JBI030: Course Software

The course material can be found on Canvas in `.html` format, but jupyter notebooks with source code will also be available on Github. The use of the notebooks is encouraged, as they allow to actively modify the code and experiment with it. In this way, you are allowed to investigate the effect of modifiying pieces of code on final results. In order to download the jupyter notebooks, a Github account is required.

Github

After installing git and creating an account, you can download the course material from the terminal in this way:

```
git clone https://github.com/davidevdt/datamining_jbi030.git
```

Updates can be downloaded by locating the terminal within the folder of the repository and subsequently launching `git pull`. Course lectures will be added at the end of each lecture. Alternatively, the repository can be downloaded as a zip file from the Github page. Students are invited (but not obliged) to find online tutorials that help them to better understand git.

Python

Before installing all the packages needed for the course, you need to make sure that the `Python` environment is installed in your machine. For JBI030, we will use Python 3. Make sure that the the latest version of Python (3.8) is the one installed on your system. Alternatively, you can install Python 3 with the Anaconda distribution. If you are new to Python, you can find several online tutorials (such as this one) that can help you to get accustomed to it.

Python Packages

scikit-learn will be the main tool used in JBI030. `scikit-learn` is a Python library that allows fitting several Machine Learning models. As you can see from their Github page, `scikit-learn` depends on other libraries in order to work. Such libraries are `Numpy` (a package that eases numerical computing), `Scipy` (a scientific library), and `joblib` (a library that enables fast computations).

Other important libraries that will be needed for the course are:

- `matplotlib` for plotting
- `pandas` for dataset manipulatoin
- `graphviz` for graph rendering

Among all these libraries, of particular importance are `numpy`, `pandas`, and `matplotlib`. Although we will provide a small introductory lecture of the libraries, you are invited to find tutorials (e.g., the tutorials from the official pages) for a deeper understanding of this material.

The packages can be downloaded and installed directly from command line:

```
pip3 install numpy scipy matplotlib pandas graphviz joblib
scikit-learn
```

or, with the Anaconda environment,

```
conda install numpy scipy matplotlib pandas graphviz joblib
scikit-learn
```

Most of the theoretical and pratical part of the course will be based on the `scikit-learn` documentation. Code will be provided during the course, but you must refer to this website for a more detailed exploration of the considered functions, or to learn the ones not covered with the lecture notes.

Keras

For Neural Networks and Deep Learning, `Keras` offers optimal performance by exploiting GPU power. You can install it by following the instructions here. Note that `Keras` requires TensorFlow; you can install it by following the instructions here.

Jupyter Notebooks

Jupyter Notebooks will be used as coding environment. They are also the suggested tool for the report of your final project, as they allow integrating exectuable Python code with written documentation (with `Markdown`). You are invited to learn the Markdown style with some online guide.

Jupyter Notebooks can be installed via

```
pip3 install jupyterlab
```

in your terminal. Students can also easily find online notebooks tutorials.