Teachers: J. Hennebert, A. Perez-Uribe
Assistants: J. Donzallaz, J. Rial, O. Zayene

Practical work 01 – September 19 2023
Let's get started

The main objective of this Practical Work (PW) is to set all things up regarding your computer environment. We then have some exercises to wrap things up regarding what was said in the chapter 1 on **Fundamentals on Machine Learning**.

In this class we are going to use Python as programming language to perform some of the PW. So, this weeks' PW will include setting up Python on your computer and getting familiar with this language.

You have basically 3 ways to work with Python :

a) Using a Python console in interactive mode. When Python is installed on your computer, just open a terminal and launch python.
b) Using a text editor or an IDE to develop scripts of functions (i.e. saving ".py" files), as you would do it in Java or C. A good IDE is Pycharm from Jet Brains.
c) Using iPython notebooks via a web browser where you can mix code and Markdown text inputs in cells.

For this class we recommend you to use option c) above and work with iPython Notebooks for your homeworks. It will allow you to answer questions and input code at the same time. Alternatively, you can also decide to give back pure Python files together with a little report.

Note : We use Python version 3 in this class.

For this PW, **there is no need to return any report**. From next week, we will assume that your environment is all set up and that you have gotten familiar with Python.

Exercice 1 Moodle subscription

Register on moodle at `https://moodle.msengineering.ch`, id 2356. To find the class navigate to Accueil du site → Région F (HES-SO) → Principes Théoriques Fondamentaux (FTP) → FTP - 2023-2024 → FTP-MachLe. Use subscription key : `moodlemsekey`.

Most of the time, you'll need to submit your PW reports on Moodle for the **next Monday at 10h00**. However we may override this rule. In any cases, the dates indicated in Moodle are the one you should follow.

Exercice 2 Python installation on your computer

Skip this part if you are already set up with Python, Jupyter notebooks and your favorite IDE for Python. Otherwise, we recommend the following installations :

— Jupyter notebooks with manual installation in virtual environments (**recommended**) - First install Python 3 (from 3.9 or up, but python3) from https://www.python.org. Then create a working directory and open a terminal in the directory :

a) Install the virtual env : `python3.9 -m venv venv`
b) Activate the virtual env on Mac : `source venv/bin/activate`
c) Activate the virtual env on Windows : `C:\> <venv>\Scripts\activate.bat`
d) Install Jupyter : `pip install jupyter`
e) Install needed libs : `pip install numpy matplotlib pandas scikit-learn`
f) Launch Jupyter : `jupyter notebook`

Whenever you need to re-launch the notebooks, repeat steps b) and f). Whenever needed install requested libraries with `pip install LIB_NAME` once the virtual env is setup step b).

— Anaconda platform - a distribution of Python with popular data science packages that are pre-installed or easily installable. https://www.anaconda.com. Select the `Anaconda Distribution`. You should obtain something similar to Figure 1 below (Fig may be outdated).

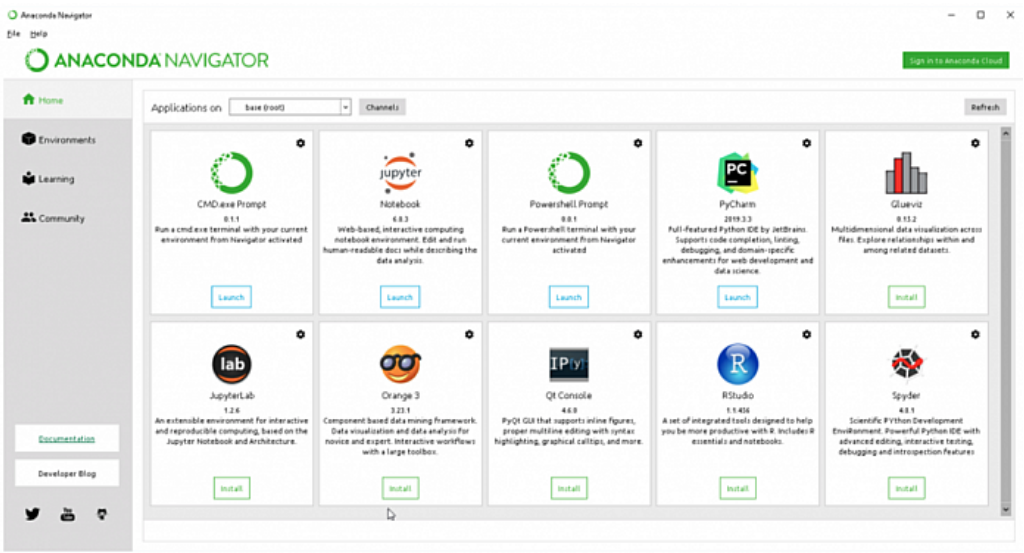

FIGURE 1 – Installing the Anaconda platform

— PyCharm IDE - an integrated development environment for Python. https://www.jetbrains.com/pycharm. You may select Professional or Community edition - the Community edition should be enough for this class. Free licenses are given to students, use your `hes-so.ch` address for the registration by JetBrains.

— If you do not want to install anything on your computer, you may use Google Colab at https://colab.research.google.com/.

Exercice 3 Python language in a nutshell

We assume here that students are knowledgeable in other programming languages such as Java or C and that basic data structure concepts are known. If you know already Python and the concept of notebooks, then you can skip this exercise.

— Open Jupyter from the command line or from the Anaconda Navigator. Jupyter Python notebooks run in your browser so, after launching Jupyter from Anaconda, a browser should show up. Open a new notebook from menu File and follow the User Interface Tour as illustrated in Figure 2.

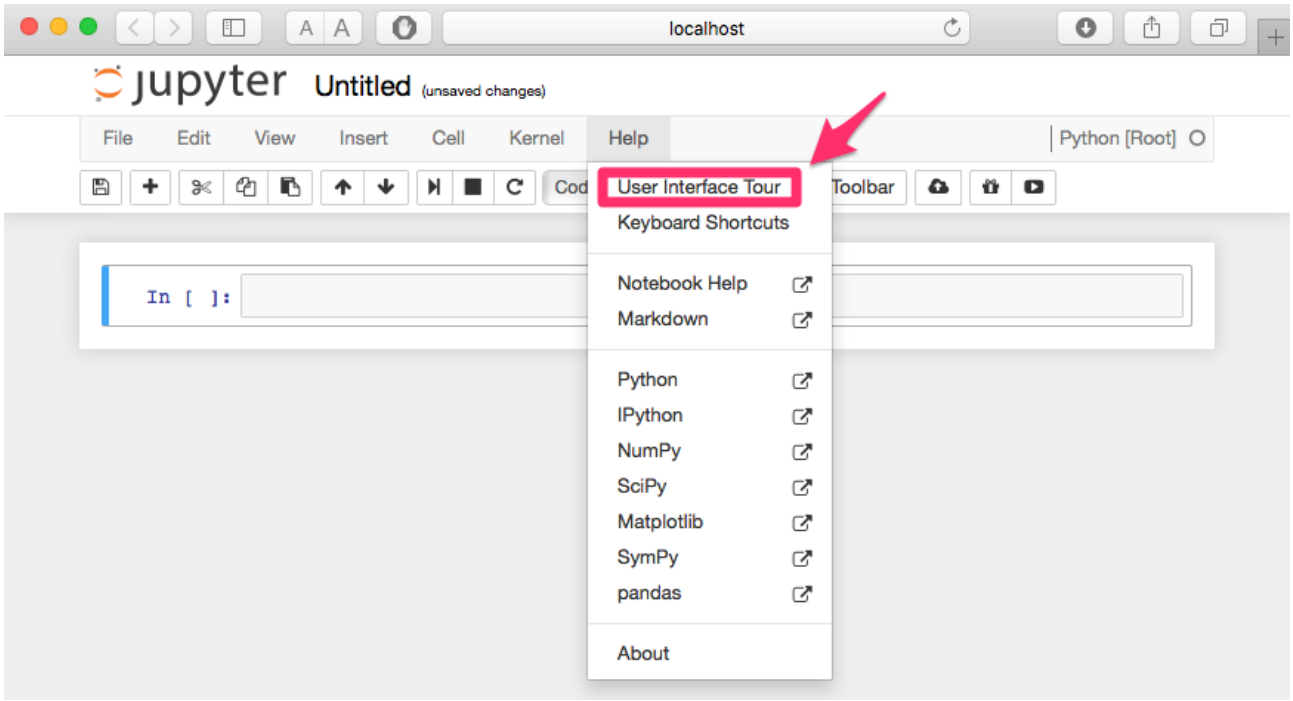

FIGURE 2 – First steps with Python notebooks.

— Download the file `intro-python-3.ipynb` from moodle and open it from Jupyter in Anaconda. You need to navigate where you stored the ipynb file. See Figure 3 below.

— Go through the content of the `intro-python-3.ipynb` notebook and play with the cells. This document should give you a quick introduction to Python assuming that you are fluent with other programming languages. For the text cells, you may also want to get familiar with Markdown syntax if not known already, for example by reading this page https://help.github.com/articles/basic-writing-and-formatting-syntax/.

— If you want a more fully fledged introduction to Python, read the official tutorial from https://docs.python.org/3.9/tutorial/.

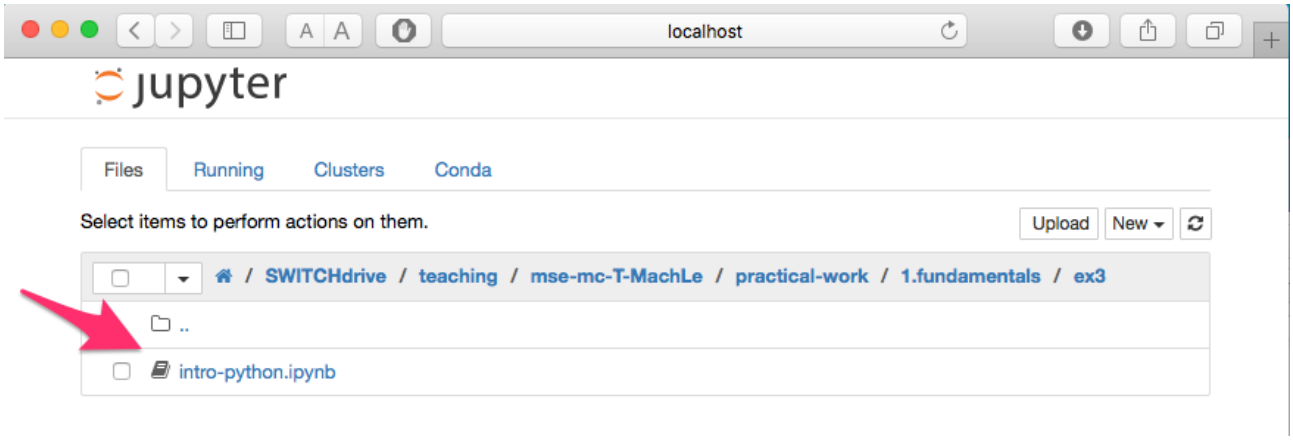

FIGURE 3 – Intro to Python language and Python notebooks.

Exercice 4 Data visualization

To train a bit yourself, we propose you to do a visualisation workout with the Iris dataset https://en.wikipedia.org/wiki/Iris_flower_data_set. Download the file `iris.txt` from moodle or from http://www.statlab.uni-heidelberg.de/data/iris/. Put it in a Python data structure and attempt to reproduce a plot close to the one in Figure 4. Advice : start by working with individual plots and then move to a grid of plots with `subplot` **or** take the shortcut of using the library `pandas`.

The iris species classification task is a classical one. The goal is to distinguish three kinds of iris : setosa, versicolor and virginica. In the data set, a botanist has extracted 4 *features* : sepal length, sepal width, petal length and petal width.

What can you say regarding class separation ? One class seems easier to distinguish from the others, which one ? How would you separate the other two ?

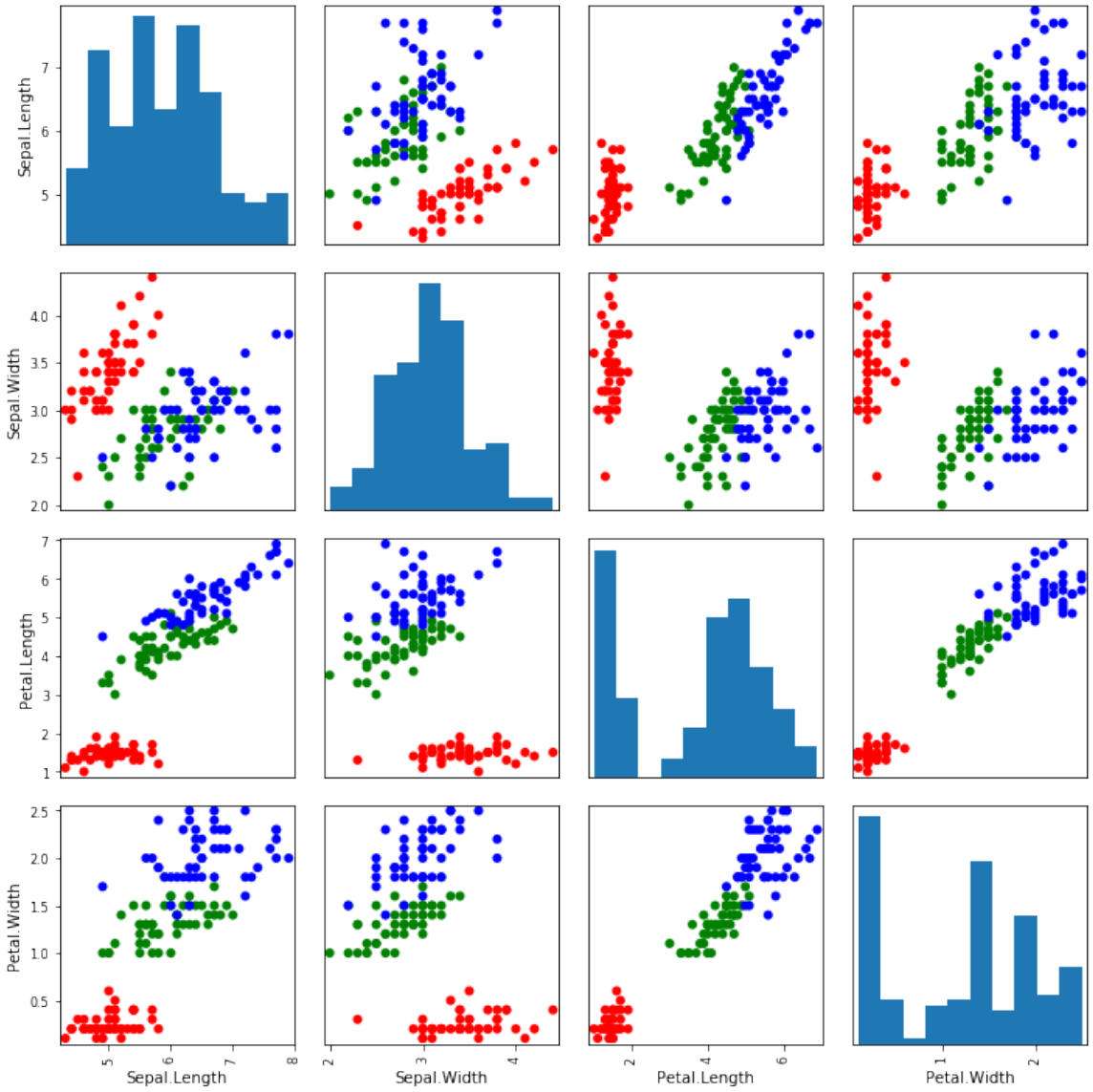

FIGURE 4 – Visualization of the Iris data as a pairwise scatter plot. The diagonal plots the marginal histograms of the 4 features. The other cells contain scatterplots of all possible pairs of features. Red circle = setosa, green diamond = versicolor, blue star = virginica.

Exercice 5 Find your own examples of Machine Learning Tasks

Complete the Table below with 2 new examples of machine learning tasks, explaining what is the task T, performance measure P and training experience E.

Tableau 1-1. Quelques exemples de machine learning proposés par Mitchell

Cas d'application	Checkers learning	Handwriting recognition	Robot driving learning
Tasks T	Playing checkers	Recognizing and classifying handwritten words within images	Driving on public four-lane highways using vision sensors
Performance measure P	Percent of games won against opponents	Percent of words correctly classified	Average distance traveled before an error (as judged by human overseer)
Training experience E	Playing practice games againts itself	A database of handwritten words with given classifications	A sequence of images and steering commands recorded while observing a human driver

FIGURE 5 – Source : Data Science - fondamentaux et études de cas, Michel Lutz et Eric Bernât, Eyrolles.

Exercice 6 Review questions

1) Supervised vs. unsupervised systems

Of the following examples, which one would you address using a supervised or an unsupervised learning algorithm ? Give some explanations for your answers.

a) Given email labeled as spam/not spam, learn a **spam filter**.
b) Given a set of news articles found on the web, group them into sets of **related articles**.
c) Given a database of customer data, automatically discover **market segments** and group customers into different market segments.
d) Given a dataset of patients diagnosed as either having **glaucoma** or not, learn to classify new patients as having glaucoma or not.

2) Classification vs. regression systems

Can we transform a regression problem into a classification problem ? What would be the benefits of doing so ?

3) Other questions

a) Why is it important that a test set is *independent* to the training set in a machine learning system ?
b) What is the meaning of the hat in the equation $\widehat{y} = h(\mathbf{x})$?
c) What is the difference between machine learning and statistics ?
d) Is a *detection* system a regression or classification system ? Give an example.

Exercice 7 Reading assignments

— Read the first chapter of Murphy's book "Machine Learning".
— Read the first chapter of Biernat and Lutz's book "Data Science – fondamentaux et études de cas".

Build your own summary of these chapters by doing a taxonomy of machine learning problems. You can find the pdfs on Moodle.