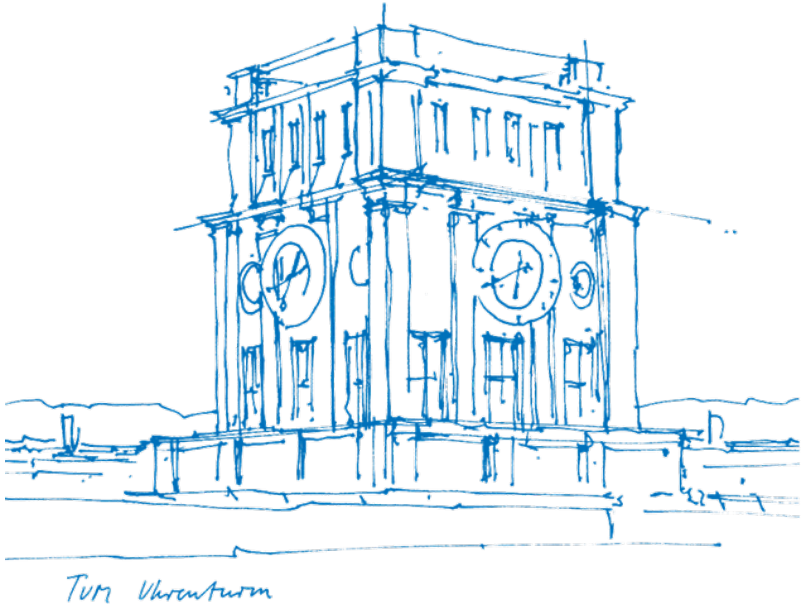
Introduction to Python for MLcomm

Fabian Steiner

Machine Learning for Communications – TUM LNT

October 18th, 2017



Overview

Introduction

Python Basics

Libraries for Numerical Computations
- NumPy
- SciPy
- Pandas

Outlook

Homework

What is Python?

- Python is an interpreted, non-statically typed language.
- It supports different programming paradigms (functional, object-oriented, imperative, etc.).
- It supports all major operating systems and comes with a huge standard library.
- Python as a language has different implementations:
  - CPython – standard, reference implementation.
  - PyPy – based on a just-in-time (JIT) compiler. Major speedups compared to CPython.
  - Cython – compiles Python to C.
- Python is open source.

Why Python for ML?

- Python has established a good reputation in the data science field.
- It is a language that is easy to start with.
- It is freely available (compare[1]: Matlab 2000 Euro + Statistics and Machine Learning toolbox: 1000 Euro).

[1] https://de.mathworks.com/pricing-licensing.html
[2] https://www.tensorflow.org/
[3] https://keras.io/
[4] http://scikit-learn.org/

Why Python for ML?

- Python has established a good reputation in the data science field.
- It is a language that is easy to start with.
- It is freely available (compare[1]: Matlab 2000 Euro + Statistics and Machine Learning toolbox: 1000 Euro).
- Three prominent machine learning libraries are developed in Python: Tensorflow[2], Keras[3], scikit-learn[4]

[1] https://de.mathworks.com/pricing-licensing.html
[2] https://www.tensorflow.org/
[3] https://keras.io/
[4] http://scikit-learn.org/

Why Python for ML?

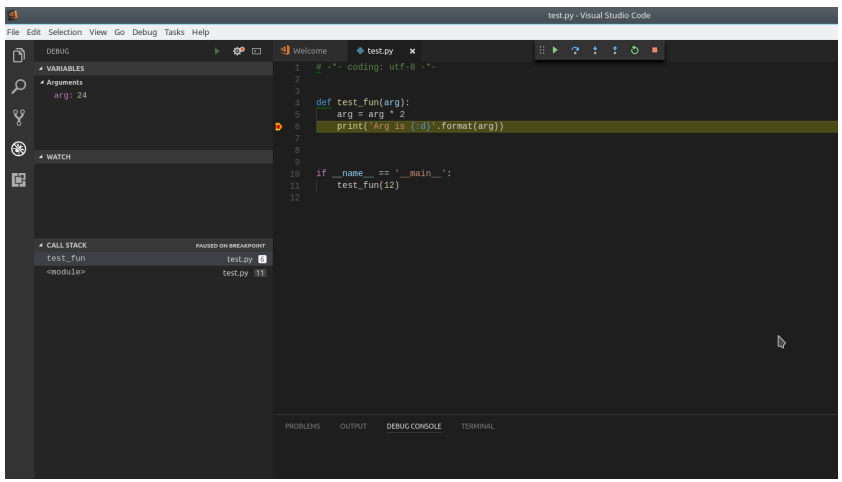
- Python has established a good reputation in the data science field.
- It is a language that is easy to start with.
- It is freely available (compare[1]: Matlab 2000 Euro + Statistics and Machine Learning toolbox: 1000 Euro).
- Three prominent machine learning libraries are developed in Python: Tensorflow[2], Keras[3], scikit-learn[4]

- Extend your horizon: There's a world beyond Matlab.

[1] https://de.mathworks.com/pricing-licensing.html
[2] https://www.tensorflow.org/
[3] https://keras.io/
[4] http://scikit-learn.org/

Programming Environment (I)

- We recommend using Microsoft Visual Studio Code[5] with installed Python support.



- Microsoft Studio Code is available for Windows, Linux and Mac.
- Various debugging possibilities and code style conformity checks.

[5] `https://code.visualstudio.com/`

Programming Environment (II)

- We provide a Ubuntu 64bit container for VirtualBox[6].
- The link can be found on the Moodle website of the MLcomm course.
- What you need to do:

[6] https://www.virtualbox.org/

Programming Environment (II)

- We provide a Ubuntu 64bit container for VirtualBox[6].
- The link can be found on the Moodle website of the MLcomm course.
- What you need to do:
  - Install VirtualBox.
  - Add a new virtual machine in VirtualBox (*Name*: mlcomm, *Type*: Linux, *Version*: Ubuntu 64bit, 8 GB RAM) and use the existing virtual hard disk file (`mlcomm.vdi`).
  - Start the virtual machine.
  - `Alt+F2` opens a command line:
    - `code` starts Visual Studio Code.
    - `terminal` starts a console, `python3` in there opens the python3 interpreter.

[6] `https://www.virtualbox.org/`

Programming Environment (III)

- We also offer Jupyter notebooks[7]:

  `http://141.40.254.115:9999`

- Password is

  `mlcomm`

- Live participation in this course!
- However: Notebooks will be deleted 1h after each lecture.

[7] `http://jupyter.readthedocs.io/en/latest/index.html`

Programming Environment (III)

- We also offer Jupyter notebooks[7]:

  `http://141.40.254.115:9999`

- Password is

  `mlcomm`

- Live participation in this course!
- However: Notebooks will be deleted 1h after each lecture.
- It is also installed in the virtual machine. Open the terminal and enter: `jupyter notebook`.

[7] `http://jupyter.readthedocs.io/en/latest/index.html`

Python Basics

Basics: Data Types

Python 3 supports the following data types:

- Integers (`int`)
- Floats (`float`)
- Booleans (`bool`)
- Strings (`str`)
- Lists (`list`)
- Tuples (`tuple`)
- Sets (`set`)
- Dictionaries (`dict`)

Basics: Data Types

Python 3 supports the following data types:

- Integers (`int`)
- Floats (`float`)
- Booleans (`bool`)
- Strings (`str`)
- Lists (`list`)
- Tuples (`tuple`)
- Sets (`set`)
- Dictionaries (`dict`)

The methods of each data type can be inspected via `help`, e.g., `help(int)`.

Basics: Data Types (Integer)

- Integer numbers (`int`):

```
>>> a = 5; type(a)
<type 'int'>
```

- Important methods: `.real()`, `.imag()`, `.conjugate()`

Basics: Data Types (Floats)

- Floating point numbers (`float`):

```
>>> a = 5.0; type(a)
<type 'float'>
```

- Important methods: `.real()`, `.imag()`, `.conjugate()`

Basics: Data Types (Booleans)

- Booleans (`bool`):

```
>>> a = True; type(a)
<type 'bool'>
```

Basics: Data Types (Strings)

- Strings (`str`):

```
>>> a = 'test'; type(a)
<type 'str'>
```

- Important methods: `.format()`, `.find()`, `.join()`, `.split()`.

Basics: Data Types (List)

- Lists (`list`):

```
>>> a = [1, 2, 3]; type(a)
<type 'list'>
```

- Sequence of arbitrary Python objects that can be modified (mutable) after it has been created.
- Builtin function `len()` returns the number of objects in the tuple.
- Important methods: `.append()`, `.extend()`, `.insert()`, `.remove()`.

Basics: Data Types (Tuples)

- Tuples (`tuple`):

```
>>> a = (1, 2, 3); type(a)
<type 'tuple'>
```

- Sequence of arbitrary Python objects that can not be modified (immutable) after it has been created.
- Builtin function `len()` returns the number of objects in the tuple.

Basics: Data Types (Sets)

- Sets (`set`):

```
>>> a = set((1,2,3,4)); type(a)
<type 'set'>
```

- A set object is an unordered, mutable collection of distinct Python objects, i.e., `set((1,2,3)) == set((3,2,1,1))`.
- Represents the mathematical concept of a set.
- Supports the associated mathematical operations `.intersection()`, `.difference()`, `.union()`.
- Builtin function `len()` returns the number of objects in the tuple.

Basics: Data Types (Dictionaries)

- Dictionary (`dict`):

```
>>> a = {'a': 1, 'b': 2, 'c': 3}; type(a)
<type 'dict'>
```

- Build dictionary from list of keys and values:

```
d = dict(zip(mykeys, myvals))
```

- Important methods: `.keys()`, `.values()`, `.items()`.

Basics: Summary Data Types

- Each data type is implemented as an object.
- Checking whether object is of a given type:

```
>>> isinstance('mystring', str)
True
>>> isinstance(2.0, int)
False
```

Basics: Printing and Formatting (I)

- Printing is done via the `print()` function.
- Each string has a corresponding `format()` method.

```
>>> print('Hello {}!'.format('Fabian'))
Hello Fabian!
>>> print('Hello {} {}!'.format('Fabian', 'Steiner'))
Hello Fabian Steiner!
>>> print('Hello {firstname} {lastname}!'.
  format(firstname='Fabian', lastname='Steiner'))
Hello Fabian Steiner!
```

Basics: Printing and Formatting (II)

- Similar to C's `printf()`, several format specifiers are supported.
- Syntax of format specifiers: `:<field_width>.<precision><data_type>`

```
>>> print('Num: {:d}'.format(2))
Num: 2
>>> print('Num: {:10d}'.format(2))
Num:          2
>>> print('Num: {:10d}'.format(223))
Num:        223
>>> print('Num: {:.3f}'.format(3.14159))
Num: 3.142
```

Basics: List comprehension

- Create lists from existing lists or an iterable object.

```
y = [x**2 for x in xrange(1, 10)]
```

- This can be combined with conditions.

```
y = [x**2 for x in mylist if x  % 2 == 0]
```

- A n-dimensional extension is possible.

```
z = [x*y for x in mylist1 for y in mylist2]
```

Basics: Generators (I)

In many cases, a new list should not be generated explicitely, because the individual list members are not needed. Hence, memory can be saved.

- Traditional

```
res = sum([xval**2 for xval in x])
```

- Generator

```
res = sum((xval**2 for xval in x))
```

Basics: Generators (II)

More elaborate example:

```
def gen_combs(set1, set2):
  l = []
  for s1 in set1:
    for s2 in set2:
      l.append((s1,s2))
  return l
```

```
def gen_combs(set1, set2):
    for s1 in set1:
        for s2 in set2:
            yield (s1, s2)
```

```
>>> for item in gen_combs((1,2,3),(4,5,6)):
    print(item)
```

Basics: Defining functions

```
def mysum(arg1, arg2):
    result = arg1 + arg2
    return result
```

- Identation is important: Use four spaces.
- Automatic cleanup tool: `autopep8`.
- Functions help to structure your program and to write reusable code.

Basics: Organizing your code (I)

- To avoid clogging your namespace, put your code into separate files and import them if required.
- Example: File `ml_tools.py` with function `entr()`.

```
>>> import ml_tools
>>> H = ml_tools.entr([0.3, 0.7])
```

Basics: Organizing your code (I)

- To avoid clogging your namespace, put your code into separate files and import them if required.
- Example: File `ml_tools.py` with function `entr()`.

```
>>> import ml_tools
>>> H = ml_tools.entr([0.3, 0.7])
```

- If the import name is too long, it can be abbreviated by `import longname as ln`.
- For larger code bases, modules are more appropriate.
- For this, we first create the module folder `mlcomm`, mark it as a module for Python by placing an empty `__init__.py` file and then add the corresponding files.

Basics: Organizing your code (II)

An example module may look like:

```
mlcomm
├── __init__.py
├── tools
│   ├── __init__.py
│   └── it.py
└── em
    ├── __init__.py
    └── em.py
```

The individual parts can be imported as (inspect your namespace with `dir()` afterwards):

```
>>> from mlcomm import em
>>> import mlcomm.tools.it
```

Basics: LBYL vs. EAFP

LBYL

Look before you leap.

Basics: LBYL vs. EAFP

LBYL

Look before you leap.

EAFP

Easier to ask for forgiveness than permission.

Basics: LBYL vs. EAFP

LBYL

Look before you leap.

EAFP

Easier to ask for forgiveness than permission.

Python's paradigm follows the EAFP style:

```
>>> d = {'name': ['Peter', 'George'], 'age': [20, 30]}
>>> try:
...     places = d['places']
... except KeyError:
...     print('No key named places')
...     places = None
No key named places
```

Basics: Unit Tests (I)

- It's a good practice in software engineering to follow a test-driven development cycle.
- Each function/module/etc. should be tested exhaustively. After each update to a file, the tests should be re-run to ensure that it is still working correctly.
- In particular, corner/pathological cases should be checked carefully.
- In Python, this can be guaranteed with the `unittest` framework[8].

[8] `https://docs.python.org/3.6/library/unittest.html`

Basics: Unit Tests (II)

We want to write a unittest for the function `mysum()` that adds its two inputs.

```
import unittest
import mysum

class TestSum(unittest.TestCase):
    def test_sum(self):
        self.assertEqual(mysum(2, 3), 5)
        self.assertEqual(mysum(0, 5), 5)
        self.assertEqual(mysum(-3, 3), 0)
        self.assertRaises(TypeError, mysum, 'a', 3)
        # add more here

if __name__ == '__main__':
    unittest.main()
```

Libraries for Numerical Computations

NumPy

- NumPy is the fundamental package for numerical computing with Python.
- It provides
  - functions for dealing with n-dimensional arrays,
  - various mathematical functions,
  - a Matlab-like interface.
- NumPy uses 0-based indexing.
- NumPy assigns by reference.
- Import NumPy into your code as

```
>>> import numpy as np
```

NumPy: Arrays (I)

- Create matrix $\begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}$:

```
>>> a = np.array([[1, 2], [3, 4]])
```

- Index single element

```
>>> a[0,1]
```

- Index first row:

```
>>> a[0,:]
```

- Index first column:

```
>>> a[:,0]
```

NumPy: Arrays (II)

- Get size

```
>>> a.shape
```

- Get number of elements

```
>>> a.size
```

- Vertically concatenate the arrays `a` and `b`:

```
>>> c = vstack((a,b))
```

- Horizontally concatenate the arrays `a` and `b`:

```
>>> c = hstack((a,b))
```

NumPy: Arrays (III)

- Serialize array

```
>>> a.flatten()
```

- Create zero 3×3 matrix

```
>>> a = np.zeros((3,3))
```

- Create 3×3 all ones matrix

```
>>> a = np.ones((3,3))
```

- Create 3×3 identity matrix

```
>>> a = np.eye(3)
```

NumPy: Arrays (IV)

- Create list of values ranging from 1.0 to 4.9 in step sizes of 0.1.

```
>>> a = np.arange(1.0,5.0,0.1)
```

- Transpose.

```
>>> a.T
```

- Conjugate transpose, i.e., Hermitian.

```
>>> a.conj().T
```

NumPy: Linear Algebra

- Innner product of two 1D vectors `a` and `b`.

```
>>> a.dot(b)
```

- Matrix-vector product of matrix `A` and vector `b`.

```
>>> A.dot(b)
```

- Matrix-matrix product of matrix `A` and matrix `B`.

```
>>> A.dot(B)
```

component wise-multiplication:
A*B

NumPy: Random Numbers

- Create vector of n normally distributed random numbers:

```
>>> N = np.random.rand(n)
```

- Create vector of n uniformly distributed, integer random numbers between `lb` and `ub`:

```
>>> N = np.random.randint(lb, ub + 1)
```

- For more, see `help(np.random)`.

NumPy: Passing by reference (I)

```
>>> a = np.array([[1,2], [3,4]])
>>> a
array([[1, 2],
       [3, 4]])
>>> b = a[:,0]
>>> b
array([1, 3])
>>> b[:] = 8
>>> a
array([[8, 2],
       [8, 4]])
```

NumPy: Passing by reference (II)

- If real copies are needed:

```
>>> a = np.array([[1,2], [3,4]])
>>> b = a.copy()
>>> c = a[:,0].copy()
```

NumPy: Importing Data

- Read simple text files:

```
>>> data = np.loadtxt('filename.txt')
```

- Save simple text files:

```
>>> data = np.savetxt('filename.txt')
```

- Detailed reference of all parameters can be found online[9].

[9] https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.loadtxt.html

NumPy: Importing Data

- Read simple text files:

```
>>> data = np.loadtxt('filename.txt')
```

- Save simple text files:

```
>>> data = np.savetxt('filename.txt')
```

- Detailed reference of all parameters can be found online[9].
- Read Matlab files:

```
>>> data = scipy.io.loadmat('filename.mat')
```

[9] https://docs.scipy.org/doc/numpy-1.13.0/reference/generated/numpy.loadtxt.html

NumPy: Plotting (I)

- Import the necessary functionality:

```
>>> import matplotlib.pyplot as plt
```

- Generate data and plot:

```
>>> x = np.linspace(1,10,10)
>>> y = 2*x
>>> plt.plot(x, y)
>>> plt.show()
```

- Result can be saved with

```
>>> plt.savefig('fig.png')
```

NumPy: Plotting (II)

- Exposed interface is similar to the Matlab plotting functionality.
- If logarithmic plots are desired:
  - `plt.semilogx(x,y)`
  - `plt.semilogy(x,y)`
  - `plt.loglog(x,y)`
- The axis can be modified via
  - `plt.xlabel('X-Label')`
  - `plt.ylabel('Y-Label')`
  - `plt.xlim((0, 10))`
  - `plt.ylim((0, 10))`

NumPy: Outlook

- Full NumPy reference[10].
- Guide for users transitioning from Matlab[11].
- Use `timeit` module for benchmarking[12] small snippets of your code.
- Further information on improving NumPy performance[13].

[10] `https://docs.scipy.org/doc/numpy/reference/`

[11] `https://docs.scipy.org/doc/numpy-dev/user/numpy-for-matlab-users.html`

[12] `https://docs.python.org/2/library/timeit.html`

[13] `http://ipython-books.github.io/featured-01/`

SciPy

What's the relation[14] of SciPy and NumPy?

"In an ideal world, NumPy would contain nothing but the array data type and the most basic operations: indexing, sorting, reshaping, basic elementwise functions, et cetera. All numerical code would reside in SciPy. However, one of NumPy's important goals is compatibility, so NumPy tries to retain all features supported by either of its predecessors. Thus NumPy contains some linear algebra functions, even though these more properly belong in SciPy. [...]"

[14] `https://www.scipy.org/scipylib/faq.html#id16`

SciPy

- The SciPy module therefore contains the actual numerical algorithms.
- Import module as

```
>>> import scipy as sc
```

  - `sc.integrate`: Numerical integration, quadrature rules.
  - `sc.optimize`: Constrained/unconstrained optimization algorithms, root finding.
  - `sc.linalg`: Supersedes `np.linalg`.
  - `sc.stats`: Implements various distributions, their PDFs, CDFs and moments.

SciPy

- The SciPy module therefore contains the actual numerical algorithms.
- Import module as

```
>>> import scipy as sc
```

  - `sc.integrate`: Numerical integration, quadrature rules.
  - `sc.optimize`: Constrained/unconstrained optimization algorithms, root finding.
  - `sc.linalg`: Supersedes `np.linalg`.
  - `sc.stats`: Implements various distributions, their PDFs, CDFs and moments.
- Instead of re-inventing the wheel (numerical algorithms can be super hard to implement reliably!), use the provided ones.
- But: Make always sure that they actually implement what you would like to have.

Pandas

- Machine learning is closely associated with “big data”.
- Before being able to work with big data, you first have to get it into Python.

Pandas

- Machine learning is closely associated with “big data”.
- Before being able to work with big data, you first have to get it into Python.
- Pandas provides convenient abstraction layers for handling data.
  - Reading and writing spreadsheets.
  - Sorting and viewing data.
  - Database-like access: joins, groups, pivoting.

Outlook

Outlook

- A lot of topics could not be covered today.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.
  - Concurrent execution.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.
  - Concurrent execution.
  - Object oriented programming: concept of objects and classes.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.
  - Concurrent execution.
  - Object oriented programming: concept of objects and classes.
  - Virtual environments.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.
  - Concurrent execution.
  - Object oriented programming: concept of objects and classes.
  - Virtual environments.
  - Extensions with own C modules.

Outlook

- A lot of topics could not be covered today.
  - Exception handling.
  - Database interaction.
  - Filesystem access.
  - Concurrent execution.
  - Object oriented programming: concept of objects and classes.
  - Virtual environments.
  - Extensions with own C modules.
- Play around yourself, write code and discuss with your colleagues.

Homework

Homework I

The purpose of this first homework is to familiarize yourself with Python and to recap some of the basics that have been introduced.

1. Setup a git repository named `mlcomm`.
2. Make a Python module out of it and set up the required directory structure. It should have submodules for each of the five course subjects and additional "tools" and "tests" folder: `nn`, `usc`, `pgm`, `var`, `dr`, `tools`, `tests`.
3. Implement a function with the signature `mlcomm.tools.it.discrete_entr`

   ```
   def discrete_entr(pX): pass
   ```

   that calculates the entropy of the provided distribution `pX`. Take care of a proper error checking and write a unit test. The entropy is defined as

$$\sum_{x \in \mathrm{supp}(P_X)} -P_X(x) \log_2(P_X(x)).$$

Homework II

4. Implement a function with the signature
   `mlcomm.tools.it.discrete_cross_entr`

```
def discrete_cross_entr(pX, pY): pass
```

   that calculates the cross-entropy of the distributions `pX` and `pY`. Take care of a proper error checking and write a unit test. The cross entropy is defined as

$$\sum_{x \in \mathrm{supp}(P_X)} -P_X(x) \log_2(P_Y(x)).$$

Homework III

5. Implement a function with the signature
   `mlcomm.tools.it.discrete_kl_dis`

```
def discrete_kl_dis(pX, pY): pass
```

   that calculates the Kullback-Leibler divergence of the distributions `pX` and `pY`. Take care of a proper error checking and write a unit test. The Kullback-Leibler divergence is defined as

$$\sum_{x \in \mathrm{supp}(P_X)} P_X(x) \log_2 \left(\frac{P_X(x)}{P_Y(x)} \right).$$

Homework IV

6. Implement a function with the signature `mlcomm.nn.utils.act_fct`

```
def act_fct(x,type_fct): pass
```

that returns the value of different activation functions evaluated at `x` depending on the `type_fct` parameter:

- Identity: $y = f(x) = x$.
- Sigmoid: $y = f(x) = \frac{1}{1+\mathrm{e}^{-x}}$.
- Tanh: $y = f(x) = \tanh(x)$.
- Rectified linear unit: $y = f(x) = \max(0, x)$.