Data Mining I

Student Projects
Outline

1. Some Advertisements for Competitions
2. Requirements for Student Projects
3. Requirements for Project Summaries
4. Final Exam
ENSAE ParisTech and ParisTech invite all data science students from Universities all around the world to participate in the 1st edition of the “Data Science Game”. By solving a data driven issue, students will be able to enlighten their data science expertise in a both competitive and friendly spirit.

Data Science Game

A joint Initiative of ParisTech and ENSAE ParisTech.

The competition is supported by two major partners: Google Inc., who will provide the scope and material of the competition, and Capgemini, who will provide an amazing setting for the competition in Paris.

A two-phase competition:

- An online non-eliminatory phase from mid-May to mid-June 2015.
- A two-day competition in Paris, the 20th and 21st of June 2015.
Goal:
Participants combine technology and innovative ideas to design prototypes of digital products that tackle today's educational challenges.
Student Projects

- **Goals**
  - Gain more practical experience with data mining
  - Get to know additional problem-specific
    - preprocessing methods
    - data mining methods

- **Expectation**
  - Select an interesting data mining problem of your choice
  - Solve the problem using
    - the data mining methods that we have learned so far plus some advanced problem-specific data pre-processing
    - other data mining methods which might be helpful for solving the problem and build on what we learned in class
Procedure

- Teams of four students
  1. realize a data mining project
  2. write a 12 page summary of the project and the methods employed in the project
  3. present the project results to the other students (12 minutes presentation + 5 minutes discussion)

- Final mark for the course
  - 30 % written summary about the project
  - 20 % project presentation
  - 50 % written exam
## Schedule

<table>
<thead>
<tr>
<th>Week</th>
<th>Wednesday</th>
<th>Thursday</th>
</tr>
</thead>
<tbody>
<tr>
<td>15.04.2015</td>
<td>Introduction to Student Projects</td>
<td>Preparation of Project Outline</td>
</tr>
<tr>
<td></td>
<td><strong>Sunday, 19.04.2015, 23:59:</strong> Submission of Project Outlines</td>
<td></td>
</tr>
<tr>
<td>22.04.2015</td>
<td>Feedback Project Outlines</td>
<td>Feedback Project Outlines</td>
</tr>
<tr>
<td>29.04.2015</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>06.05.2015</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td>13.05.2015</td>
<td>Project Work</td>
<td>- Public holiday -</td>
</tr>
<tr>
<td>20.05.2015</td>
<td>Project Work</td>
<td>Feedback on demand</td>
</tr>
<tr>
<td></td>
<td><strong>Friday, 22.05.2015, 23:59:</strong> Submission of Project Reports</td>
<td></td>
</tr>
<tr>
<td>27.05.2015</td>
<td>Presentation of Project Results</td>
<td>Presentation of Project Results</td>
</tr>
</tbody>
</table>
Where to find interesting Data Sets?

- **The Data Hub**
  - [http://thedatahub.org/](http://thedatahub.org/)
  - Large European data set catalog
- **Data.gov.uk, Data.gov.us**
  - Public sector data provided by the UK and US governments
- **KDD Cup and Data Mining Cup**
  - Data mining competitions providing data sets and solutions
- **Kraggle**
  - Website running commercial and educational data science competitions
  - Offers datasets as well as solutions for older competitions
  - [https://www.kaggle.com/](https://www.kaggle.com/)
- **Programmable Web**
  - Website giving an overview about 13000 public Web APIs
Where to find Information about additional Methods?

1. Pang-Ning Tan, Michael Steinback, Vipin Kumar: Introduction to Data Mining, Pearson / Addison Wesley.


Where to find Information about additional Methods?

– Check out the solutions to your problem that other people have tried.
  • for instance by looking at submissions of the KDD CUP or Data Mining Cup as well as Kraggle
  • or search for relevant scientific papers using:

![Google Logo]
Some Project Ideas (not binding)

- **Web Log Mining**
  - Learn a classifier for the categorizing the visitors of your website.
  - Which features matter? Number of pages visited, time on site, .. (Bing Liu Chapter 12.x)
  - Preprocess some web log data outside RapidMiner
  - Learn and evaluate classifier within RapidMiner

- **Wikipedia Contributors**
  - Examine the edit history of Wikipedia contributors
  - Cluster users by different attributes (no of edits, edits/day, topic, ..)
  - Or learn a classifier for the categorizing Wikipedia contributors

- **Sentiment Analysis for Discussion Forum / Rating Site / Tweets**
  - Are people positive or negative about topic / product? (Bing Liu 11.x)

- **SPAM Detection**
  - eMail, blog or discussion forum (Bing Liu 6.10, 11.9)
Some Projects realized in previous Semesters

- **Mannheim Police Reports**
  - Learn classifiers for police reports
  - Identify type of incident, severity of incident, location of incident

- **Bundesliga Betting Rules**
  - Find rules that help you to predict the outcome of a bundesliga game

- **last.fm Playlist Analysis**
  - Cluster last.fm users according to the style of the songs they are listening to
  - Find commons sets of songs for the different clusters

- **Analysis of Training Data of a Fitness Center**
  - Find different customer groups by clustering exercise data

- **Sentiment Analysis of Tweets about Movies**
  - Learned classifier from IMDB movie reviews
  - Applied and tested with tweets afterwards

- **Classifying a Document’s Perspective**
  - using the example of Israeli – Palestinian Essays
Project Outlines

- 4 pages (exactly!) including title, using DWS master thesis layout
- due Sunday, 19.04.2015, 23:59
- send by eMail to Chris, Volha & Oliver
- answer the following questions:
  1. What is the problem you are solving?
  2. What data will you use?
     - Where will you get it?
     - How will you gather it?
  3. How will you solve the problem?
     - What preprocessing steps will be required?
     - Which algorithms you plan to use?
     - Be as specific as you can!
  4. How will you evaluate, measure success?
  5. What do you expect your results to look like?
Coaching Sessions

- We will give you tips and answer questions concerning your project.
- Registration via email is mandatory!
  - until Tuesday night!
  - including the questions that you like to discuss
  - including which session you prefer (Thursday B2/B3)
- We will assign you a time slot afterwards and inform you about the slot via email.
- Every team has to attend at least one coaching session!
Project Summaries

- 12 pages (exactly!) including title, TOC page and references ➔ max. 9 pages text/tables/diagrams. Each extra page and each day of late submission downgrades your mark by 0.3!
- due Friday, 22.05.2015, 23:59
- send by email to Chris, Volha & Oliver
- Outline for project summaries:
  1. Application area and goals
  2. Structure and size of the data set
  3. Preprocessing
  4. Actual Data Mining
    1. including evaluation results
    2. including different approaches that you tried
  5. Discussion of Results
- Requirements
  • You must use the DWS master thesis layout.
  • Please cite sources properly. Preferred citation style [Author, year].
  • Also submit your RapidMiner processes and (a subset) of your data.
Final Exam

- Date: 05.06.2015
- Duration: 60 minutes
- Structure: 5 - 6 open questions that
  - check whether you have understood the content of the lecture
  - require you to describe the ideas behind algorithms and methods
  - might require you to do some simple calculations