Introduction to Python for Data Science

NumPy, Pandas

Thomas Torku, Ph.D.

Week 3-4

Outline

Data Types and Control Structures

- Python is a dynamic, interpreted language used in a variety of programming environments.
- Basic Data Types:
  - Integers, floating-point numbers, strings, and booleans.
  - Operations and expressions.
- Data Structures:
  - Lists: Ordered and changeable collections.
  - Dictionaries: Key-value pairs, unordered, and mutable.
  - Sets: Unordered collections of unique elements.
  - Tuples: Ordered and unchangeable collections.
- Control Structures:
  - Conditional statements (if, elif, else).
  - Loops (for, while) and iteration over collections.
- Functions: Define reusable code blocks with def and return statements.
- Modules: Import and use code from Python libraries.

Introduction to NumPy

- NumPy is a fundamental package for scientific computing in Python.
- Provides support for large, multi-dimensional arrays and matrices.
- Rich collection of mathematical functions to operate on these arrays.
- Creating and Manipulating Arrays:
  - np.array, np.zeros, np.ones, np.arange, np.linspace.
  - Array operations: element-wise and matrix operations.
- Basic Operations and Broadcasting:
  - Arithmetic operations, comparisons, logical operations.
  - Broadcasting rules for combining arrays of different sizes.
- Indexing, Slicing, and Iterating:
  - Accessing and modifying array elements.
  - Slicing arrays to create sub-arrays.
  - Iterating over multi-dimensional arrays.

Introduction to Pandas

- Pandas is an open-source library providing high-performance, easy-to-use data structures.
- Designed to make working with "relational" or "labeled" data intuitive.
- Series and DataFrame: The core data structures for one-dimensional and two-dimensional data respectively.
- Data Importing and Exporting:
  - Reading from and writing to different file formats (CSV, Excel, SQL databases, etc.).
- Data Cleaning and Preparation:
  - Handling missing data, dropping or filling NA values.
  - Data transformation with operations such as merging, reshaping, and pivot tables.
- Basic Data Analysis with Pandas:
  - Descriptive statistics, grouping data, applying functions.
  - Visualizing data with the help of Matplotlib integration.

Data Manipulation and Analysis

- Combining the power of NumPy and Pandas for effective data analysis.
- NumPy for numerical and mathematical computation.
- Pandas for structured data manipulation and analysis.
- Example: Data Analysis Workflow
  - Import data using Pandas.
  - Clean and prepare data: handling missing values, filtering rows/columns, and data type conversion.
  - Analyze data: using NumPy for statistical analysis, array operations, and Pandas for group by operations, merge/join datasets.
  - Visualization: creating plots and charts to visualize trends and patterns.
  - Exporting results: saving processed data to files or databases.
- Case Study:
  - Brief introduction to a real-world dataset.
  - Demonstrating data manipulation and cleaning techniques.
  - Applying statistical methods to draw insights.
  - Visualizing the results for presentation.