



UNIVERSITY
OF TRENTO



Machine Learning with Python

Numpy / Matplotlib / Scikit-learn

Luca Erculiani

Setup

On lab machines



Download and extract the Scikit-learn lecture material from:

<http://disi.unitn.it/~passerini/teaching/2019-2020/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 jupyter
```

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

<http://disi.unitn.it/~passerini/teaching/2019-2020/MachineLearning/>

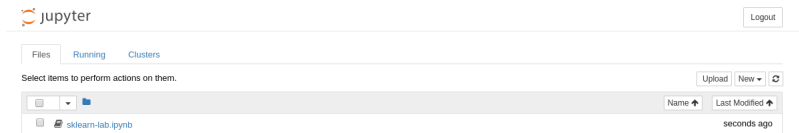
Open the terminal in the folder containing the extracted files and run:

```
> jupyter notebook
```

Setup

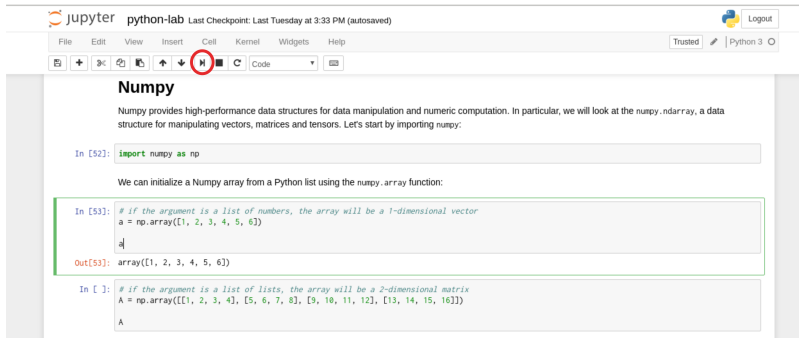
Jupyter notebook

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



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

Jupyter notebook



The screenshot shows the Jupyter Notebook interface. At the top, it says "jupyter python-lab" and "Last Checkpoint: Last Tuesday at 3:33 PM (autosaved)". There is a "Logout" button in the top right. Below the title bar is a menu bar with "File", "Edit", "View", "Insert", "Cell", "Kernel", "Widgets", and "Help". To the right of the menu bar are "Trusted" and "Python 3" indicators. Below the menu bar is a toolbar with various icons, including a red circle around the "Run" button (a square with a play symbol). The main content area has a title "Numpy" and a paragraph: "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.

Assignment

For the second Machine Learning assignment you will solve a classification task using **Scikit-learn** over some given dataset. Each available dataset is already split into training and test sets. Your task is to choose a dataset, train a classifier on the training set and predict the labels on the test set. To pass the assignment, your classifier has to classify the examples in the test set with higher accuracy than the reference baseline for the chosen dataset. Additionally, you need to test your algorithm via cross-validation over the training set and produce a report containing the results obtained.

Assignment

Datasets

OCR

Optical Character Recognition



Spambase

Spam email classification



Presidential campaign tweets

Classification of tweets from D. Trump and H. Clinton



Assignment

Material

Download the assignment material:

<http://disi.unitn.it/~passerini/teaching/2019-2020/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.

Assignment

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;
6. Place the labels in a file, in the same order as you read the test examples and in the same format of the labels in the training set.

Assignment

Report

Write a report describing the learning algorithm used and discussing the results obtained; The report should contain at least:

- The average precision, recall, and F_1 over the cross validation folds and over the test set.

Using `cross_val_score` you can specify 'precision', 'recall' and 'f1' for the scoring parameter.

For the OCR dataset, in which you do multiclass classification, use weighted averaging, i.e. using 'precision_weighted', 'recall_weighted' and 'f1_weighted';

- The plot of the learning curve, as shown in the lecture;

Assignment

Submit

- After completing the assignment submit it via email
- Send an email to mllab@unitn.it
- Subject: sklearnSubmit2019
- Attachment: `id_name_surname.zip` containing:
 - The text file, named `test-pred.txt`, containing the final predictions;
 - The code used to produce the predictions, the results and the plots;
 - The report in PDF format.

NOTE

- **No group work**
- This assignment is mandatory in order to enroll to the oral exam