F20/21DL. Data Mining and Machine Learning

Lab 1. Introduction to Data Mining

Covering Practical work to be done by students in Week 1

The purpose of this lab sheet is:

1. to set you up for Python programming in this course
2. give you some practice with exploring data mining and machine learning data sets
3. to help you to start with your **DM & ML portfolio**, in groups.

1 Python Tutorial and Programming Practice

This part is for your individual programming practice during the week.

Go to the **Python Tutorials** section on **Canvas**. Cover the following steps:

- Make sure that Python and Jupyter Notebooks run on your computer or the lab computer that you will be using as your working machine during the term. Canvas contains the installation instructions, should you need them.
- Complete Python Tutorial P0, in case you are not yet familiar with Python
- Study a data set application: Classifying Iris Species (in Chapter 1 section 1.7 - Book: Introduction to Machine Learning with Python, Mueller & Guido, 2017). GitHub code for Chapter 1 is available here: `https://github.com/amueller/introduction_to_ml_with_python/blob/master/01-introduction.ipynb`.

2 Machine Learning and Data Mining Portfolio

This part is to be completed in groups.

2.1 Exploring data sets

Explore at least 3 data sets of your interest and describe features, samples and output variables. Use provided data descriptions.

You might find the following useful:

- There are some well-known benchmark data sets, such as Iris, MNIST, CIFAR10, CIFAR100 (an overview will be given in the lectures). You could use those.
- you can also search for data sets on `https://www.kaggle.com/` and UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets.php`).
- You can also use google dataset search at `https://datasetsearch.research.google.com/` . For example if you are interested in the problem of 'customer churn analysis', you could enter this keyword and you will receive several suggestions for available data sets

- If your data sets are in .csv format, run the Python Notebook dataset_explore.ipynb to help you understand the data characteristics.
- Look up the documentation of the Python functions below: read_csv(), head(), info() and describe()

2.2 What to Do during the lab:

- Meet your group.
- Explore 3 data sets from at least two sources. Discuss them among your groups, compare similarities and differences.
- Note down your observations. Write a brief description of the 3 data sets, and conclusions of all discussions taken during the lab. Have it ready to show to your tutor

Further recommendations:

- Each student is advised to keep an individual copy of the python notebook for further experimentation
- Group members can communicate virtually to complete the lab tasks.
- Groups are advised to upload a summary of their findings on your group space on canvas (this could be reviewed by your instructor/tutor)

2.3 Preview of next week Lab tasks

- Next week, you will be asked to replicate the contents of Python Tutorial P1 using a different data set.
- To start on the task, watch and run Python Tutorial P1, and start using your data set with the code
- Watch out for the next Lab handout with full instructions!