Introduction to Data Analysis and Machine Learning in Physics

Jörg Marks, Klaus Reygers

11-14 April 2023
9:00 - 12:00 and 14:00 - 17:00

Outline

- **Day 1**
  - Introduction, software and data fitting
- **Day 2**
  - Machine learning - basics
- **Day 3**
  - Machine learning - decision trees
- **Day 4**
  - Machine learning - convolutional networks and graph neural networks
- **Organization** and **Objective**
  - 2 ETC: Compulsory attendance is required
    Active participation in the exercises
  - Course in CIP pool in a tutorial style
  - Obtain basic knowledge for problem-oriented self-studies

Course Information (1)

- Course requirements
  - Python knowledge needed / good C++ knowledge might work
  - Userid to use the CIP Pool of the faculty of physics
- Course structure
  - Course in CIP pool using the jupyter3 hub
  - Lectures are interleaved with tutorial/exercise sessions in small groups (up to 5 persons / group)
- Course homepage which includes and distributes all material

  `https://www.physi.uni-heidelberg.de/~reygers/lectures/2023/ml/`

/transparencies	Transparencies of the lectures
/examples	iPython files shown in the lectures
/exercises	Exercises to be solved during the course
/solutions	Solutions of the exercises

Course Information (2)

`TensorFlow` and `Keras` are now also installed in the CIP jupyter hub. In addition, with a google account you can run jupyter notebooks on Google Colab:

`https://colab.research.google.com/`

Missing python libraries can be included by adding the following to a cell (here for the pypng library):

```
!pip install pypng
```

Course Information (3)

- Your installation at home:
  - Web Browser to access jupyter3
  - Access to the CIP pool via an ssh client on your home PC
- No requirements for a special operating system
- Software:
  - firefox or similar
  - Cisco AnyConnect
  - ssh client (MobaXterm on Windows, integrated in Linux/Mac)
- Local execution of python / iPython
  - Install `anaconda3` and download / run the iPython notebooks (also python scripts are available)
- Hints for software installations and CIP pool access

  `https://www.physi.uni-heidelberg.de/~marks/root_einfuehrung/Folien/CIPpoolAccess.PDF`

Course Information (4)

Alternatively, you can install the libraries needed on your local computer.

Here are the relevant instruction for macOS using `pip`:

Assumptions: `homebrew` is installed.
Install python3 (see https://docs.python-guide.org/starting/install3/osx/)

```
$ brew install python
$ python --version
Python 3.8.5
```

Make sure pip3 is up-to-date (alternative: conda → don't mix conda and pip installations)

```
$ pip3 install --upgrade pip
```

Install modules needed:

```
$ pip3 install --upgrade jupyter matplotlib numpy pandas
scipy scikit-learn xgboost iminuit tensorflow tensorflow_datasets Keras
```

Topcics and file name conventions

0. Introduction (this file) (introduction.pdf)
1. Introduction to python (01_intro_python_*)
2. Data modeling and fitting (02_fit_intro_*)
3. Machine learning basics (03_ml_intro_*)
4. Decisions trees (04_decision_trees_*)
5. Neural networks (05_neural_networks_*)

Programm Day 1

- Technicalities
- Summary of NumPy
- Plotting with matplotlib
- Input / output of data
- Summary of pandas
- Fitting with iminuit and PyROOT
- Transparencies with activated links, examples and exercises
  - Software: 01_intro_python.pdf
  - Fitting: 02_fit_intro.pdf

Programm Day 2

- Introduction to machine learning 03_ml_intro.pdf
  - Tensorflow / Keras, datasets
  - Supervised learning
  - Classification

- Multivariate analysis 03_ml_intro_mva.pdf
  - Regression
  - Linear regression
  - Logistic regression
  - Softmax regression (multi-class classification)

Programm Day 3

- Decision trees
- Bagging and boosting
- Random forest
- XGBoost

04_decision_trees.pdf

Programm Day 4

- Neural networks
- Convolutional neural networks
- Hand-written digit recognition with Keras

05_neural_networks.pdf