA); developed by the Department of Computer Science, University of Waikato, New Zealand

• Data mining / Machine learning software written in Java (distributed under the GNU Public License)

• Used for research, education, and applications

http://www.cs.waikato.ac.nz/ml/weka/

• Ian Witten, Eibe Frank
WEKA - Tools

- Pre-processing Filters
- Attribute selection
- Classification/Regression
- Clustering
- Association discovery
- Visualization
Examples of Systems for Data Mining

- IBM: QUEST and Intelligent Miner
- Silicon Graphics: MineSet
- SAS Institute: Enterprise Miner
- SPSS / Integral Solutions Ltd.: Clementine
- Oracle Miner
- Rapid Miner (YALE)
- Orange
- Other systems
  - Information Discovery Inc.: Data Mining Suite
  - SFU: DBMiner, GeoMiner, MultiMediaMiner
RapidMiner (YALE)

TESTIMONIALS

"I have encountered various learning environments, but none so broad, powerful, and easy-to-use as RapidMiner / YALE. Many of us who are not skilled in programming are thankful."

Roberto E. Ferrer, Venezuela

DOWNLOADS

RapidMiner / YALE
RapidMiner / YALE Plugins
RapidMiner / YALE Documentation
RapidMiner / YALE Interactive Tour

TRAINING SEMINARS

Data Mining for Marketing and Customer Service
Data Mining Techniques: Theory and Practice
Extending RapidMiner and Integration as a Data
Orange (Slovenia)

Following are screenshots of Orange Widgets and Orange’s visual programming interface for data mining.

Classification tree viewer with a navigator.
Data mining: what is on the market?

• Consultancy
  • There are some people which offer data mining consulting.

• Software
  SAS, Statistica, Oracle, SGI Miner, IBM intelligent miner, …
Summary:

• Technology trends lead to data flood
  • data mining is needed to make sense of data

• Data Mining has many applications, successful and not

• Knowledge Discovery Process

• Data Mining Tasks
  • classification, prediction, clustering, associations, …
Short Break
Lecture 1b.

Few remarks on the nature of data to be mined
Data Mining: On What Kind of Data?

- **Flat files** → *Attribute-value data tables*
- Relational databases and data warehouses → relations.
- Multi-relational data / first order predicate calculus
- Structured data (graphs, workflows, ontologies, …)
- Sequence data bases
- Other advanced data repositories
  - Object-oriented and object-relational databases
  - Spatial databases
  - Time-series data and temporal data
  - Text databases and multimedia databases
  - WWW resources
  - …
Flat files

- Actually the most common data source for data mining, especially at the research level.

- Simple data files in text or binary format with a structure known by the data mining algorithm to be applied.

- The data in these files can be transactions, time-series data, scientific measurements, etc.

- Big data – efficiency of access and management.

<table>
<thead>
<tr>
<th>Instance</th>
<th>f 1</th>
<th>...</th>
<th>f_k</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>V 1,1</td>
<td>...</td>
<td>V 1,k</td>
<td>V 1,k+1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>x_i</td>
<td>V i,1</td>
<td>...</td>
<td>V i,k</td>
<td>V i,k+1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>x_n</td>
<td>V n,1</td>
<td>...</td>
<td>V n,k</td>
<td>V n,k+1</td>
</tr>
</tbody>
</table>
A data table is a set of measurements taken from some environment of process \((n \times k \text{ table})\).

In general, the input takes the form of: concepts-classes, instances, attributes.

- **Target concepts (classes):** kinds of things that can be discovered
  - **Aim:** intelligible and operational concept description.

- **Instances:** the individual, independent examples of a concept
  - „Objects” of the same type.
  - **Note:** more complicated forms of input are possible.

- **Attributes:** (features) measuring aspects of an instance
  - Could be defined on various measurement scales – however we will focus mainly on nominal and numeric ones.
  - In statistics → variables.
Again how to get a data table?

- Instance: specific type of example
  - Thing to be classified, associated, or clustered
  - Individual, independent example of target concept
  - Characterized by a predetermined set of attributes
- Input to learning scheme: set of instances / data table
  - Represented as a single relation/flat file
- Rather restricted form of input
  - No relationships between objects
- Most common form in practical data mining
- Process of flattening a file “denormalization”
  - Several relations are joined together to make one
- Possible with any finite set of finite relations
The weather problem (Quinlan’s play sport)

<table>
<thead>
<tr>
<th>Outlook</th>
<th>Temperature</th>
<th>Humidity</th>
<th>Windy</th>
<th>Play</th>
</tr>
</thead>
<tbody>
<tr>
<td>sunny</td>
<td>85</td>
<td>85</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>sunny</td>
<td>80</td>
<td>90</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>overcast</td>
<td>83</td>
<td>86</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>70</td>
<td>96</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>68</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>65</td>
<td>70</td>
<td>true</td>
<td>no</td>
</tr>
<tr>
<td>overcast</td>
<td>64</td>
<td>65</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>sunny</td>
<td>72</td>
<td>95</td>
<td>false</td>
<td>no</td>
</tr>
<tr>
<td>sunny</td>
<td>69</td>
<td>70</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>75</td>
<td>80</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>sunny</td>
<td>75</td>
<td>70</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>overcast</td>
<td>72</td>
<td>90</td>
<td>true</td>
<td>yes</td>
</tr>
<tr>
<td>overcast</td>
<td>81</td>
<td>75</td>
<td>false</td>
<td>yes</td>
</tr>
<tr>
<td>rainy</td>
<td>71</td>
<td>91</td>
<td>true</td>
<td>no</td>
</tr>
</tbody>
</table>

Given past data, Can you come up with the rules for Play/Not Play?

Classics in ML
Types of attributes

• The most common distinction comes from measurement scale and statistics:
  • Nominal (also binary)
  • Ordinal
  • Interval-scaled
  • Ratio-scaled.

• Other names:
  • Categorical vs. numeric/continuous ones.

• Other types:
  • Criteria (preference-ordered), hierarchical, …
Transforming text documents into a standard form

- Transformation into Vector Representation

The \( d=7 \) documents:
D1: Large Scale Singular Value Computations
D2: Software for the Sparse Singular Value Decomposition
D3: Introduction to Modern Information Retrieval
D4: Linear Algebra for Intelligent Information Retrieval
D5: Matrix Computations
D6: Singular Value Analysis of Cryptograms
D7: Automatic Information Organization

The \( t=5 \) terms:
T1: Information
T2: Singular
T3: Value
T4: Computations
T5: Retrieval

\[
A = \begin{pmatrix}
0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 1.00 \\
0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\
0.58 & 0.71 & 0.00 & 0.00 & 0.00 & 0.71 & 0.00 \\
0.58 & 0.00 & 0.00 & 0.00 & 1.00 & 0.00 & 0.00 \\
0.00 & 0.00 & 0.71 & 0.71 & 0.00 & 0.00 & 0.00 \\
\end{pmatrix}
\]
% ARFF file for weather data with some numeric features
%
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature numeric
@attribute humidity numeric
@attribute windy {true, false}
@attribute play? {yes, no}

@data
sunny, 85, 85, false, no
sunny, 80, 90, true, no
overcast, 83, 86, false, yes
...
Attribute types in WEKA

• ARFF supports numeric and nominal attributes.
• Interpretation depends on learning scheme:
  • Numeric attributes are interpreted as
    • interval scales if less-than and greater-than are used,
    • ratio scales if distance calculations are performed (normalization/standardization may be required).
  • Instance-based schemes define distance between nominal values (0 if values are equal, 1 otherwise)
• Integers: nominal, ordinal, or ratio scale?
Any questions, remarks?