CS-401

Applied data analysis

Catasta Michele

Cursus	Sem.	Type
Computer science	MA1, MA3	Opt.
Digital Humanities	MA1	Obl.
SC master EPFL	MA1, MA3	Opt.

Language of teaching	English
Credits	6
Session	Winter
Semester	Fall
Exam	Written
Workload	180h
Weeks	14
Hours	**4 weekly**
Courses	2 weekly
Project	2 weekly
Number of positions	

Summary

This course teaches the basic techniques and practical skills required to make sense out of a variety of data, with the help of the most acclaimed software tools in the Data Science world: pandas, scikit-learn, Spark, TensorFlow, etc.

Content

Thanks to a new breed of software tools that allows to easily process and analyze data at scale, we are now able to extract invaluable insights from the vast amount of data generated daily. As a result, both the business and scientific world are undergoing a revolution which is fueled by one of the most sought after job profiles: the data scientist.
This course covers the fundamental steps of the Data Science pipeline:
Data Acquisition

- Variety as one of the main challenges in Big Data: structured, semi-structured, unstructured
- Data sources: open, public (scraping, parsing and down-sampling)
- Dataset fusion, filtering, slicing & dicing
- Data granularities and aggregations

Data Wrangling

- Data manipulation, array programming, dataframes
- The many sources of data problems (and how to fix them): missing data, incorrect data, inconsistent representations
- Schema alignment, data reconciliation
- Data quality testing with crowdsourcing

Data Interpretation

- Stats in practice (distribution fitting, statistical significance, etc.)
- Co-occurrence grouping (market-basket analysis)
- Machine learning in practice (supervised and unsupervised, feature engineering, more data vs advanced algorithms, curse of dimensionality, etc.)
- Text mining: vector space model, topic models, word embedding
- Profiling (fraud / anomaly detection)
- Social Network Analysis (influencers, community detection, etc.)

Data Visualization

- Introduction to different plot types (1, 2 and 3 variables), layout best practices, network and geographical data
- Visualization to diagnose data problems, scaling visualization to large datasets, visualizing uncertain data

Reporting

- Results reporting, infographics
- How to publish reproducible results
- Anonymiziation, ethical concerns

The students will learn the techniques during the ex-cathedra lectures, and will then get familiar with the software tools to complete the homework assignments (which will be in part executed under the supervision of the teacher and the assistants, during the lab hours).
In parallel, the students will embark in a semester-long project, split in agile teams of 2-3. The outcome of such team efforts will be unified towards the end of the course, to build a project portfolio that will be made public (and available as open-source).
At the end of the semester, students will also take a 3h final exam in a classroom with computers, where they will be asked to complete a data analysis pipeline (both with code and extensive comments) on a dataset they have never worked with before.

Keywords

data science, data analysis, data mining, machine learning

Learning Prerequisites

Required courses

The student MUST have passed an introduction to databases course, OR a course in probability & statistics, OR two separate courses that include programming projects.

Recommended courses

- CS-423 Distributed Information Systems
- CS-433 Pattern Classification and Machine Learning

Important concepts to start the course

Algorithms, object oriented programming, basic probability and statistics

Learning Outcomes

By the end of the course, the student must be able to:

- Construct a coherent understanding of the techniques and software tools required to perform the fundamental steps of the Data Science pipeline
- Perform data acquisition (data formats, dataset fusion, Web scrapers, Rest APIs, Open Data, Big Data platforms, etc.)
- Perform data wrangling (fixing missing and incorrect data, data reconciliation, data quality assessments, etc.)
- Perform data interpretation (statistics, knowledge extraction, critical thinking, team discussions, ad-hoc visualizations, etc.)
- Perform result dissemination (reporting, visualizations, publishing reproducible results, ethical concerns, etc.)

Transversal skills

- Evaluate one's own performance in the team, receive and respond appropriately to feedback.

- Give feedback (critique) in an appropriate fashion.
- Demonstrate the capacity for critical thinking
- Write a scientific or technical report.

Teaching methods

- Physical in-class recitations and lab sessions
- Homework assignments
- Course project

Expected student activities

Students are expected to:

- Attend the lectures and lab sessions
- Complete a weekly homework assignment
- Read / watch the pertinent material before a lecture
- Engage during the class, and present their results in front of the other colleagues

Assessment methods

- 30% continuous assessment during the semester (homework)
- 30% final exam, data analysis task on a computer (3h)
- 40% final project, done in groups of 2-3

Supervision

Office hours	Yes
Assistants	Yes
Forum	Yes
Others	http://ada.epfl.ch

Resources

Websites

- http://ada.epfl.ch