

Lab1:

Preparation

1. Download and install Anaconda from anaconda.com
2. Open up the Anaconda Prompt and run the following commands to create a virtual environment for the lab called “aai_lab”, and then install some common packages into it

```
conda create -n aai_lab python=3.10
conda activate aai_lab
conda install tensorflow notebook pandas matplotlib numpy
conda install scikit-learn
conda install -c conda-forge nb_conda_kernels
```

3. Launch Jupyter Notebook by typing the following into the same prompt (with aai_lab activated)

```
jupyter notebook
```

4. Jupyter Notebook should open within your browser. Within the Jupyter Notebook, navigate to a folder location where you would like to store your lab files and then create a new notebook by clicking on New and selecting your virtual environment.
5. When the new notebook opens, rename it from “Untitled” at the top.
6. For a quick overview of the use of a Jupyter Notebook, you can have a look here: <https://realpython.com/jupyter-notebook-introduction/>
7. For an introduction to python you can have a look here: https://www.w3schools.com/python/python_syntax.asp

Python basics

For each of the following tasks, create a new cell in your notebook and write python code to generate the desired result:

1. Use a for loop to print the following output to the screen:

```
The number is 0
The number is 1
The number is 2
The number is 3
```

2. Write a function called `one_plus_sqrt` that takes a number x as input and returns $1 + \sqrt{x}$ using the `sqrt` function of the `math` module. Then print the result of the function rounded to 3 decimal places, when it is called with an input value of $x=2$.
3. Redefine your function `one_plus_sqrt` to test for negative inputs and return `None` if this is the case. Save the result of the function for the input value of -1 to a variable called `result`, and then print the datatype of the variable `result`.

4. Use a list comprehension to generate a list type variable that contains the values of x^2 for x from 0 to 100. Save the result to a variable called result and then print the 10th to 14th elements of the list.

Numpy basics

Just as you did in the previous section, for each of the following tasks, create a new cell in your notebook and write python code to generate the desired result:

1. Use the command “import numpy as np” to import the numpy package, and then print the version of numpy you have installed.
2. Use numpy to create a variable that contains a one dimensional array with the numbers 1 through 4 and then print both the value of the array and the shape of the array.
3. Create two 2-D arrays, the first with 1 row and 3 columns and the second with 3 rows and 1 column. Then print for each array both the value of the array and the shape of the array
4. Create a 2D array variable called big_array that has 6 rows and 6 columns containing all zeros. Print the value of the array.
5. Modify the value of big_array such that all of the elements in the middle two rows are set to 1, and then print the value of the entire array.

Matplotlib basics

Create a new cell for each of the following tasks and write python code to generate the desired result.

1. First import matplotlib using the command “import matplotlib.pyplot as plt”, then use the “fig,ax = plt.subplots()” command to instantiate a Figure with one Axes and get handles to the figure and axes objects. Next, create a variable called x which is a one dimensional numpy array of length 100 containing uniformly spaced numbers between -10 and 10, and then plot the function $y = \sin x$. Add the title ‘y=sin(x)’ to your plot and label the horizontal axis ‘x’ and the vertical axis ‘y’.

Pandas Basics

Once again, create a new cell for each of the following tasks and write python code to generate the desired result.

1. Create three the numpy arrays, the first called x is the same as the x from the previous question (i.e. a one dimensional numpy array of length 100 containing uniformly spaced numbers between -10 and 10), the second is an array that contains $\sin x$ and the third is $\sin^2 x$. Make a dictionary object called data where the keys are the strings ‘x’, ‘sin_x’ and ‘sin_squared_x’, and the values are the associated numpy arrays.
2. Import pandas using the command “import pandas as pd”, and then create a pandas dataframe called df using the data dictionary you created in the previous step. Then use the dataframe head() method to print the first few rows of your dataframe.
3. Print the mean of the ‘sin_squared_x’ column of the dataframe.

Data Investigation

Before the next exercise read chapter 2 End-to-End Machine Learning Project, section 1-4 of the textbook [1]

Download from Brightspace the csv file called **Med_insurance.csv** [2]. This csv file contains data of medical information and insurance costs ('charges'). Read this csv file using pandas and create a dataframe called *insurance_data_raw*

Perform the following steps and provide answers to all of the questions (use a mix of Markdown and Code cells in your notebook):

1. Investigate the data, what attributes do you see? How many rows do you have? Does the data contain missing values?
2. Compute the correlation between the "**charges**" and the rest of the attributes. Create a plot using the package matplotlib showing these correlations. Use an appropriate title for the plot and label the x and y axis.
Question: What do you observe?
Question: How is the insurance cost correlated with the age?
3. Separate the labels from the data. Make two variables *insurance_data_label* and *insurance_data* The first variable contains the labels of the data i.e, the expected output for which you want to train a machine learning model and the second variable will contain the rest of the data.

[1] Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2n Ed. By Aurelien Geron

[2] <https://www.kaggle.com/mirichoi0218/insurance>