Data Mining I
Introduction and Course Outline

Heiko Paulheim
Hello

- Prof. Dr. Heiko Paulheim
- Assistant Professor
- Research Interests:
  - Semantic Web and Linked Open Data
  - Data Mining with Linked Open Data
  - Ontology Matching
  - Data Quality and Data Cleaning
  - Outlier Detection
- Room: B6 – B1.16
- Consultation: by appointment
- Heiko will teach the lectures
Hello

• M.Sc. Nicolas Heist
• Graduate Research Associate
• Research Interests:
  – Semantic Web Technologies
  – Knowledge Graphs and Linked Data
• eMail: nico@informatik.uni-mannheim.de
• Nico will teach the *Python* exercises and co-supervise the team projects.
Hello

• M.Sc. Wi.-Inf. Oliver Lehmberg
• Graduate Research Associate
• Research Interests:
  – Data and Web Mining
  – Network Analysis
  – Web Data Integration
• eMail: oli@informatik.uni-mannheim.de
• Oliver will teach the RapidMiner exercises and co-supervise the team projects.
Hello

• M.Sc. Kiril Gashteovski
• Graduate Research Associate
• Research Interests:
  – Data Mining
  – Pattern Extraction
  – NLP
  – Knowledge Extraction
• eMail: k.gashteovski@uni-mannheim.de
• Kiril will teach the *Python* exercises and co-supervise the team projects.
Introduction and Course Outline

• Course Outline and Organization
• What is Data Mining?
• Methods and Applications
• The Data Mining Process
Course Organization

• 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

• Exercise
  – students experiment with data sets using RapidMiner or Python

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

• Final grade
  – 60 % written exam
  – 40 % project work

If you fail the exam, but do a good project, you pass.
Exercises of Your Choice

• Exercises in Python
  – Thursday, 12 – 13.30 and 15.30 – 17.00
  – Requires basic programming knowledge

• Exercise in RapidMiner
  – Thursday, 13.45 – 15.15
  – Require no programming knowledge

• All exercises start next week!
## Contents and Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>03.09.</td>
<td>Introduction/Course Outline</td>
<td>–</td>
</tr>
<tr>
<td>10.09.</td>
<td>Lecture: Clustering</td>
<td>Exercise: Clustering</td>
</tr>
<tr>
<td>17.09.</td>
<td>Lecture: Classification 1</td>
<td>Exercise: Classification</td>
</tr>
<tr>
<td>24.09.</td>
<td>Lecture: Classification 2</td>
<td>Exercise: Classification</td>
</tr>
<tr>
<td>01.10.</td>
<td>Holiday</td>
<td>–</td>
</tr>
<tr>
<td>08.10.</td>
<td><em>Introduction to Student Projects</em></td>
<td>Work on Project Outline</td>
</tr>
<tr>
<td>15.10.</td>
<td>Lecture: Classification 3</td>
<td>Exercise: Classification 3</td>
</tr>
<tr>
<td>22.10.</td>
<td>Lecture: Text Mining</td>
<td>Exercise: Text Mining</td>
</tr>
<tr>
<td>29.10.</td>
<td>Lecture: Association Rule Mining</td>
<td>Exercise: Association Rule Mining</td>
</tr>
<tr>
<td>05.11.</td>
<td><em>Project work</em></td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>12.11.</td>
<td><em>Project work</em></td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>19.11.</td>
<td><em>Project work</em></td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>26.11.</td>
<td>Submission of Project Results</td>
<td>–</td>
</tr>
<tr>
<td>03.12.</td>
<td><em>Presentation of project results</em></td>
<td>Presentation of project results</td>
</tr>
</tbody>
</table>
Deadlines

• Submission of project work proposal
  – Sunday, October 14\textsuperscript{th}, 23:59

• Submission of final project work report
  – Sunday, December 2\textsuperscript{nd}, 23:59

• Project presentations
  – schedule to be announced
  – everyone has to attend
Course Organization

• Lecture Webpage: Slides, Announcements
  – hint: look at version tags!

• Additional Material

• Time and Location
  – Lecture: Wednesday, 12.00 – 13.30, A5, C0.14
  – Exercises: Thursday, 12.00 – 13.30 (Python), 13.45 – 15.15 (RapidMiner), 15.30 – 17.00 (Python)
    B6 23-25, A1.04
  • these are three parallel groups, you only have to attend one
Course Organization

• Registration
  – you have registered via Portal2
  – ILIAS group will be opened soon

• There is a waiting list
  – if you decide not to continue, please email Ms. Czanderle
  – we will reassign your place
# Lecture Contents

<table>
<thead>
<tr>
<th>Introduction to Data Mining</th>
<th>What is Data Mining? Methods and Applications The Data Mining Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clustering</td>
<td>K-means Clustering, Hierarchical Clustering, Proximity Measures</td>
</tr>
<tr>
<td>Classification</td>
<td>Nearest Neighbor, Decision Trees, Rule Learning, Model Evaluation, Naïve Bayes, Support Vector Machines</td>
</tr>
<tr>
<td>Association Analysis</td>
<td>Frequent Item Set Generation, Rule Generation, Interestingness Measures</td>
</tr>
<tr>
<td>Text Mining</td>
<td>Preprocessing Text, Feature Creation, Feature Selection, RapidMiner Text Extension</td>
</tr>
<tr>
<td>Introduction to student projects</td>
<td>Overview of provided data sets Overview of proposed tasks</td>
</tr>
</tbody>
</table>
Literature & Slide Sources

• Pang-Ning Tan, Michael Steinbach, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.
  – 10 copies in university library.
  – we provide scans of important chapters via ILIAS

  – several copies in university library
  – we provide scans of important chapters via ILIAS
Literature & Slide Sources

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

• Gregory Piatetsky-Shapiro, Gary Parker: KDNuggets Data Mining course:
  http://www.kdnuggets.com/data_mining_course/
   - Explains along case studies how to use simple and advanced Rapidminer features.
   - Website with data and processes: http://rapidminerbook.com

   - Free PDF version available online.

   - Introduction to user interface and basic features
Literature – Python

• McKinney: Python for Data Analysis
• Severance: Python for Everybody: Exploring Data in Python 3
• Coelho and Richert: Building Machine Learning Systems with Python – Free Online Access via university library

• Online Sources:
  – https://www.learnpython.org/
  – https://docs.python.org/3/tutorial/
Additional Material

- Video recordings from FSS 2015
Outlook: Data Mining II

- Taught every FSS
- Topics
  - Regression
  - Sequential Pattern Mining, Time Series Prediction
  - Anomaly Detection
  - Online Data Analysis
  - Advanced Data Preprocessing
- Practical project
  - The annual Data Mining Cup
  - Worldwide competition of student teams
  - Real-world data mining tasks
Questions?
A Bit of History

• *We are drowning in data, but starving for knowledge.*
  
  *(John Naisbitt, 1982)*

• *Computers have promised us a fountain of wisdom but delivered a flood of data.*

• *It has been estimated that the amount of information in the world doubles every 20 months.*
  
  *(Frawley, Piatetsky-Shapiro, Matheus, 1992)*
“We are Drowning in Data...”

More and more data is generated:

- Transaction data from banking, telecommunication, e-commerce
- Scientific data from astronomy, physics, biology
- The public Web, Twitter, ...
- Social network sites
- Application logs
Data, Information, Knowledge, and Wisdom

Gene Bellinger, Durval Castro and Anthony Mills. "Transforming Data to Wisdom."
A Historical Example

• Cholera disease
• From beginning of 19th century
• ~100,000 deaths per year
  – until today!
• For a long time, there was little knowledge
  – on ways of infection
  – on causes of the disease

http://fieldnotes.unicefusa.org/2008/09/newsnet_combating_cholera_1.html
A Historical Example

- August Heinrich Petermann
- 1822-1878
- Geographer and Cartographer
- Geographic maps as a means
  - to understand data
  - to gather knowledge
A Historical Example

- 1848 map of Cholera deaths in London
  - finding: Cholera is more likely in densely populated areas
  - where there is no functioning sewage system
  - conclusion: Cholera is transmitted through contaminated water

A More Recent Example: SARS

- SARS: Severe acute respiratory syndrome
- Outbreak: 2012 in Hong Kong

A More Recent Example: SARS

- Which paths do SARS infection take?
- Max Planck Institute for Dynamics and Self-Organization:
  - SARS infections follow major airline routes

http://www.mpg.de/483574/pressemitteilung20041014
A Very Recent Example: the NSA

- Communication data from all over the world
- Searching for suspects and terrorists

http://www.theguardian.com/world/2013/jul/31/nsa-top-secret-program-online-data
A Very Recent Example: the NSA

The NSA collects metadata from phone records, enabling it to identify terrorists without examining the calls’ contents. Amid millions of calls, patterns can emerge, as our hypothetical scenario below demonstrates.

1. The phone records of a known terrorist supporter in Saudi Arabia form a cluster of possible accomplices.
2. A call from the known terrorist supporter is made to a person of interest in the United States, a U.S. citizen.
3. The phone metadata from the person of interest in the United States forms a cluster of associates in California.
4. Phone records show one of the associates in the California cluster called someone in the Saudi Arabia cluster. The NSA alerts the FBI to the connection, enabling the agency to obtain a wiretap.

https://www.popularmechanics.com/military/a9465/nsa-data-mining-how-it-works-15910146/
“We are Drowning in Data...”

Wikipedia

\[ \approx 10 \text{ TB of data} \]

(May 2016 Dump)

1 Wiki = 1 Wikipedia

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

Human Genome

\[ \approx 4 \text{ GB/person} \]
\[ \approx 0.0004 \text{ Wiki/person} \]
\[ \approx 2.4M \text{ Wiki/humankind} \]
“We are Drowning in Data...”

US Library of Congress
≈235 TB archived
≈23.5 Wiki
“We are Drowning in Data...”

Sloan Digital Sky Survey
≈ 200 GB/day
≈ 73 TB/year
≈ 7.3 Wiki/year
“We are Drowning in Data...”

NASA Center for Climate Simulation
≈ 32 PB archived
≈ 3,200 Wiki
“We are Drowning in Data...”

Facebook
≈ 12 TB/day added
≈ 1.2 Wiki/day
≈ 438 Wiki/year
(as of Mar. 2010)
“We are Drowning in Data…”

Large Hadron Collider
≈15 PB/year
≈1,500 Wikipedias/year
We are Drowning in Data...

Google

≈ 20 PB/day processed
≈ 2,000 Wiki/day
≈ 730,000 Wiki/year

(Jan. 2010)
“We are Drowning in Data...”

Internet (2016)
≈1.3 ZB/year
≈130,000,000 Wiki/year
(2016 IP traffic; Cisco est.)
“We are Drowning in Data...”
...but starving for knowledge!

← Rate at which data are produced

← Rate at which data can be understood
manual interpretation is hardly feasible!
Data Mining: Definitions

- Idea: mountains of data
  - where knowledge is mined
Data Mining: Definitions

- Data Mining is a non-trivial process of identifying
  - valid
  - novel
  - potentially useful
  - ultimately understandable
  patterns in data.
  
  (Fayyad et al. 1996)

- Data mining is nothing else than torturing the data until it confesses
  (Fred Menger, year unknown)

- ...and if you torture it enough, you can get it to confess to anything.
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, distributed nature of data
Data Mining Application Fields

• Business
  – Customer relationship management, e-commerce, fraud detection, manufacturing, telecom, targeted marketing, health care, …

• Science
  – Data mining helps scientists to analyze data and to formulate hypotheses.
  – Astronomy, physics, bioinformatics, drug discovery, …

• Web and Social Media
  – advertising, search engine optimization, spam detection, web site optimization, personalization, sentiment analysis, …

• Government
  – surveillance, crime detection, profiling tax cheaters, …
A Hype Topic: Big Data

• Everybody can analyze large amounts of data at low costs in the cloud

• Technical realization
  – massive parallelization using hundreds or thousands of machines
  – using tools like Hadoop, Hive, Hbase, Mahout, ...

• Open Data and Data Market Places
  – infochimps: >14,000 data sets
  – CKAN DataHub: >5,000 data sets
Data Mining Methods

• Descriptive methods
  – find patterns in data
  – e.g., *which products are often bought together?*

• Predictive methods
  – predict unknown or future values of a variable
    • given observations (e.g., from the past)
  – e.g., *will a person click a banner?*
    • given his/her browsing history

• Machine learning terminology:
  – descriptive = unsupervised
  – predictive = supervised
Data Mining Tasks

- Clustering (descriptive)
- Classification (predictive)
- Association Rule Mining (descriptive)
- Text Mining (both descriptive and predictive)

- Covered in Data Mining 2
  - Regression (predictive)
  - Anomaly Detection (descriptive)
  - Sequential Pattern Mining (descriptive)
  - Time Series Prediction (predictive)
Clustering

• Given a set of data points, and a similarity measure among them, find clusters such that
  – Data points in one cluster are similar to one another
  – Data points in separate clusters are different from each other

• Result
  – A descriptive grouping of data points
Clustering: Applications

• Application area: Market segmentation
• Goal: Subdivide 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:
  – Collect information about customers
  – Find clusters of similar customers
  – Measure the clustering quality by observing buying patterns of customers in same cluster vs. those from different clusters
Clustering: Applications

- Application area: Document Clustering
- Goal: Find groups of documents that are similar to each other based on the important terms appearing in them
- Approach
  - Identify frequently occurring terms in each document
  - Define a similarity measure based on the frequencies of different terms
- Application Example: Grouping of stories in Google News
Classification

• 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 class attribute as a function of the values of other attributes

• Goal: previously unseen records should be assigned a class as accurately as possible
  – A test set is used to validate the accuracy of the model
  – Training set may be split into training and validation data
Classification Example

Class/Label Attribute

<table>
<thead>
<tr>
<th>Tid</th>
<th>Refund</th>
<th>Marital Status</th>
<th>Taxable Income</th>
<th>Cheat</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Yes</td>
<td>Single</td>
<td>125K</td>
<td>No</td>
</tr>
<tr>
<td>2</td>
<td>No</td>
<td>Married</td>
<td>100K</td>
<td>No</td>
</tr>
<tr>
<td>3</td>
<td>No</td>
<td>Single</td>
<td>70K</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Yes</td>
<td>Married</td>
<td>120K</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>No</td>
<td>Divorced</td>
<td>95K</td>
<td>Yes</td>
</tr>
<tr>
<td>6</td>
<td>No</td>
<td>Married</td>
<td>60K</td>
<td>No</td>
</tr>
<tr>
<td>7</td>
<td>Yes</td>
<td>Divorced</td>
<td>220K</td>
<td>No</td>
</tr>
<tr>
<td>8</td>
<td>No</td>
<td>Single</td>
<td>85K</td>
<td>Yes</td>
</tr>
<tr>
<td>9</td>
<td>No</td>
<td>Married</td>
<td>75K</td>
<td>No</td>
</tr>
<tr>
<td>10</td>
<td>No</td>
<td>Single</td>
<td>90K</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Refund | Marital Status | Taxable Income | Cheat
--- | --- | --- | ---
No | Single | 75K | ?
Yes | Married | 50K | ?
No | Married | 150K | ?
Yes | Divorced | 90K | ?
No | Single | 40K | ?
No | Married | 80K | ?

Unseen Data

Training Set → Learn Classifier → Model
Classification: Applications

- Application area: Direct Marketing
- Goal: Reduce cost of mailing by targeting a set of consumers which are likely to buy a new cell phone
- Approach:
  - Use the data for a similar product introduced before
  - We know which customers decided to buy and which did not
  - Collect various demographic, lifestyle, and company-interaction related information about all such customers
    - Type of business, where they stay, how much they earn, etc.
  - Use this information as input attributes to learn a classifier model
Classification: Applications

• Application area: Fraud Detection
• Goal: Recognize fraudulent cases in credit card transactions
• Approach:
  – Use credit card transactions and the information on its account-holder as attributes
    • When and where does a customer buy? What does he buy?
    • How often he pays on time? etc.
  – Label past transactions as *fraud* or *fair* transactions
    This forms the *class attribute*
  – Learn a model for the class of the transaction
  – Use this model to detect fraud by observing credit card transactions on an account
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, Di</td>
</tr>
</tbody>
</table>

Rules Discovered

{Diaper, Milk} → {Beer}
{Milk} → {Coke}
Association Rule Discovery: Applications

• Application area: Marketing and Sales Promotion
• Example rule discovered:
  \{Bagels, Coke\} \rightarrow \{Potato Chips\}
• Insights:
  – promote bagels to boost potato chips sales
  – if selling bagels is discontinued, this will affect potato chips sales
  – coke should be sold together with bagels to boost potato chips sales
Association Rule Discovery: Applications

• Customers who bought this product also bought…
  – ...do terrorists order bomb building parts on Amazon?

Association Rule Discovery: Applications

• Content-based recommendation
  – requirement: much data
  – e.g., Amazon transactions, Spotify logfiles
Association Rule Discovery: Applications

- Real world example:
  - Customer loyalty programs

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

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

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: Applications

• Bottom line of the Target teenage girl story:
  – Janet Vertesi, Princeton university
  – Tried to hide her pregnancy from computers

• Measures taken:
  – using Tor for online surfing
  – no social media posts about her pregnancy
  – paying all pregnancy/baby related products in cash
  – a fresh Amazon account delivering to a local locker
    • paying with cash-paid gift cards

• Outcome:
  – massive buying of gift cards in a convenience store
    was reported to tax authorities

read the full story at
The Data Mining Process

Source: Fayyad et al. (1996)
The Data Mining Process

• Note that none of those steps actually requires a computer
• Recall Petermann's Cholera maps
  – Data Selection: find data on cholera deaths
  – Data Preprocessing: organize data by geographic area
  – Transformation: draw data on a map
  – Data Mining: look at the map and find patterns
    • possibly step back: add more data (population, water system, ...)
  – Interpretation: Cholera is transmitted via contaminated water

• However, computers make things easier
  – mainly: scalability
  – large data sets
  – large number of possible patterns
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
Selection and Exploration

- Visual Data Mining
  - For example as maps
  - Map showing the most popular photo locations in the world, generated from Panoramio logs

http://www.sightsmap.com/
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
  - 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 preparation estimated to take 70-80% of the time and effort of a data mining project!
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 different alternative methods
   – Improve preprocessing and feature generation
   – Combine different methods
Interpretation / Evaluation

• Output of Data Mining
  – Patterns
  – Models

• In the end, we want to derive value from that, e.g.,
  – gain knowledge
  – make better decisions
  – increase revenue
What you will learn in this lecture

• Common data mining tasks
  – How they work
  – When and how to apply them
  – How to interpret their output
Questions?