Python data analysis

Nicolas Rousset

Monday, 21st of November

Python eco-system

- Python is widely used, with a strong eco-system, that provides everything you can do in programming with a “open-source” flavour (excluding interaction with proprietary software)
- You will have a strong integration in all the major web hosting providers (AWS, google cloud, OVH)
- Python libraries, especially pandas, are very user friendly and provide a lot of built-in possibility (like exporting a data-set to xml)
- You have some librairies to read / write excel, words,latex . . .

Numpy : efficient arrays

This library is a key to all of the python scientific / data analysis eco-system, as it provides the low-level C-like structure that allows high performance computation.

Numpy is about manipulating, array, matrix, tensor (matrix of any dimensions).

It also provides basic analysis tools on theses structures (correlation, mean, standard deviation, etc . . .)

You might not use directly, as it is quite low level, but you should know it is used by all libraries, even picture manipulation. For example if you want to make a machine learning library interact with pandas, it will go through numpy structure that they both use.

Pandas : excel for programmers

It is the main library used by data analysis, especially in study mode. The main object is the Dataframe, which is an excel sheet for python programmers, it allows to handle dataset with mixed types variables : string, int, float,date . . .

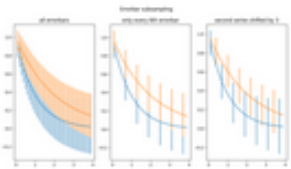
It also provides a lot of handy built-in features, including, but not only :

- histograms
- export to html / xml / latex - simple data analysis

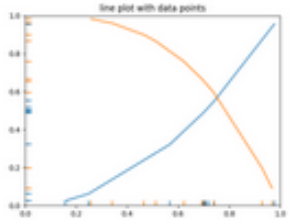
Matplotlib : drawing beautiful graphs

The main base library for drawing graphs. Main page for it is :

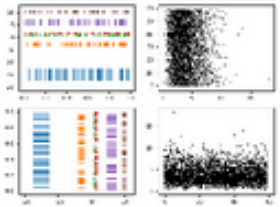
https://matplotlib.org/stable/gallery/index



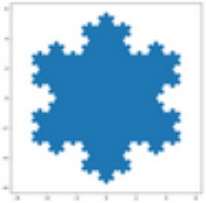
Errorbar subsampling



EventCollection Demo

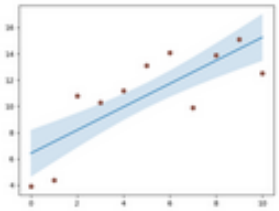


Eventplot Demo



Filled polygon

Fill Between and Alpha



Filling the area between lines

Altair : interactive html graphs

Another drawing graph library. One of its great strong point compare to Matplotlib is that it can provide interactive graphs

Other librairies

The following libraries won't be presented in this training.

Scipy

Scipy is a catalog of mathematical function and algorithms :
- integration
- optimization
- differential equations
- eigen values

So it is useful in a lot of specialized domains, but no so much in data analysis. I don't think it is worth investing in understanding it directly, you will use it when typing "doing something mthematical" in stackoverflow that will provide some scipy code.

Scikit learn : machine learning

A very popular library for machine learning algorithms :

- linear and logistic regression
- clustering (KNN)
- classification
- dimensionnality reduction (PCA)

Tensorflow / pytorch / keras

One of the reason of the peaks of popularity of python at the moment, they are the main libraries for neural network / artificial intelligence.

Tensorflow / keras are developped by google
Pytorch is developped by facebook

Seaborn

A library for graph and data visualisation

Geopandas / cartopy

For handling and visualisation of geographical datas