Statistical analysis of experimental data
Introduction to Python

Artur Kalinowski

FACULTY OF
PHYSICS
UNIVERSITY
OF WARSAW

Lecture 02

October 12, 2023

Introduction

Lecture concept

The goal of the lecture is not only to present the theoretical concepts but also to demonstrate how they can be applied in practical terms to data analysis.

Different concepts as well as analysis methods will be discussed using code examples.

We will address many (relatively simple) problems and try to look at their numerical solutions.

All examples presented are based on Python and multiple additional packages, which will be shortly introduced today.

Introduction

Web resources

Kampus platform will be used for lecture slides, Python notebooks, as well as home exercises (and final exam):

https://kampus-student2.ckc.uw.edu.pl/course/view.php?id=14456

All files will also be available from the dedicated web page:

http://www.fuw.edu.pl/~zarnecki/SAED/

accessible without USOS account...

Environment

Containers

To avoid the burden of installing and configuring all the required packages on your personal computers, we will use virtual environments known as containers. Containers are lightweight versions of old time virtual machines. Two popular environments used to run the containers:

- Apptainer (previously singularity; still difficult to find in Google)
- Docker

Both Docker and Apptainer can use the same input files.
Docker is available for any operating system, Apptainer only for Linux ones.

Notebooks presented at this lecture are developed with akalinow/root-fedora35 container prepared for the other course some time ago. This is a container based on Fedora Linux ditribution, and conains all packages we will use during the classes.

Detailed instructions on installing and starting Docker are given here

Environment

Git and Google Colaboratory

All scripts discussed during the course will be available for download at the Kampus platform and the dedicated lecture web page. They will also be uploaded to github repository:

https://github.com/zarnecki/SAED

You also open them in your browser using the Google Colab platform.
To do so, use CO Open in Colab link and select the notebook you want to load.

Please remember that you need to save the file to (your) Google Disc before you make any modifications. Otherwise, all your changes will be lost!..

Python packages

Python tools

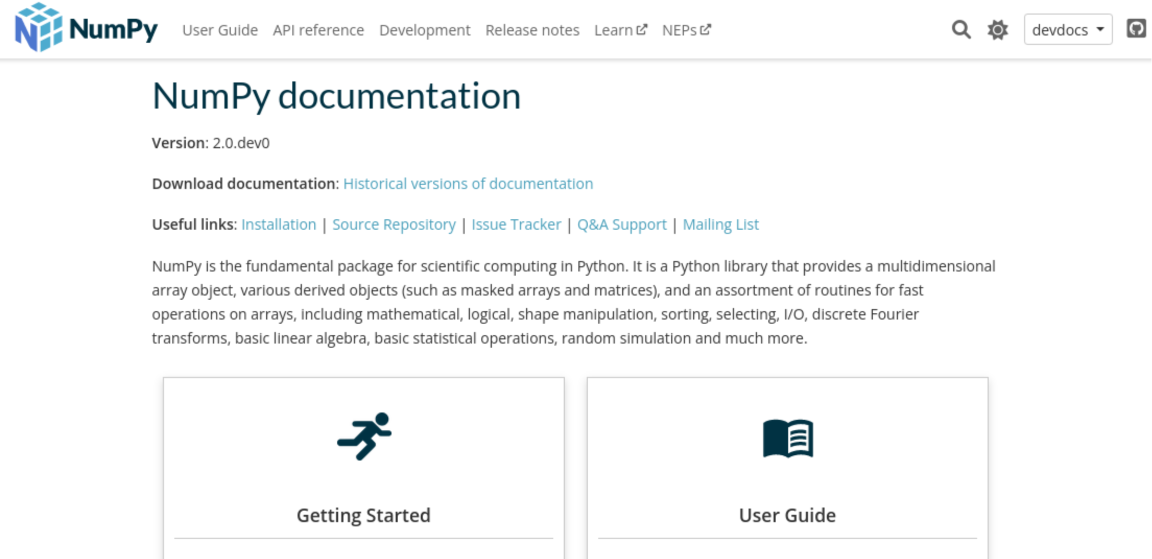
One of the main advantages of Python is the large number of diverse packages developed for various applications...

During the course we will use the following packages for computation and plotting:

- numpy - The fundamental package for scientific computing with Python
- matplotlib - Visualization with Python
- SciPy - Fundamental algorithms for scientific computing in Python
- scikit-learn - Machine Learning in Python, simple and efficient
- pandas - Handling more complex data structures

See lecture notebook for more details and examples: CO Open in Colab

Python packages



NumPy documentation

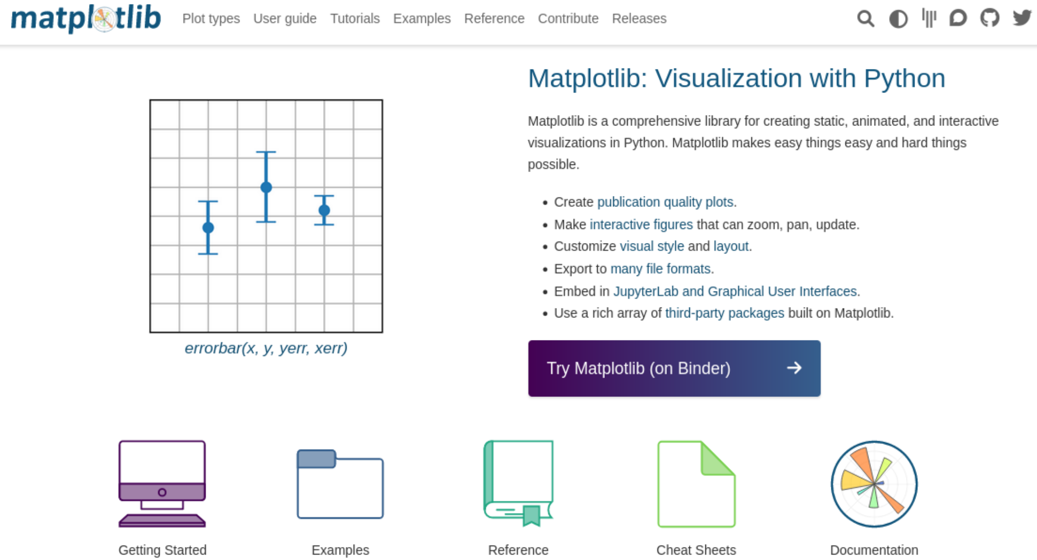
Version: 2.0.dev0

Download documentation: Historical versions of documentation

Useful links: Installation | Source Repository | Issue Tracker | Q&A Support | Mailing List

NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape manipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

Python packages



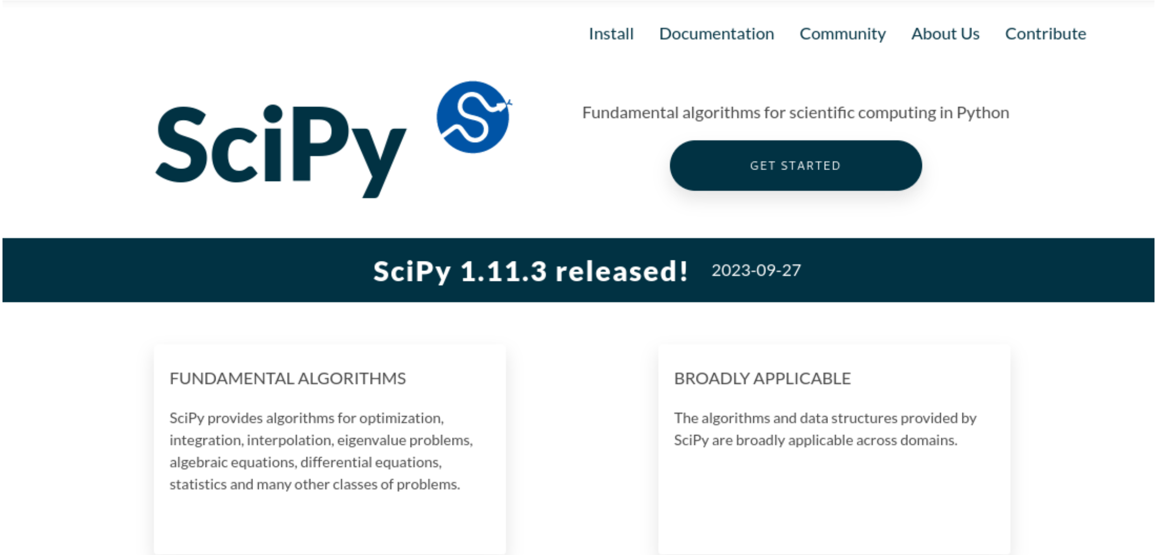
Matplotlib: Visualization with Python

Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python. Matplotlib makes easy things easy and hard things possible.

- Create publication quality plots.
- Make interactive figures that can zoom, pan, update.
- Customize visual style and layout.
- Export to many file formats.
- Embed in JupyterLab and Graphical User Interfaces.
- Use a rich array of third-party packages built on Matplotlib.

errorbar(x, y, yerr, xerr)

Python packages



Fundamental algorithms for scientific computing in Python

SciPy 1.11.3 released! 2023-09-27

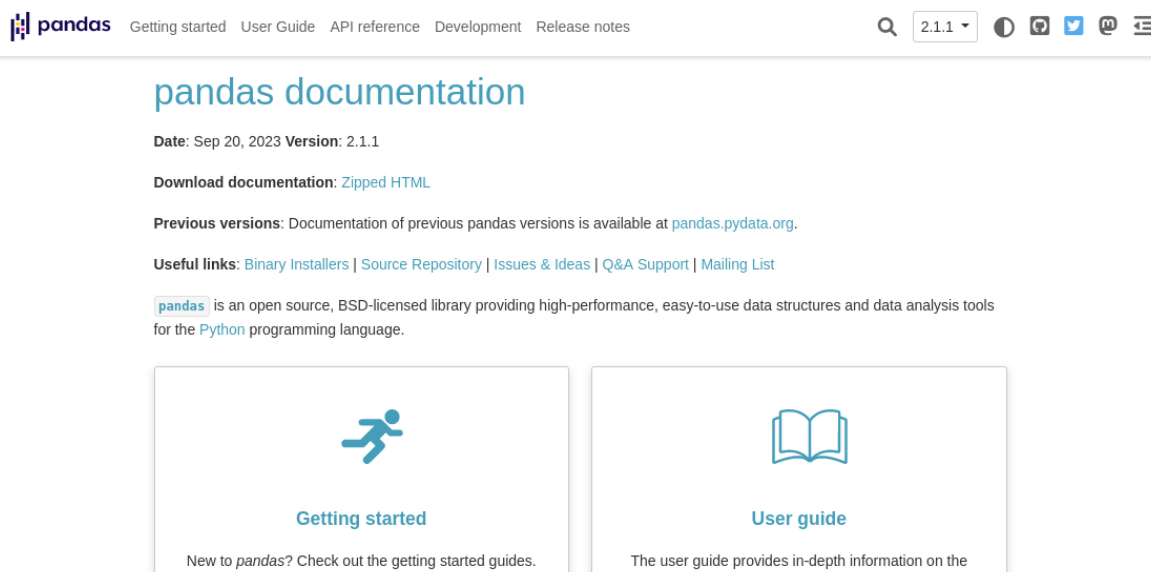
FUNDAMENTAL ALGORITHMS

SciPy provides algorithms for optimization, integration, interpolation, eigenvalue problems, algebraic equations, differential equations, statistics and many other classes of problems.

BROADLY APPLICABLE

The algorithms and data structures provided by SciPy are broadly applicable across domains.

Python packages



pandas documentation

Date: Sep 20, 2023 **Version**: 2.1.1

Download documentation: Zipped HTML

Previous versions: Documentation of previous pandas versions is available at pandas.pydata.org.

Useful links: Binary Installers | Source Repository | Issues & Ideas | Q&A Support | Mailing List

`pandas` is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

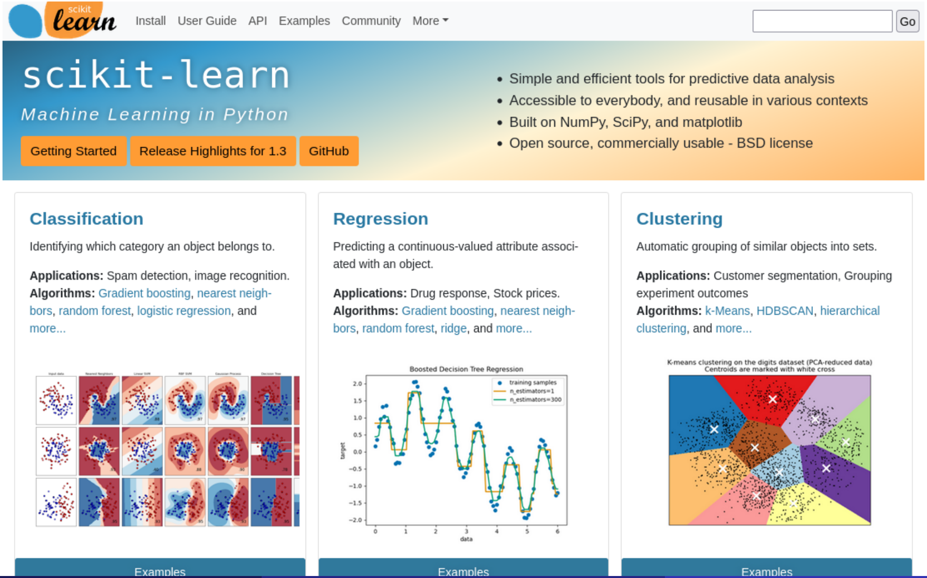
Getting started

New to *pandas*? Check out the getting started guides.

User guide

The user guide provides in-depth information on the

Python packages



scikit learn Install User Guide API Examples Community More Go

scikit-learn

Machine Learning in Python

Getting Started | Release Highlights for 1.3 | GitHub

- Simple and efficient tools for predictive data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

Classification

Identifying which category an object belongs to.

Applications: Spam detection, image recognition.
Algorithms: Gradient boosting, nearest neighbors, random forest, logistic regression, and more...

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.
Algorithms: Gradient boosting, nearest neighbors, random forest, ridge, and more...

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes
Algorithms: k-Means, HDBSCAN, hierarchical clustering, and more...

K-means clustering on the digits dataset (PCA-reduced data)
Centroids are marked with white cross

Examples | Examples | Examples

Homework

Graphical presentation of the data

Look for the recent data describing age and gender structure of the population of Poland. One of the possibile sources is the Central Statistical Office (Polish: Główny Urzad Statystyczny) or Poland's Data Portal.

1. Prepare the plot showing the gender balance (men to women ratio) as a function of age.
2. Can the observed dependence be used to draw any conclusions concerning the life expectancy? Why do you think so?

Solutions should be uploaded until October 26.

Solutions (dedicated files or screenshots of the notebook final output) should be uploaded to Kampus in readable format (PDF, JPG, PNG). You should add the link to your notebook in Google Colab as a comment.
Please share your Google Colab space with azarnecki@uw.edu.pl, so I can have a look at your notebook, if needed.