Outline “Python for Machine Learning”, PhD course

The course allows PhD and advanced Master’s students to start their own empirical Python projects and introduces various Machine Learning models. It also teaches some highly relevant python packages, version control, project management. Each afternoon (except on the first day) the students work on their own applied exercises - the course is thus totally applied.

Students with no or little Python experience have to complete a self-paced, online and interactive tutorial. I recommend one of the two

1. datacamp.com (“Intro to Python for Data Science”, Sessions 1, 2, 3)
2. codeacademy.com (“Learn Python 2”, Sessions 1, 2, 3, 4, 5, 7, 8)

Outline

First unit: Getting python started, Working with Pandas

- Basic concepts in Python (recap from pre-course)
- pandas as key module for data wrangling in Python
- Plotting with matplotlib

Second unit: Seaborn, Getting data, debugging

- Statistical plotting with seaborn
- Examples of RESTful APIs
- How to debug

Third unit: project management

- Version control with git
- Collaborative development with GitHub
- Project organization
- Script style

Fourth unit: Unsupervised ML and record-matching and deduplication with dedupe

- Scaling
- PCA
- k-means clustering, agglomerative clustering and DBSCAN
- dedupe as applied tool to match or deduplicate large structured data

Fifth day: Supervised ML

- Introduction on ML, general workflow for supervised ML, model evaluation
- K-nearest neighbors, linear models (ridge and lasso), Decision Tree, Random Forest, Neural Networks
- Grid Search and Cross-Validation with Pipelines

Sixth session: Natural Language Processing

- Preprocessing and vectorizing
- Topic Modelling with LDA and model evaluation
- Simple text analysis (Word clouds, sentiment analysis, noun phrases, readability)

Literature

- Downey, Allen B.: “How to think like a Computer Scientist”
- Steven Bird, Ewan Klein, Edward Loper: "Natural Language Processing with Python"
- Gentzkow, M., B. Kelly and M. Taddy: “Text as Data”

Teacher

- Michael E. Rose, Post Doctoral Researcher at MPI for Innovation and Competition
- Experience in teaching ML and Big Data in Python to PhD students (Georgia Tech and LMU), Computational Mathematics in Matlab to Master’s students, Time Series Econometrics in Python to Master’s students, SQL, Excel and VBA to Master’s students
- Daily usage of Python, ML in own research
- Lead development of two python packages (scopus and sosia)