Introduction to Machine learning with scikit-learn

Instructor

- Andreas Mueller @amuellerml - Columbia University; Book: Introduction to Machine Learning with Python

---

This repository will contain the teaching material and other info associated with the "Introduction to Machine Learning with scikit-learn" course.

About the workshop

Machine learning has become an indispensable tool across many areas of research and commercial applications. From text-to-speech for your phone to detecting the Higgs boson, machine learning excells at extracting knowledge from large amounts of data. This talk will give a general introduction to machine learning, as well as introduce practical tools for you to apply machine learning in your research. We will focus on one particularly important subfield of machine learning, supervised learning. The goal of supervised learning is to "learn" a function that maps inputs x to an output y, by using a collection of training data consisting of input-output pairs. We will walk through formulating a problem as a supervised machine learning problem, creating the necessary training data and applying and evaluating a machine learning algorithm. This workshop should give you all the necessary background to start using machine learning yourself.

Prerequisites

This workshop assumes familiarity with Jupyter notebooks and basics of pandas, matplotlib and numpy.

Content

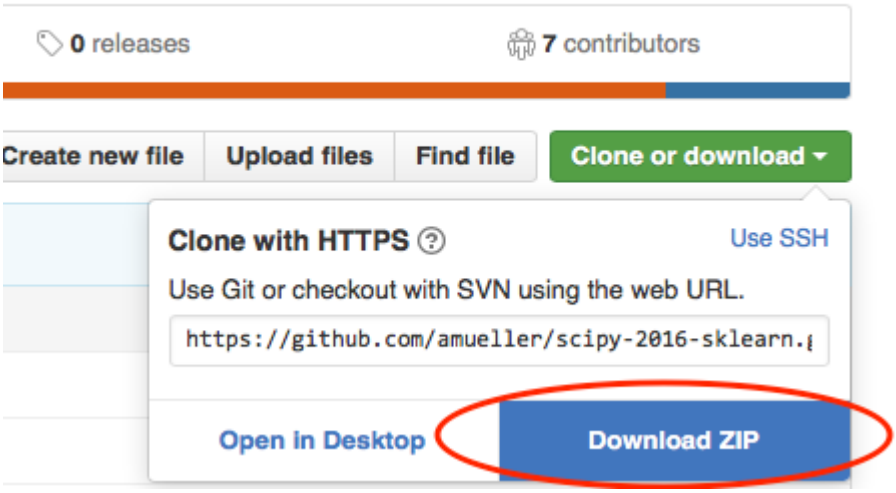
- Types of machine learning, basics of supervised learning
- Data loading with pandas
- Data requirements for scikit-learn
- Building classification and regression models
- The scikit-learn API
- Fundamentals of Data Preprocessing

Obtaining the Tutorial Material

If you are familiar with git, it is most convenient if you clone the GitHub repository. This is highly encouraged as it allows you to easily synchronize any changes to the material.

```
git clone https://github.com/amueller/ml-workshop-1-of-4.git
```

If you are not familiar with git, you can download the repository as a .zip file by heading over to the GitHub repository (https://github.com/amueller/ml-workshop-1-of-4) in your browser and click the green “Download” button in the upper right.



Please note that I may add and improve the material until shortly before the tutorial session, and we recommend you to update your copy of the materials one day before the tutorials. If you have an GitHub account and forked/cloned the repository via GitHub, you can sync your existing fork with via the following commands:

```
git pull origin master
```

Installation Notes

This tutorial will require recent installations of

- NumPy
- SciPy
- matplotlib
- pillow
- pandas
- scikit-learn (>=0.18.1)
- IPython
- Jupyter Notebook

The last one is important, you should be able to type:

```
jupyter notebook
```

in your terminal window and see the notebook panel load in your web browser. Try opening and running a notebook from the material to see check that it works.

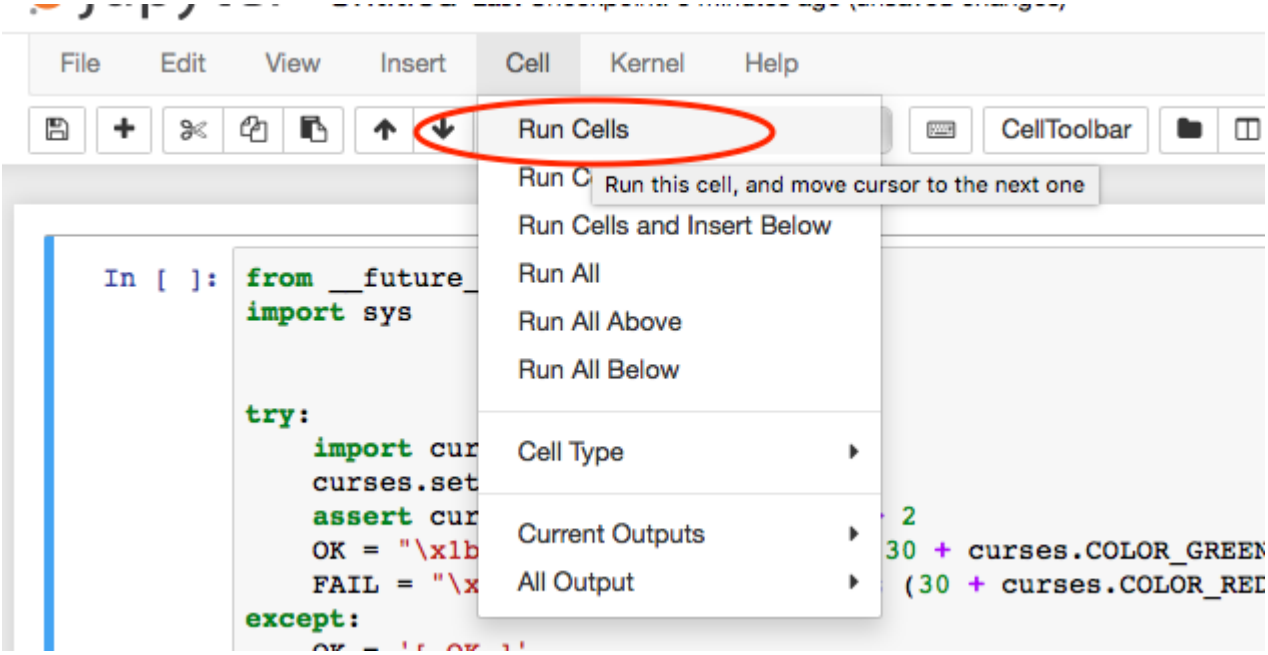
For users who do not yet have these packages installed, a relatively painless way to install all the requirements is to use a Python distribution such as Anaconda, which includes the most relevant Python packages for science, math, engineering, and data analysis; Anaconda can be downloaded and installed for

free including commercial use and redistribution. The code examples in this tutorial should be compatible to Python 2.7, Python 3.4 and later. However, it's recommended to use a recent Python version (like 3.5 or 3.6).

After obtaining the material, we **strongly recommend** you to open and execute a Jupyter Notebook `jupter notebook check_env.ipynb` that is located at the top level of this repository. Inside the repository, you can open the notebook by executing

```
jupyter notebook check_env.ipynb
```

inside this repository. Inside the Notebook, you can run the code cell by clicking on the "Run Cells" button as illustrated in the figure below:



Finally, if your environment satisfies the requirements for the tutorials, the executed code cell will produce an output message as shown below:

```
Using python in /Users/Sebastian/miniconda3
3.5.1 |Continuum Analytics, Inc.| (default, Jun 15 2016, 16:14:02)
[GCC 4.2.1 Compatible Apple LLVM 4.2 (clang-425.0.28)]

[ OK ] IPython version 4.2.0
[ OK ] numpy version 1.11.0
[ OK ] watermark version 1.3.1
[ OK ] matplotlib version 1.5.1
[ OK ] scipy version 0.17.1
[ OK ] yaml version 3.11
[ OK ] PIL version 1.1.7
[ OK ] sklearn version 0.17.1
[ OK ] pydot version 1.2.2
```