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, and will be assseed during the labs. Marking scheme: this lab will bring you up to 2 points. 1 point for completing the task, 1 additional point for any non-trivial analytical work with the material.

2.1 The data set

You are starting to work on your own Data Mining and Machine learning portfolio: therefore, the choice of the data set is yours! We suggest that you choose a domain that interests you most. For example, if you are interested in computer games, why not finding or generating your own data set that analyses the player behaviour! When choosing your favourite data set, you may want to take into consideration the following factors:

- Data sets that contain images will require you to learn some basics of image processing, and will be particularly good for working with Convolutional neural nets at the end of the term.
- Data sets that contain nominal (non-numeric) data will work more impressively with Bayes nets and Decision trees than computer vision data sets.

- There are some well-known benchmark data sets, such as Iris, MNIST, CIFAR10, CIFAR100 (an overview will be given in the lectures). We ask that you do not take benchmark data sets, as their properties are already well-explained on-line.
- When searching for data sets, `https://www.kaggle.com/` and UCI Machine Learning Repository (`https://archive.ics.uci.edu/ml/datasets.php`) are good sources.
- You can create your own data set, and this effort will score marks;
- Generally, more challenging data sets will be more likely to require more advanced DM&ML methods, and for this reason may bring higher marks.

Note: we recommend you use the same data set throughout the whole of your portfolio. So, invest a bit of time into this decision. Ask lab tutors if in doubt.

2.2 What to Do:

2.2.1 During the lab:

- Meet your group. Decide on a repository where you will write your DM&ML portfolio. Assign 1 group rep – a person who will be responsible for maintaining the access rights to the DL&ML portfolio and will be able to share sources with the course leader at the end of the term.
- Explore 3 data sets from at least two sources. Discuss them among your groups, compare similarities and differences.
- Choose one data set out of those 3, discuss reasons, pros and cons, for choosing it.
- Explain your findings to your designated lab tutor/instructor, they will be able to assign you marks for this work, and give advice as to the right choice of the data set for you.

2.3 After the lab:

- *Group rep:* Make sure all group members have access to the repository where they will collaborate
- *Everyone:* Write a breif description of the 3 data sets, and conclusions of all discussions taken during the lab, in your joint Jupyther Notebook.
- *Everyone:* Explain your reasons for choosing your data set. If you created the data set yourself, explain all technical difficulties and pitfalls in generating the data set.

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)

2.4 Preview of next week Lab tasks

- Next week, you will be asked to replicate the contents of Python Tutorial P1 using your chosen 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!