Lab 1 – Preprocessing

Data Mining, Spring 2016

IT UNIVERSITY OF COPENHAGEN

Welcome to the first lab of the course!

Hello from your two TAs!

- Anders Hartzen
  - 2010: B.Sc. Software Development, ITU
  - 2012: M.Sc. Games Technology, ITU
  - Working at ITU since as
    - Research and Teaching assistant
    - This semester also External Lecturer for Data Mining
  - Trying to get PhD project funded

- Kasra Tahmasebi
  - 2014: B.Sc. Software Development, ITU
  - Final semester of Games Technology.
  - Trying to defeat the Master's Thesis.

Data Mining Labs

- Most exercises during the course will focus on you implementing algorithms in a programming language of your choice.
  - We recommend Java, as you will be able to use code starter packages we will provide you in the different labs.
- We will be here to help you and sometimes Sebastian will be around as well to help you if needed.
- Doing the labs will help you do the groundwork for the individual mandatory assignment you have to hand in on Friday March 18th before 23:55 (Digital submission via learnIT).
- Use the Q&A Forum on the learnIT page to ask questions about anything about the course (labs, assignment etc.) and help each other out.
- Basic info:
  - Labs take place from 14:00-16:00 in rooms 4A54 and 4A58.
  - Labs are optional, but you will be expected to know the algorithms covered in labs really well at the exam!

Today's Lab: Preprocessing

Preprocessing – aka cleaning data

- Today you will be working with data from the questionnaire you filled out last week
- You will be cleaning up the data by using pre-processing techniques you get to implement yourself
- The data is the data dump from the online questionnaire containing the freetext (i.e. comments) participants wrote. No assumptions or corrections were made
- Therefore the data needs heavy cleaning and preprocessing to be more useful for further experiments

Overview of today's Lab

- Part 1 – Load the data set using code.
- Part 2 – Clean the data set using code.
- Part 3 – Normalize attributes using code.
- Part 4 – Use descriptive statistical methods to describe the data set using code.

Part 1 – Load Data

- Java code is available from the course page on learnIT to help you get started loading in the data.
- Is pretty basic, but works.
- Feel free to write your own code and/or use another programming language.
  - Though we are most able to help with Java/C# questions.

Part 2 – Clean the Data Set

- Using code!
- Issues worth considering:
- Missing values?
- Different formats?
- Noise? Outliers?
- Data transformation?

Part 3 – Normalization

- In your own code normalize the numerical values you find most interesting
  - Min-max
  - Z-score
  - Decimal scaling

Part 4 – Descriptive Statistics

- In your own code try to describe the data using descriptive statistics.
- Central tendency of the data
  - Mean
  - Median
  - Mode
  - Etc. (See pg. 45 in book for overview)
- Dispersion of the data
  - Standard deviation
  - Five-number summary
    - Min
    - Quartiles
    - Median
    - Max
  - Etc. (see pg. 48 in book)

Hidden truths? Large datasets

At the end of the lab think about the following:

- Did you find any meaningful correlations between parts of the data?
- Are there other methods I could have used to detect possible correlations?
- Would my preprocessing code work well if applied to a very large dataset?
  - Any changes I would make in my code?
  - Any new constructs I could add to my code to make it work with a large dataset?
    - Mixed-initiative via a GUI?
    - Output troublesome data via File-output?
    - ???

Thanks for listening!