Introduction to Data Analysis and Machine Learning in Physics

Martino Borsato, Jörg Marks, Klaus Reygers

6-9 April 2021
9:00 - 12:00 and 14:00 - 17:00

Outline

- **Day 1**
  - Introduction, software and data fitting
- **Day 2**
  - Machine learning - classification
- **Day 3**
  - Machine learning - decision tree
- **Day 4**
  - Machine learning - convolutional networks
- **Organization** and **Objective**
  - 2 ETC: Compulsory attendance is required, checked within heiCONF active participation in the exercises
  - Online course in 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
  - Online Course using the jupyter2 hub of the CIP Pool and heiCONF for communication
  - 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/2021/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)

- Your installation at home:
  - Web Browser to access heiCONF and jupyter2
  - Headset and Webcam for communication via heiCONF
  - (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
- Hints for software installations and CIP pool access
  `https://www.physi.uni-heidelberg.de/~reygers/lectures/2021/ml/transparencies/CIPpoolAccess.PDF`

Course Information (3)

Alternatively, you can install the needed libraries 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)

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

Install needed modules:

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

Course Information (4)

`TensorFlow` and `Keras` are not installed on the CIP jupyter hub. With a google account you can run jupyter notebooks with these libraries on Google Colab:

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

One can install missing python libraries by adding the following to a cell (here for the pypng library):

```
!pip install pypng
```

heiCONF Access

We will have a lecture room and 10 break out rooms. In the break out rooms all participants are moderators. Here, the screens of the small groups should be shared in order to design programms and discuss solutions.

Main room:

lectures[1]

Break out rooms:

room 1	room 2	room 3	room 4	room 5
room 6	room 7	room 8	room 9	room 10

Technical hints:

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

[1] All rooms have activated heiCONF links.

Topcics and file name conventions

1. Introduction to python (01_intro_python_*)
2. Data modeling and fitting (02_fit_intro_*)
3. Machine learning basics (03_ml_basics_*)
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

- Supervised learning
- Classification and regression
- Linear regression
- Logistic regression
- Softmax regression (multi-class classification)

Programm Day 3

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

Programm Day 4

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