CS-C3250 Data Science Project Autumn 2021

Jorma Laaksonen

13.9.2021: Introduction, arrangements, topics

Teachers

- Jorma Laaksonen, D.Sc.(Tech.), Docent, Senior University Lecturer
- Janne Sinkkonen, D.Sc.(Tech.), Senior Data Scientist, Reaktor
- Saku Suuriniemi, D.Sc.(Tech.), Data Scientist, Reaktor
- Henrik Aalto, M.Sc.(Tech.), Data Scientist, Reaktor
- Vilen Looga, D.Sc.(Tech.), Senior Data Scientist, Futurice
- Juha Vesanto, D.Sc.(Tech.), Principal Data Scientist, OP Financial Group
- Alena Shchevyeva, Mylinh Nguyen, Thong Tran, Letizia Iannucci
- Use LinkedIn to connect with us!
- Jorma’s tg: @jormalaaksonen, room: B326

Learning outcomes

- After the course, you can work as a data scientist in a team.
- You understand the structure and technical and non-technical challenges of data science projects.
- Furthermore, you learn to apply data analysis tools in a real-world data analysis project.
- Finally, you learn to document your project work and its outcomes, and to present its results and conclusions thereof both in writing and verbally.

Contents

- The course consists of a data science project which will be done in a small group for a real client from industry or academia.
- The activities include:
  - project management
  - requirements specification
  - design
  - coding
  - data collection
  - data curation
  - data storing
  - experimenting
  - documentation
  - presenting

Workload

- 5 x 2-hour lectures (1 by Jorma Laaksonen, 4 by visitors)
- 10 x 2-hour supervised group meetings
- 1 x 2-hour final presentations for other groups and the companies
- participation in all above is **mandatory** and can be attended **remotely**
- 3 x 2-hour **physical** visits to the companies **(possibly, non-mandatory)**
- **rest**: group and individual work, to be reported in MyCourses
- 135 hours total

Assessment methods

- Outcome of the project work done in a group of 6–7 students
- Presentation of the project work in the group
- Documentation of the project in the group’s Final report
- Progress reports in MyCourses
- Self- and peer-reported activity
- Grade 0–5 by default the same for the whole group, but exceptions possible

Final report (to be specified more in detail)

- Written by the group
- Summarize the data, methods and results of your project work
- Assess what was easy and what was difficult in the project
- Explain how the group worked as individuals and as a whole

Lectures Mondays online 12:15-14:00

- L1 13.9. Jorma Laaksonen: *introduction, arrangements, project topics*
- L2 20.9. Letizia Iannucci: *python, numpy, pandas, scikit*
- L3 27.9. Vilen Looga: *Data science toolbox*
- L4 4.10. Juha Vesanto: *Privacy & ethics*
- L5 11.10. Janne Sinkkonen: *Cases & observations*

Group meetings online Thursdays 12:15-14:00

- G1 16.9. Company representatives: *project kick-off* (G2: +17.9. 10:00-11:00)
- G2 23.9. Teaching assistants: *progress monitoring*
- G3 30.9. Teaching assistants: *progress monitoring*
- G4 7.10. Company representatives: *2nd company meeting*
- G5 14.10. Teaching assistants: *progress monitoring*
- G6 21.10. Teaching assistants: *progress monitoring*
- G7 4.11. Company representatives: *3rd company meeting*
- G8 11.11. Teaching assistants: *progress monitoring*
- G9 18.11. Teaching assistants: *progress monitoring*
- G10 25.11. Teaching assistants: *progress monitoring*
- + your own group meetings !
- + decide whether/when/how you meet physically/remotely

Visits to the companies

- Mon 15.11. *possibly to one of the companies*
- Mon 22.11. *possibly to one of the companies*

Final presentations

- Mon 29.11. *final presentations* (**tentative**)
- Thu 2.12. *final presentations* (**tentative)**

Project reports – one per group

- Fri 26.11. Deadline for the first version – TAs will give you feedback
- Thu 9.12. Deadline for the final version
- TAs, Jorma & company representatives will assess the reports

Group 1 Reaktor: Communicating climate change

- Company representatives: Janne Sinkkonen, Saku Suuriniemi, Henrik Aalto
- TA: Alena Shchevyeva tg: @aishchev
- Bùi, Hà My (not present)
- Kucheria, Aayush
- Nguyen, Thi Minh Nguyet (not present)
- Nguyen, Khue
- Niva, Verna
- Sauer, Hanne
- Wojnicki, Mikolaj

Group 1 Reaktor: Communicating climate change

- Climate change, mainly because of CO2 and methane emissions, currently looks like one of humanity's largest challenges
- An imaginary client who would like to have an interactive web page to communicate the emission reduction momentum to Nordic audience
- This includes *finding data*, this time from public sources rather than from an organization, *understanding what is important and has impact, modeling and visualization, and planning and building an implementation* with continuity in mind
- Let's concentrate on past temperatures first, because observations are more convincing than forecasts
- Use public sources, such as the World Bank, Berkeley Earth, NASA GISS, ...

Group 2 Futurice:

- Company representative: Vilen Looga
- TA: Linh Nguyen tg: @tienrang
- Hallonbacka, Matthew
- Le, Son
- Nguyen, Long
- Pham, Binh
- Salehi, Hafsa
- Strozanski, Pawel
- Zakuraev, Sergey

Group 2 Futurice: ML pipeline to predict the virality of tweets

- Your task will be to build a cloud-based machine learning pipeline that predicts the virality of a tweet (likes, retweets).
- Offline solution, where you take a static dataset of tweets and their related statistics, and predict the virality of a tweet +1 day after getting posted
- Cloud-based ML pipeline, where the model predicts the virality of new incoming tweets. The pipeline should receive new data, preprocess it, get the predictions and collect some performance statistics.
- The pipeline has to be capable of automatically acquiring new data, making predictions and, if needed, re-training the model.
- You can start with an offline dataset of tweets scraped from Twitter API and later augment it with online data updated every day.

Group 3 OP: Forecasting future house prices

- Company representative: Juha Vesanto
- TA Thong Tran tg: @anhthongtran
- Kee, Taeyoung
- Nguyen, Bruce (Trung Quan)
- Nguyen, Son
- Ray, Atreya
- Riippa, Ken
- Tran, Duong

Group 3 OP: Forecasting future house prices

- The development of real estate prices, especially apartments and detached houses, is a highly relevant topic both for private individuals and in terms of the national economy
- You will need to gather related data from public sources and make a forecast model of the development of house prices
- The house price market has clearly different behaviour in different regions and for different types of apartments, so different forecasts will be needed
- The results of the forecasts should be credible and presented clearly and justified by various visualizations and analysis from different viewpoints
- The primary data for the task can be obtained from Tilastokeskus
- Assisting data can be acquired from various apartment marketing web portals

MyCourses

- Find materials of the lectures, group meetings and reading tasks
- Every week’s Thursday morning: report the time spent during the week
- Mondays 20.9.–11.10.: report completion of reading tasks
- Return your final reports, deadlines 26.11. and 9.12.2021
- See: Syllabus, Attendances, Forums, Questionnaires, Resources

Reading task #1

Wes McKinney: *Python for Data Analysis, 2nd Edition*. O'Reilly 2017.

- Available online for Aalto students, link in MyCourses
- Some of the topics will be discussed on lecture 20.9.
- Read the following chapters:
  - 1–3 (pages 1–84) by 20.9.
  - 4–5 (pages 85–165) by 27.9.
  - 6–8 (pages 167–251) by 4.10.
  - 9–11 (pages 253–364) by 11.10.
- Mark in MyCourses assignments when done

Reading task #2

- Visit scikit-learn's website https://scikit-learn.org/stable/
- Go to *User Guide*
  - Expand all items 1.1–8.3
  - Read through the **titles** of all topics in the guide
  - Select one topic of each section 1–8, read and study them in detail
- Go to *Examples*
  - Select 8 examples, read and study them in detail
- Mark in MyCourses assignments which topics and examples you studied
- Deadline Monday 27 September

What will happen on Thursday September 16th?

- We’ll start at 12:15 in Jorma’s Zoom room
- Each group will move to the TAs’ meeting rooms
- TAs will check that you have reported your working hours
- The company representatives will tell about the projects
- You agree with the company representatives about the next actions
- Group 2 Futurice will meet the company representative on Fri 17.9. 10:00–

What to do before Monday September 20th?

- Agree on your group's communication methods
  - **Telegram** / Email / Slack / Yammer / (MyCourses) …
  - Invite Jorma, your assistant and the company representative to the communication group
- Agree on your software repository
  - **GitHub** / Bitbucket / SourceForge / version.aalto.fi
  - Invite Jorma, your assistant and the company representative to the repository
- Agree on your document sharing/writing/wiki platform
  - **Google Drive** / version.aalto.fi / …
  - Invite Jorma, your assistant and the company representative to the repository
- (Later possibly decide about data storage and computing environment.)