

Machine Learning with Python

Numpy / Matplotlib / Scikit-learn

Giovanni De Toni
giovanni.detoni@unitn.it

November 23, 2022

Setup

On lab machines



Download and extract the Scikit-learn lecture material from:

https://github.com/geektoni/ml_labs/archive/refs/heads/master.zip

Open the terminal in the folder containing the extracted files and run:

```
> ./jupyter - scikit . sh
```

Setup

On your own machine

Make sure you are using Python 3 for the following steps.

Install Numpy, Scipy, Matplotlib, Scikit-learn and Jupyter:

```
> pip install numpy scipy matplotlib sklearn pandas  
> pip install jupyter
```

Download and extract the material for the Scikit-learn lab:

https://disi.unitn.it/~passerini/teaching/2021-2022/MachineLearning_AIS/index.html

Open the terminal in the folder containing the extracted files and run:

```
> jupyter notebook
```

Setup

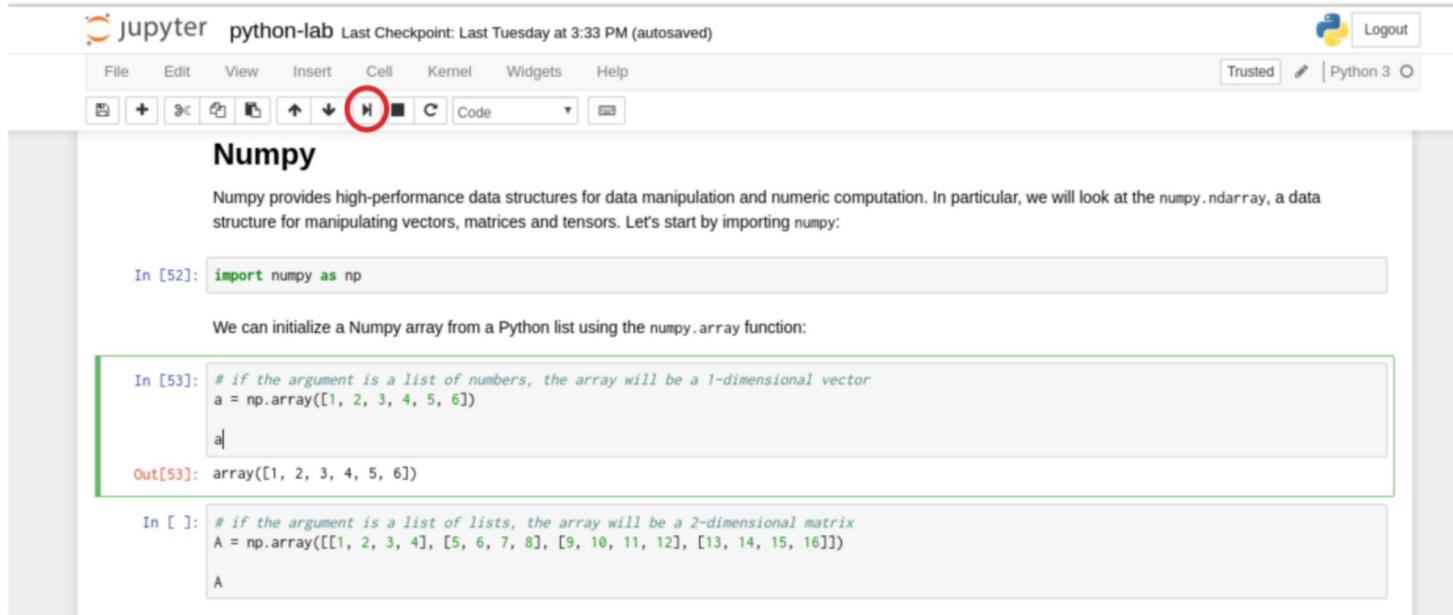
Jupyter notebook

Open the browser at the given address and you'll see something like:



Open the `sklearn-lab.ipynb` file containing the lecture notebook.

Jupyter notebook



The screenshot shows a Jupyter Notebook interface. At the top, the header bar includes the Jupyter logo, the text "python-lab", and a status message "Last Checkpoint: Last Tuesday at 3:33 PM (autosaved)". On the right side of the header, there is a "Logout" button. Below the header is a menu bar with options: File, Edit, View, Insert, Cell, Kernel, Widgets, and Help. To the right of the menu bar, there is a "Trusted" status indicator and a "Python 3" kernel selector. Below the menu bar is a toolbar with various icons. The "Run" button, which is a play icon, is circled in red. The main content area of the notebook displays a code cell with the title "Numpy". The text in the cell explains that Numpy provides high-performance data structures for data manipulation and numeric computation, and introduces the `numpy.ndarray` data structure. It then shows two code snippets. The first snippet, labeled "In [52]:", is `import numpy as np`. Below it, the text says "We can initialize a Numpy array from a Python list using the `numpy.array` function:". The second snippet, labeled "In [53]:", is a multi-line code block: `# if the argument is a list of numbers, the array will be a 1-dimensional vector`, `a = np.array([1, 2, 3, 4, 5, 6])`, and `a`. Below this code, the output is shown: "Out[53]: array([1, 2, 3, 4, 5, 6])". A third snippet, labeled "In []:", is another multi-line code block: `# if the argument is a list of lists, the array will be a 2-dimensional matrix`, `A = np.array([[1, 2, 3, 4], [5, 6, 7, 8], [9, 10, 11, 12], [13, 14, 15, 16]])`, and `A`.

Execute commands by selecting a cell and clicking the **Run button** on the header of the page or by **Shift+Enter**. You will see the output of the command just below the cell.

You can tweak and modify the code as you wish and execute it again.

Exercise

For the exercise, you will solve a classification task using **Scikit-learn** over some given dataset. Each available dataset is already split into training and test sets. Choose a dataset, train a classifier on the training set and predict the labels on the test set. Hopefully, your classifier will classify the examples in the test set with higher accuracy than the reference baseline for the chosen dataset.

Exercise

Datasets

OCR

Optical Character Recognition



Spambase

Spam email classification



Presidential campaign tweets

Classification of tweets from D. Trump and H. Clinton



Exercise

Material

Download the material:

https://disi.unitn.it/~passerini/teaching/2021-2022/MachineLearning_AIS/index.html

The material contains the three datasets, each one containing:

- ▶ The training set examples;
- ▶ The training set labels;
- ▶ The test set examples;
- ▶ The test set labels;
- ▶ A README containing info about the dataset.
this file also contains the reference baseline accuracy;
- ▶ Other info files.

Exercise

Step-by-step

1. Choose a dataset;
2. Experiment with a classification algorithm of your choosing;
3. Test your classifier using cross-validation over the training set
4. Train your classifier over the full training set;
5. Use the classifier to predict the examples in the test set;