Lab 1 – Lab Preprocessing

Data Mining, Spring 2017

Welcome to the first lab of the course!

- Daniel Sørensen
  - 2016: B.Sc. Software Development, ITU
  - Second semester of Games Technology

- Mathias Löwe
  - 2016: B.Sc. Software Development, ITU
  - Third semester of Software Development and Technology

- 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 Monday 17th April before 14:00 (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 12:00-14:00 in rooms 2A12-14.

  - Labs are optional, but you will be expected to know the algorithms covered in labs really well at the exam!

Today's Lab: Preprocessing

- 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 free text (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

- 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.

- Java code is available from the course page on LearnIT to help you get started loading in the data.

- It's pretty basic, but it works.

- Feel free to write your own code and/or use another programming language.

- Though we are most able to help with Java/C#/Python questions.

- Using code!

- Issues worth considering:

- Missing values?

- Different formats?

- Noise? Outliers?

- Data transformation?

- In your own code normalize the numerical values you find most interesting

  - Min-max

  - Z-score

  - Decimal scaling

- 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)

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 you 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?

- Experiment with different degrees of normalization of continuous attributes.

  - How does this affect the central tendency? Does this change your perception of the data?

Thanks for listening!