Machine Learning with Python

Numpy / Matplotlib / Scikit-learn

Giovanni Pellegrini

November 29, 2020

Setup

On lab machines

Download and extract the Scikit-learn lecture material from:

`http://disi.unitn.it/~passerini/teaching/2020-2021/MachineLearning/`

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
> pip install jupyter
```

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

`http://disi.unitn.it/~passerini/teaching/2020-2021/MachineLearning/`

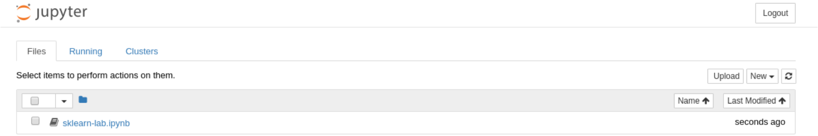
Open the terminal in the folder containing the extracted files and run:

```
> jupyter notebook
```

Setup

Jupyter notebook

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

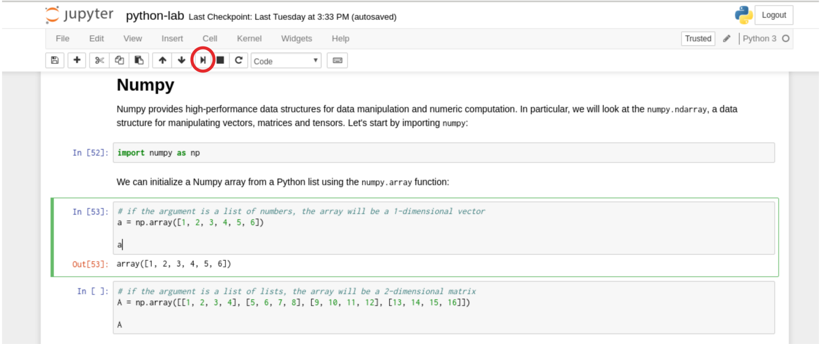


Select items to perform actions on them.

	Name	Last Modified
□	sklearn-lab.ipynb	seconds ago

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

Jupyter notebook



Numpy

Numpy provides high-performance data structures for data manipulation and numeric computation. In particular, we will look at the `numpy.ndarray`, a data structure for manipulating vectors, matrices and tensors. Let's start by importing `numpy`:

```
In [52]: import numpy as np
```

We can initialize a Numpy array from a Python list using the `numpy.array` function:

```
In [53]: # 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])

         a
Out[53]: array([1, 2, 3, 4, 5, 6])
```

```
In [ ]: # 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]])

        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.

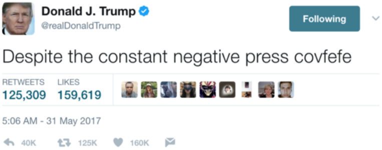
Exercise

Datasets

OCR
Optical Character Recognition

Spambase
Spam email classification

Presidential campaign tweets
Classification of tweets from D. Trump and H. Clinton



Donald J. Trump
@realDonaldTrump
Following

Despite the constant negative press covfefe

RETWEETS 125,309 LIKES 159,619

5:06 AM - 31 May 2017

40K 125K 160K

Exercise

Material

Download the material:

`http://disi.unitn.it/~passerini/teaching/2020-2021/MachineLearning/`

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;