Big Data and Machine Learning
Syllabus

Fabian Slonimczyk[1]

December 29, 2018

1 Lectures and Office Hours

- Lectures:
- Labs:
- Office Hours: Fridays 15:00–17:00. Room 3222

2 Course description

Big Data and Machine Learning (M.Sc. level) is an advanced elective course designed for masters students at ICEF. The main objective of the course is to endow students with fundamental skills related to data mining and analytics, as well as with designing and implementing machine learning predictive models. The course is open to all second year M.Sc. students. Basic knowledge of the Python programming language is strongly advised but not required. Students without Python knowledge will be expected to exert additional effort during the first few weeks of the course to catch up. The course is taught in English.

The course has three broad sections:

I. Building skills using Python libraries to solve common problems in the analysis offinancial data.

II. Learning how to mine the web, social media, and other big data sources in searchfor data.

III. Designing and implementing machine learning models.

3 Teaching methods

The following methods and forms of study are used in the course:

- Lectures: description and explanation of techniques, algorithms, and tools

[1] fabian.slonimczyk@gmail.com

- Practice in computer lab: solving problems applying what was explained in lecture

  • Self-study in computer lab: doing home assignments using Python

- Self-study with literature and the internet: in depth research, specially related to the final project

4 Assessment

1. Attendance to lectures
2. Homework Assignments
3. Final Project

Grade determination

The main form of evaluation is the final project. In particular, only students who show evidence of having spent sufficient time and energy developing their project can expect an excellent grade.

Students will be expected to make a short project proposal by the end of week 3 (10% of final grade). An intermediate report after week 8 (20%) and a final report and presentation during the last meeting of the semester (30%).

The final grade is also partly determined by attendance and participation during lectures (10%), and home assignments (30%).

5 Readings

Main Readings

Each week's assigned readings will be coming mostly from the following books (see the course outline below for details):

P4DA Wes McKinney (2018). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. 2nd Ed., O'Reilly Media.

P4Fin Yves Hilpisch (2014). *Python for Finance: Analyze Big Financial Data*. 1st Ed., O'Reilly Media.

Mine J. Han, M. Kamber, and J. Pei (2012). Data Mining: Concepts and Techniques. 3rd Ed., Morgan Kaufmann.

TML Matthew Kirk (2017). *Thoughtful Machine Learning with Python*. 1st Ed., O'Reilly Media.

HOML Aurlien Gron (2017). *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. 1st Ed., O'Reilly Media.

Additional Readings

For some specific topics, I will assign chapters from the following:

Yan Yuxing Yan (2017). *Python for Finance: Apply powerful finance models and quantitative analysis with Python*. 2nd Ed., Packt Publishing.

Big James Ma Weiming (2015). *Mastering Python for Finance*. 1st Ed., Packt Publishing.

- Megan Squire (2016). *Mastering Data Mining with Python: Find patterns hidden in your data*. Packt Publishing.

Soc1 S. Chatterjee and M. Krystyanczuk (2017). *Python Social Media Analytics: Analyze and visualize data from Twitter, Youtube, GitHub, and more*. Packt Publishing.

Soc2 Matthew A. Russell and Mikhail Klassen (2019). *Mining the Social Web Data Mining Facebook Twitter LinkedIn Instagram*. 3rd Ed. O'Reilly Media.

PML Dipanjan Sarkar, Raghav Bali and Tushar Sharma (2018). *Practical Machine Learning with Python*. Apress.

Internet Resources and Databases

- The home page of matplotlib: http://matplotlib.org
- McKinney's Python for Data Analysis repository: https://github.com/wesm/ pydata-book
- Hilpisch's Python for Finance repository: https://github.com/yhilpisch/py4fi
- Yan's Python for Finance repository: https://github.com/PacktPublishing/ Python-for-Finance-Second-Edition

6 Basic Course Outline

Part I. Finance Analytics using Python

Week 1. Introduction to Python
Introductory examples. Basic data types and structures. Functions and control flow. Classes and objects. Functional programming and Object-oriented programming. Readings: *P4Fin*, Chapters 1–4.

Week 2. Python's Scientific Stack: NumPy, Pandas, and SciPy
NumPy arrays and code vectorization. Indexing and slicing. Series and DataFrame objects. Basic data wrangling: missing data, transformed data, reshaped data, merged data. Data summarization.
Readings: *P4Fin*, Chapter 6. *P4DA*, Chapters 4–5.

Week 3. Data Visualization

Plots in two and three dimensions. The matplotlib and the seaborn libraries. Readings: *P4Fin*, Chapter 5. *P4DA*, Chapter 9.

Weeks 4 and 5. Financial and Other Applications

The value of money. A financial calculator. Bond and stock valuation. The CAPM. Readings: *Yan*, Chapters 3, 5, and 6.

Week 6. Mathematical tools and numerical calculus

Finding roots. Convex optimization. Monte Carlo simulations. Portfolio optimization. Readings: *P4Fin*, Chapters 9–11.

Part II. Big Data

Week 7. Big Data

Accessing regular data with Python. How is big data different? Apache Hadoop and the HDFS data format. SQL and No-SQL databases. Readings: *P4Fin*, Chapter 7. *Big*, Chapter 7.

Week 8. Introduction to Data Mining

Data mining methodologies.
Readings: *Mine*, Chapters 1–4.

Weeks 9 and 10. Mining the Social Web

Restful APIs. Mining Twitter, Facebook, Instagram, and GitHub to learn what's trending.
Readings: *Soc1*, Chapters 1–3. *Soc2*, Chapters 1–3.

Week 11. Textual Analysis

Mining Text Files: Computing Document Similarity, Extracting Collocations. Readings: *Soc2*, Chapters 4–5.

Part III. Machine Learning

Weeks 12 and 13. Machine Learning Classification Methods

Gradient descent. Cross Validation. K-nearest neighbor. Naive Bayesian Classification. Decision Trees.
Readings: *TML*, Chapters 1–5.

Week 14. Machine Learning Regression Methods

OLS, LASSO, Ridge regression.
Readings: *HOML*, Chapters 3–4.

Weeks 15 and 16. Neural Networks. Forecasting Stock and Commodity Prices

Comparing traditional time series analysis and Recursive Neural Nets. Readings: *TML*, Chapter 8. *PML*, Chapter 11.

7 Distribution of Hours

Unit	Topic title	Lectures	Contact hours	Self-study
1	Introduction to Python	1	4	4
2	Python's Scientific Stack	1	4	4
3	Data Visualization	1	4	4
4	Financial Applications	2	8	8
5	Math tools	1	4	4
6	Big Data	1	4	4
7	Intro to Data Mining	1	4	4
8	Mining the Social Web	2	8	8
9	Textual Analysis	1	4	4
10	ML Classification Methods	2	8	8
11	ML Regression Methods	1	4	4
12	Neural Networks	2	8	8
	Total	16	64	64