Data Mining I

Introduction and Course Organisation
Hallo

- Prof. Dr. Christian Bizer
- Professor for Information Systems V
- Research Interests:
  - Data and Web Mining
  - Web Data Integration
  - Linked Data Technologies
- Room: B6 - B1.15
- eMail: chris@informatik.uni-mannheim.de
- Consultation: Wednesday, 13:30-14:30

- I will teach the lecture introducing the principle methods of data mining.
Hallo

- **Kiril Gashteovski**
- Graduate Research Associate
- Research Interests:
  - Text Analytics
  - Machine Learning
  - Open Information Extraction
- Address: Parkring 39, Room 107
- eMail: k.gashteovski@uni-mannheim.de

Kiril will teach the exercise group at 10:15 and co-supervise the student projects.
Hallo

- Daniel Ruffinelli
- Graduate Research Associate
- Research Interests:
  - Data Mining
  - Ontologies and Reasoning
  - Optimization
- Room: B6, 26, C 1.11
- eMail: daniel@informatik.uni-mannheim.de
- Daniel will teach the exercise group at 12:00 and co-supervise the student projects.
Hallo

- Oliver Lehmberg
- Graduate Research Associate
- Research Interests:
  - Data and Web Mining
  - Network Analysis
  - Web Data Integration
- Room: B6, 26, C 1.04
- eMail: oli@informatik.uni-mannheim.de
- Oliver will teach the exercise group at 13:45 and co-supervise the student projects.
Outline of Today‘s Lecture

1. Introduction to Data Mining
   1. What is Data Mining?
   2. Methods and Applications
   3. The Data Mining Process

2. Course Organisation
1. Introduction to Data Mining
The Data Deluge

More and more data is generated:

- Transaction data from banking, telecommunication, e-commerce
- Scientific data from astronomy, physics, biology
- The public Web, Twitter, the Blogosphere
- Social network sites
- Application logs
“We are Drowning in Data...”

Wikipedia = Reference Size
≈ 5.9 TB of data
(Jan. 2010 Dump)

Source: The following slides are taken from Aidan Hogan's course on “Massive Data Processing”
“We are Drowning in Data...”

Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year
≈ 12 Wikipedias/year

Analyze
• Type of sky object:
  Star or galaxy?
“We are Drowning in Data…”

US Library of Congress
≈ 235 TB archived
≈ 40 Wikipedias

Analyze
• Topic distributions
• Citation networks
• Historic trends*

“We are Drowning in Data...”

Facebook

≈ 12 TB/day added
(as of Mar. 2010)
≈ 2 Wikipedias/day
≈ 782 Wikipedias/year

Analyze

• Current interests and behavior of over one billion people.
“We are Drowning in Data…”

Google
~ 20 PB/day processed (Jan. 2010)
~ 3,389 Wikipedias/day
~ 7,300,000 Wikipedias/year

Analyze
• Browsing behavior and interests of users.
“We are Drowning in Data…”

Analyze
- Current behavior and interests of mankind.
“We are Drowning in Data...”

NSA
Unknown amount of communication data from all over the world.

Analyze
• Identify suspects and terrorists.
“...but starving for knowledge!”

← Amount of data that is produced.

← Amount of data that can be looked at by humans.
Necessity of Data Mining

- Data is a valuable resource for companies.
- **But:** Raw data is mostly useless.
- The valuable information is “hidden” in the raw data.
- Data Mining methods are needed in many cases to
  • make sense of data.
  • take business decisions based on data.
1.1 What is Data Mining?

- 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 needed? Methods that
  1. detect patterns and regularities in data
  2. support business decisions based on data patterns
What is Data Mining?

- More informal definition

„Data mining is nothing else than torturing the data until it confesses.“

(Fred Menger)

- ... and if you torture it enough, you can get it to confess to anything.
- Thus, we need to be critical and question the pattern that we find.
Origins of Data Mining

- Draws ideas from machine learning, statistics, and database systems.
- Traditional techniques may be unsuitable due to
  - large amount of data
  - high dimensionality of data
  - heterogeneous and distributed nature of data
Data Mining Application Fields

- Business
  - Customer relationship management, marketing, fraud detection, manufacturing, telecom, health care, …

- Science
  - Data mining helps scientists to formulate hypotheses.
  - Astronomy, physics, drug discovery, social sciences, …

- Web and Social Media
  - Advertising, search engine optimization, spam detection, web site optimization, sentiment analysis, …

- Government
  - Surveillance, crime detection, finding tax cheaters, …
Hype Topic: Big Data

- Today, everybody can mine large amounts of data at low costs in the cloud.

- Technical Realization
  - massive parallelization using hundreds or thousands of machines
  - using tools like Hadoop, Spark, Mahout, Hbase

- Open Data Catalogs
  - TheDataHub: 9,100 data sets
  - publicdata.eu: 47,000 data sets

- Conference
  - O‘Reilly STRATA Conference
  - http://strataconf.com/public/content/home
LinkedIn analyzed the skills of members who started new jobs or received interest from recruiters in 2014.

1. Statistical Analysis and Data Mining
2. Middleware and Integration Software
3. Storage Systems and Management
4. Network and Information Security
5. SEO/SEM Marketing
6. Business Intelligence
7. Mobile Development
8. Web Architecture and Development Framework
9. Algorithm Design
10. Perl/Python/Ruby

1.2 Methods and Applications
Data Mining Methods

– **Descriptive Methods**
  - Goal: Find patterns in the data.
  - Example: *Which products are often bought together?*

– **Predictive Methods**
  - Goal: Predict unknown values of a variable
    - given observations (e.g., from the past)
  - Example: *Will a person click an online advertisement?*
    - given her browsing history

– **Machine Learning Terminology**
  - descriptive = unsupervised
  - predictive = supervised
Data Mining Tasks

1. Clustering [Descriptive]
2. Classification [Predictive]
3. Association Rule Discovery [Descriptive]
4. Regression [Predictive, Data Mining II]
5. Time Series Prediction [Predictive, Data Mining II]
6. Sequential Pattern Discovery [Descriptive, Data Mining II]
7. Anomaly Detection [Descriptive, Data Mining II]
1.2.1 Clustering: Definition

- Given a set of data points, each having a set of attributes, and a similarity measure among them, find clusters such that
  - data points in one cluster are more similar to one another
  - data points in separate clusters are less similar to one another

- Similarity Measures
  - Euclidean distance if attributes are continuous
  - Other problem-specific similarity measures

- Goals
  - Intracluster distances are minimized
  - Intercluster distances are maximized

- Result
  - A descriptive grouping of data points
Clustering: Application 1

- **Application area:** Market segmentation
- **Goal:** Divide a market into distinct subsets of customers
  - where any subset may be conceived as a marketing target to be reached with a distinct marketing mix
- **Approach:**
  1. Collect information about customers
  2. Find clusters of similar customers
  3. Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters
Clustering: Application 2

- Application area: Document Clustering
- Goal: To find groups of documents that are similar to each other based on important terms appearing in them.
- Approach
  1. Identify frequently occurring terms in each document.
  2. Form a similarity measure based on the frequencies of different terms.
- Application Example: Grouping of articles in Google News
1.2.2 Classification: Definition

- **Goal**: Previously unseen records should be assigned a class from a given set of classes as accurately as possible.

- **Approach**: Given a collection of records (*training set*)
  - each record contains a set of attributes
  - one of the attributes is the *class (label)* that should be predicted.

- Find a *model* for the class attribute as a function of the values of other attributes.
Classification: Example

• Training set:

"tree"  "tree"  "tree"

"not a tree"  "not a tree"  "not a tree"

• Learned model: "Trees are big, green plants"
Classification: Workflow

Class/Label Attribute

**Training Set**

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attrib1</th>
<th>Attrib2</th>
<th>Attrib3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Large</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Medium</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Small</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Medium</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Large</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Medium</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Large</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Small</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Medium</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Small</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

**Unseen Records**

<table>
<thead>
<tr>
<th>Tid</th>
<th>Attrib1</th>
<th>Attrib2</th>
<th>Attrib3</th>
<th>Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>No</td>
<td>Small</td>
<td>55K</td>
<td>?</td>
</tr>
<tr>
<td>12</td>
<td>Yes</td>
<td>Medium</td>
<td>80K</td>
<td>?</td>
</tr>
<tr>
<td>13</td>
<td>Yes</td>
<td>Large</td>
<td>110K</td>
<td>?</td>
</tr>
<tr>
<td>14</td>
<td>No</td>
<td>Small</td>
<td>95K</td>
<td>?</td>
</tr>
<tr>
<td>15</td>
<td>No</td>
<td>Large</td>
<td>67K</td>
<td>?</td>
</tr>
</tbody>
</table>

Learning algorithm

Induction

Learn Model

Apply Model

Deduction

Model
Classification: Application 1

- Application area: Fraud Detection
- Goal: Predict fraudulent cases in credit card transactions.
- Approach:
  1. Use credit card transactions and information about account-holders as attributes.
     - When and where does a customer buy? What does he buy?
     - How often he pays on time? etc.
  2. Label past transactions as fraud or fair transactions.
     This forms the class attribute.
  3. Learn a model for the class attribute from the transactions.
  4. Use this model to detect fraud by observing credit card transactions on an account.
Classification: Application 2

- Application area: Direct Marketing
- Goal: Reduce cost of mailing by targeting the set of consumers likely to buy a new cell-phone product.
- Approach:
  1. Use the data for a similar product introduced before.
     - We know which customers decided to buy and which decided otherwise.
     - This \{buy, don’t buy\} decision forms the \textit{class attribute}.
  2. Collect various demographic, lifestyle, and company-interaction related information about all such customers.
     - Age, profession, location, income, marriage status, etc.
  3. Use this information as input attributes to learn a classification model.
### 1.2.3 Regression

- Predict a value of a given **continuous valued variable** based on the values of other variables, assuming a linear or nonlinear model of dependency.

- Greatly studied in statistics and neural network field.

- **Examples:**
  - Predicting sales amounts of new product based on advertising expenditure.
  - Predicting wind velocities as a function of temperature, humidity, air pressure, etc.

- **Difference to classification:** The class attribute is continuous, while classification is used for nominal class attributes (e.g., *yes/no*).
1.2.4 Association Rule Discovery: Definition

- Given a set of records each of which contain some number of items from a given collection
- produce dependency rules which will predict occurrence of an item based on occurrences of other items

<table>
<thead>
<tr>
<th>TID</th>
<th>Items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread, Coke, Milk</td>
</tr>
<tr>
<td>2</td>
<td>Beer, Bread</td>
</tr>
<tr>
<td>3</td>
<td>Beer, Coke, Diaper, Milk</td>
</tr>
<tr>
<td>4</td>
<td>Beer, Bread, Diaper, Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke, Diaper, Milk</td>
</tr>
</tbody>
</table>

Rules Discovered

- \{Diaper, Milk\} \rightarrow \{Beer\}
- \{Milk\} \rightarrow \{Coke\}
Application area: Supermarket shelf management.
  - Goal: To identify items that are bought together by sufficiently many customers.
  - Approach: Process the point-of-sale data collected with barcode scanners to find dependencies among items.
  - A classic rule and its implications:
    • If a customer buys diapers and milk, then he is likely to buy beer as well.
    • So, don’t be surprised if you find six-packs stacked next to diapers!
    • Promote diapers to boost beer sales.
    • If selling diapers is discontinued, this will affect beer sales as well.

Application area: Sales Promotion
Application area: Advertising

Real example:
- Target (American grocery store)
- Analyzes customer buying behavior
- Sends personalized advertisements and coupons

Famous case in the USA:
- Teenage girl gets advertisement for baby products
- ... and her father is mad

Articles about this case
- http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html
- http://www.forbes.com/sites/kashmirhill/2012/02/16/how-target-figured-out-a-teen-girl-was-pregnant-before-her-father-did/
Association Rule Discovery: Application 3

- Application area:
  Inventory Management:

- Goal: A consumer appliance repair company wants to anticipate the nature of repairs on its consumer products and keep the service vehicles equipped with right parts to reduce on number of visits to consumer households.

- Approach: Process the data on tools and parts required in previous repairs at different consumer locations and discover the co-occurrence patterns.
1.2.5 Sequential Pattern Discovery: Definition

- Given a sequence of events (or sets of events)
- Find typical temporal patterns:
  - 1. (A,B)  2. (C)  3. (D,E)
  - 1. (A)  2. (B,C)  3. (D)
  - 1. (C)  2. (A,D)  3. (B)

- Typical pattern: Event C usually takes place before event D.
Sequential Pattern Mining: Applications 1

- Application area: Marketing
- Recurring customers
  - Book store example: (Twilight) (New Moon) → (Eclipse)

- Sequential patterns allow more fine grained suggestions than frequent pattern mining without sequence information

- Example:
  - mobile phone → charger vs. charger → mobile phone
    - are indistinguishable by frequent pattern mining
  - customers will buy a charger after a mobile phone
    - but not the other way around!
Sequential Pattern Mining: Applications 2

- Application area: Web usage mining
- Input
  - Web server logs
- Patterns
  - typical sequences of pages visited
- Goal: Improve structure and navigation of website
1.3. The Data Mining Process

Source: Fayyad et al. (1996)
1.3.1 Selection and Exploration

- Selection
  - What data is available?
  - What do I know about the provenance of this data?
  - What do I know about the quality of the data?

- Exploration
  - Get an initial understanding of the data
  - Calculate basic summarization statistics
  - Visualize the data
  - Identify data problems such as outliers, missing values, duplicate records
1.3.2 Preprocessing and Transformation

- Transform data into a representation that is suitable for the chosen data mining methods
  - number of dimensions
  - scales of attributes (nominal, ordinal, numeric)
  - amount of data (determines hardware requirements)

- Methods
  - integrate data from multiple sources
  - aggregation, sampling
  - dimensionality reduction / feature subset selection
  - attribute transformation / text to term vector
  - discretization and binarization

- Good data preparation is key to producing valid and reliable models.

- Data integration and preparation is estimated to take **70-80%** of the time and effort of a data mining project!
1.3.3 Data Mining

- Input: Preprocessed Data
- Output: Model / Patterns

1. Apply data mining method.
2. Evaluate resulting model / patterns.
3. Iterate
   - Experiment with different parameter settings.
   - Experiment with multiple alternative methods.
   - Improve preprocessing and feature generation.
   - Combine different methods.
1.3.4 Deployment

- Use model in the business context.

CRISP-DM Process Model
How Do Data Scientists Spend Their Days?

Literature Reference for this Chapter


Chapter 1: Introduction

Chapter 2: Data
2. Course Organisation

- Lecture
  - introduces the principle methods of data mining
  - discusses how to evaluate generated models
  - presents practical examples of data mining applications from the corporate and Web context.

- Three alternative Exercise Groups
  - students experiment with data sets using RapidMiner
  - you need to register for one of them

- Project Work
  - teams of five students realize a data mining project
  - teams may choose their own data sets and tasks (in addition, I will propose some suitable data sets and tasks)
  - teams write a summary about their project and present the project results

- Grading
  - 60% written exam, 30% project report, 10% presentation of project results
Course Organisation

- Course Webpage
  - provides up-to-date information, lecture slides, and exercise material

- Solutions to the Exercises
  - ILIAS eLearning System, https://ilias.uni-mannheim.de/

- Time and Location
  - Lecture: Wednesday, 10.15 - 11.45
    Room A5, B144
  - Exercise:
    - Thursday, 10.15 - 11.45
      Room B6, A104 (Kiril Gashteovski)
    - Thursday, 12.00 - 13.30,
      Room B6, A104 (Daniel Ruffinelli)
    - Thursday, 13.45 - 15.15,
      Room B6, A104 (Oliver Lehmberg)
Waiting List

- We currently have 15 people on the waiting list
  - if you decide not to attend, please leave ILIAS group this week
  - so others can have your place

- Policy: Two strikes out
  - you have to attend this lecture or the lecture next Wednesday (22.02.2017)
  - If you are not attending, you will be deleted from the participants list.
  - Please sign now next to your name on the participants list.

- If you are on the waiting list
  - you may get assigned a place next week
  - The waiting list is cleared after this semester (i.e., no priority for next semester!)
<table>
<thead>
<tr>
<th>Lectures Contents</th>
</tr>
</thead>
</table>
| **1. Introduction to Data Mining** | What is Data Mining?  
| | Methods and Applications  
| | The Data Mining Process  
| **2. Clustering** | K-means Clustering, Density-based Clustering, Hierarchical Clustering, Proximity Measures  
| **3. Classification** | Nearest Neighbor, Decision Trees  
| | Model Evaluation, Rule Learning, Naïve Bayes, Neural Networks, Support Vector Machines  
| **4. Association Analysis** | Frequent Item Set Generation, Rule Generation  
| | Interestingness Measures  
| **5. Text Mining** | Preprocessing Text, Feature Generation, Feature Selection, RapidMiner Text Extension  
| **6. Introduction to Student Projects** | Requirements and Organization  
<p>| | Overview of proposed data sets and tasks  |</p>
<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.02.2017</td>
<td>Introduction to Data Mining</td>
<td>Introduction to RapidMiner</td>
</tr>
<tr>
<td>22.02.2017</td>
<td>Lecture Clustering</td>
<td>Exercise Clustering</td>
</tr>
<tr>
<td>01.03.2017</td>
<td>Lecture Classification 1</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>08.03.2017</td>
<td>Lecture Classification 2</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>15.03.2017</td>
<td>Lecture Classification 3</td>
<td>Exercise Classification</td>
</tr>
<tr>
<td>22.03.2017</td>
<td>Lecture Text Mining</td>
<td>Exercise Text Mining</td>
</tr>
<tr>
<td>29.03.2017</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td>05.04.2017</td>
<td>Lecture Association Analysis</td>
<td>Exercise Association Analysis + Feedback Student Projects</td>
</tr>
<tr>
<td>- Easter break -</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26.04.2017</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>03.05.2017</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>10.05.2017</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>17.05.2017</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>24.05.2017</td>
<td>Submission of project results</td>
<td>- Holiday -</td>
</tr>
<tr>
<td>31.05.2017</td>
<td>Presentation of project results</td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Deadlines

- Submission of project proposal
  - Monday, April 3rd, 23:59

- Submission of final project report
  - Friday, May 26th, 23:59

- Project presentations
  - week of May 31st
  - everyone has to attend the presentations

- Final written exam
  - date will be announced
   - main reference book for the course!
   - 10 copies in university library.
   - we provide scans of important chapters via ILIAS

2. Vijay Kotu, Bala Deshpande: *Predictive Analytics and Data Mining: Concepts and Practice with RapidMiner*. Morgan Kaufmann.
   - several copies in university library.
   - electronic edition available via the library

   - several copies in university library
   - electronic edition available via the library

4. Website: KDnuggets
   - Overview of tools, online courses, events
   - http://www.kdnuggets.com/
Software

- Powerful data mining suite
- We are using Version 7 in the exercise
Gartner 2016 Magic Quadrant for Advanced Analytics Platforms
   - Explains along case studies how to use simple and advanced Rapidminer features.
   - Online access via the university library.
   - Website with data and processes: http://rapidminerbook.com

   - Free PDF version available online.

3. **Rapidminer – Documentation**
   - Introduction to user interface and basic features
   - http://docs.rapidminer.com
Lecture Videos and Screencasts

1. Video recordings of all lectures from FSS 2015
2. Step-by-step introduction to relevant RapidMiner features

http://dws.informatik.uni-mannheim.de/en/teaching/lecture-videos/
Questions?