uc3m | Universidad Carlos III de Madrid

OPENCOURSEWARE
ADVANCED PROGRAMMING
STATISTICS FOR DATA SCIENCE
Ricardo Aler

Python for data analysis

The purpose of this topic is to give a quick introduction to the Python language, as well as the most important libraries for doing data analysis and visualization : numpy, pandas, matplotlib, and seaborn.

Lectures start by introducing base (or core) Python :

- Its main data types : scalars, strings, lists, dictionaries, sets.
- Use of modules and functions that belong to them.
- Control flow : loops, if-then-else, list comprehension
- File input-output

The second topic is the main data types to be used in data analysis : numpy and pandas. With respect to numpy, the main concepts such as arrays and matrices, universal and reduction functions, and broadcasting, among others, will be explained. With respect to numpy, the dataframe data structure will be explained, together with its main operations : indexing and modification.

The third topic is the library scikit-learn. The main stages of machine learning using this library will be explained : model training, model evaluation, hyper-parameter tuning, and pre-processing. The concept that unifies the whole procedure, pipelines, will be explained in detail, and how it avoids data leakage, and how pipeline hyper-parameters can be tuned. An advanced topic for creating pipeline steps, based on Python object orientation, will also be introduced.

The fourth and last topic is data visualization using two Python libraries : Matplotlib and Seaborn. Matplotlib is low level, but allows complete control of the plots. Seaborn is higher level and permits constructing plots with shorter plot descriptions. An important concept for building Seaborn plots will be introduced : wide versus long formats for dataframes.

Associated Material :

- Slides and some exercises (check the lecture notes).
- Three of the labs addresses the topics of this section :
    - Feature Extraction in Python for text analysis
    - Scikit-learn
    - Pipelines and plotting