CS7052-Machine Learning

Workshop 1: Data Processing in NumPy and Pandas, Exploring Titanic dataset

You will learn:

- To work with Jupyter Notebook
- To work with data using the Python language and the NumPy and Pandas packages
- To explore Titanic database

Introduction

Now Python is one of the most common programming languages. One of its advantages is a large number of packages that solve a variety of problems. In our course, we recommend using Pandas, NumPy, and other libraries, which greatly simplify the reading, storage, and processing of data. In future work, you will also be introduced to packages, which implement many machine learning algorithms.

Task 1- Beginning of work

To start the workshop, you need to install or run Jupyter Notebook.

Trying Jupyter Notebook in a browser

Please refer to the link below and choose "try in in a browser":

https://jupyter.org

I do recommend this option, if you are using a lab computer.

Installing Jupyter Notebook on your computer

There are multiple ways for the installation of Jupyter Notebook.

1. You can install Jupyter Notebook by visiting the link below and choose to "install the Notebook": https://jupyter.org. I do recommend this option or trying Jupyter Notebook in a browser if you are using a lab computer.
2. You can install Anaconda by visiting https://www.anaconda.com and then as a result you will have Python, Jupyter Notebook, and many other applications installed in one go. I do recommend this option if you are planning to use your laptop during workshops.

3. If you have python already installed on your computer you can open the terminal or the command-line interface, make sure the version of python installed on your computer is 3 or higher.

```
python --version
```

and then install Jupyter Notebook by running this command:

```
pip install notebook
```

if the version of Python installed on your computer is lower than 3 refer to this tutorial to install the latest version first.

https://www.youtube.com/watch?v=YYXdXT2l-Gg

Then open the terminal or command-line and type the following command

```
pip install notebook
```

Task 2- Get started with Jupyter Notebook

Open the terminal or command-line interface and run the following command to open Jupyter Notebook:

```
jupyter notebook
```

if you are using the browser version, Jupyter Netbook is already open on your computer.

Refer to the link below and follow the instructions on “creating a new notebook document” and familiarise yourself with the “Notebook user interface” and how you can execute the content of a cell (“structure of a notebook document”).

https://jupyter-notebook.readthedocs.io/en/latest/notebook.html

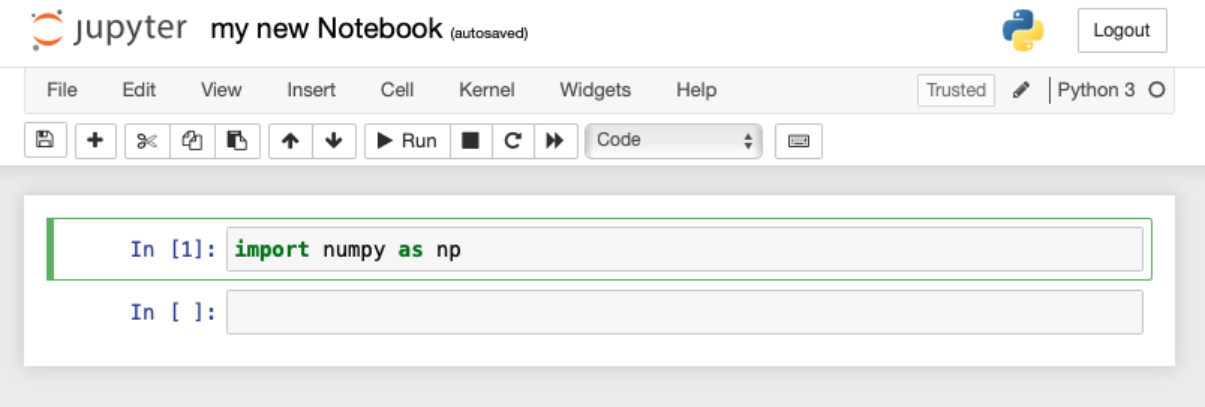
Task 3- Learn about the NumPy package in Jupyter Notebook

Refer to the Reading List section of the Machine Learning module on Weblearn.

Open “Python data analytics: with Pandas, NumPy, and Matplotlib” book by Fabio Nelli. I will refer to this book as **Nelli’s book** from now on.

Browse to chapter 3

Start by typing “import numpy as np” in line 1 of your Jupyter notebook as you can see on page 50 and in the following picture.



Then you can follow exercises of this chapter to familiarise yourself with NumPy.

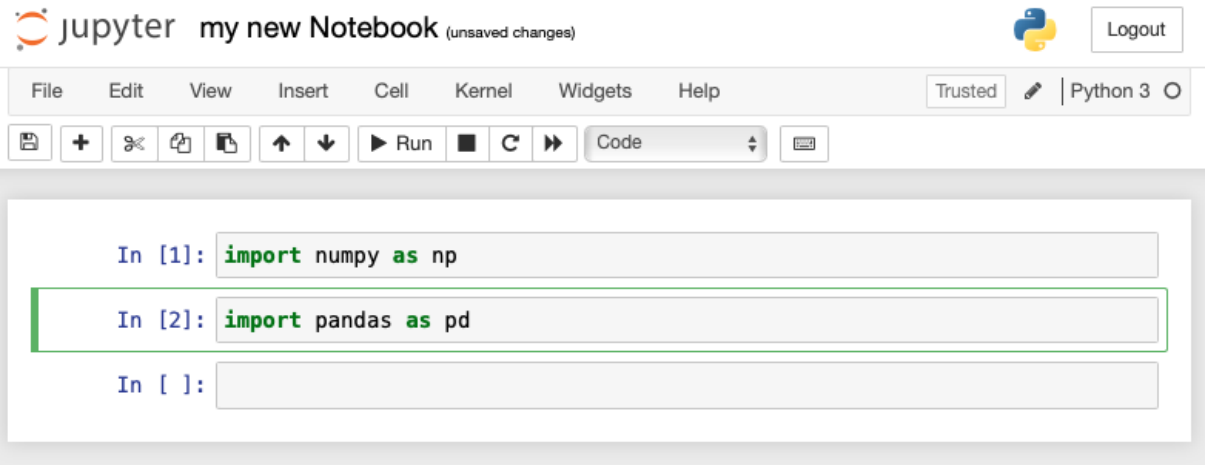
I particularly would like to ask you to run the code from the following sections:

- Ndarray: the heart of the library, page 50
- You can see data types supported by NumPy on page 54
- Matrix product, page 59
- Aggregate function, page 62
- Indexing, page 63

Task 4- Pandas Library

Start with

“import pandas as pd” and “import NumPy as np” in your Jupyter Notebook as you can see below:



As mentioned in chapter 4 (The pandas Library) of the same book (Nelli’s book)

Again, you can try all exercises of this chapter in your own time, but make sure you cover the following sections during the workshop

- DataFrame, page 102
- Defining a dataframe, page 103
- Selecting elements, page 105
- Assigning values, page 107

- “Not a Number” data, page 131

Task 5- Exploring Titanic dataset with Pandas

To start working with data, you must first load them from a file. In this task, we will work with data in CSV format intended for storing tabular data: columns are separated by a comma, the first row contains the column names.

An example of loading data into Pandas:

```
data = pd.read_csv('titanic.csv')
```

Data will be loaded as a **`DataFrame`**, with which you can conveniently work.

To see what the data is, you can use code in several styles:
• if only one index is specified, then rows are selected:

```
data[:10]
```

• or use the data frame method:

```
data.head()
```

One way to access columns in a data frame is to use square brackets and a column name:

```
data['Pclass']
```

To calculate some statistics (quantity, average, maximum, minimum), you can also use data frame methods:

```
data['Pclass'].value_counts()
```

A more detailed list of data frame methods can be found in the documentation.

Download the **titanic.csv** dataset and, using the data methods described above, find answers to the following questions.

Questions

1. How many men and women were on the ship? As an answer, give two numbers separated by a space.

2. What part of the passengers managed to survive? Calculate the proportion of surviving passengers. Answer in percentage rounded to two decimals.

3. What is the proportion of first-class passengers among all passengers? Answer in percentage rounded to two decimals.

4. How old were the passengers? Calculate the average and median age of passengers. As an answer, give two numbers separated by a space.

5. Does the number of siblings correlate with the number of parents/children? Calculate the Pearson correlation between the SibSp and Parch features.

6. What is the most popular female name on the ship? Extract the passenger's full name (Name column) from his name (First Name). This task is a typical example of what a data analysis specialist is faced with. The data is very heterogeneous and noisy, but you need to extract the necessary information from it. Try to manually parse several values of the Name column and work out a rule for extracting names, as well as dividing them into female and male.

More online resources related to workshop 1:

1. You can refer to the link below for more information regarding the installation of Jupyter Notebook: https://jupyter.org/install.html

2. For more information on Jupyter notebook refer to the following links:
   - get started with the Jupyter Notebook: https://jupyter.readthedocs.io/en/latest/content-quickstart.html
   - documents on Jupyter Notebook: https://jupyter-notebook.readthedocs.io/en/latest/?badge=latest

3. If you are unfamiliar with NumPy or matplotlib, we recommend reading the first chapter of the SciPy Lecture Notes.

   http://scipy-lectures.org/_downloads/ScipyLectures.pdf