Prerequisites

Course materials on GitHub

The course materials are available on GitHub, so that you always have the latest updates. To download them, you first need to install git (if you haven't already).

You can 'clone' the course as follows from the command line (you can also use a GUI)

```
git clone https://github.com/joaquinvanschoren/ML-intro.git
```

To download updates, run

```
git pull
```

For more details on using git, see the GitHub 10-minute tutorial. If you want to suggest improvements, you can! Just create a pull request and we'll review it.

Alternatively, you can download the course as a .zip file. Click 'Clone or download'. Download individual files with right-click -> Save Link As…

Python

You first need to set up a Python environment (if you do not have done so already). The easiest way to do this is by installing Anaconda. We will be using Python 3, so be sure to install the right version.

If you are completely new to Python, we recommend to take an introductory online course, such as the DataCamp Intro to Python for Data Science, or the Whirlwind Tour of Python.

Required packages

Next, you'll need to install several packages that we'll be using extensively. You'll need to run these commands on the command line.

Installing packages with conda

If you are using Anaconda, you can use the `conda` package manager to install all packages:

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

and then *also*

```
pip install graphviz
```

Installing packages with pip

With most other setups (not conda), you can use pip to install all packages. Pip is the Python Package index. It is included in most Python installations.

```
pip install numpy scipy scikit-learn matplotlib pandas pillow graphviz
```

Note: we'll be using scikit-learn 0.18, which is currently the latest version.

You also need to install the graphiz C-library:

- OS X: use homebrew: `brew install graphviz`
- Ubuntu/debian: use apt-get: `apt-get install graphviz`.

- Installing graphviz on Windows can be tricky and using conda / anaconda is recommended.

Note: if you are not using Anaconda, and you already have a custom Python environment set up, possibly using a different Python version, it may be wise to set up a virtual environment for this course so that it does not affect your existing environment.

Installing OpenML

OpenML is used to easily import datasets and share models and experiments. The OpenML package is not yet available through pip. We will need to install it from github, as well as the latest version of Python's ARFF reader.

With git installed, run the following:

```
pip install git+https://github.com/renatopp/liac-arff@master
pip install git+https://github.com/openml/openml-python.git@develop
```

You'll also need an OpenML account to download/upload data. If you don't have one, go ahead and create one.

Installing Jupyter notebooks

As our coding environment, we'll be using Jupyter notebooks. They interleave documentation (in markdown) with executable Python code, and they run in your browser. That means that you can easily edit and re-run all the code in this course.

If you use Anaconda, Jupyter is already installed. If you use pip, you can install it with

```
pip3 install jupyter
```

To test if it works, run

```
jupyter notebook
```

A browser window should open showing the files in your current directory. You can shut down the notebook by typing CTRL-C in your terminal.

If you are new to notebooks, take this quick tutorial, or this more detailed one. Optionally, for a more in-depth coverage, try the DataCamp tutorial.

Testing

To test if everything works, download the course materials and run Jupyter (ideally from the directory where you downloaded the course):

```
jupyter notebook
```

A browser window should open with all course materials. Open one of the chapters and check if you can execute all code by clicking Cell > Run all.

Alternative: Everware

In case you run into any issues, you can also run all materials in the cloud. This is a special (private beta) service provided by Everware with computing resources provided by the Yandex School of Data Analytics. It is not (yet) meant for large-scale use, so it may be slow.

You'll need a GitHub account to authenticate. If you dont, you can make one now. To spin up the service, just click here (it may take a few minutes to boot):
Start server