COMPDATA.810

Master's Thesis Seminar

CBDA, DS, SDA

Spring 2021, 5 ECTS credits

Jaakko Peltonen

https://coursepages2.tuni.fi/mttts11/

Learning Goals

The student can

- **perform independent research in computational big data analytics** (including statistical and computational aspects of the analysis)
- **report** the underlying theory of the chosen methods and the achieved results in an understandable fashion

The student can

- **comprehend scientific areas of knowledge**
- **follow ongoing developments in the student's field**

The student is able to

- **present** the topic of the pro gradu thesis and the produced analysis results to other students in an understandable manner during the seminar

Stages of the Master's Thesis

Preliminary selection of the topic – data, goal, methods?

Background research / literature study

Narrowing down of the topic based on background research

Acquiring data and becoming initially familiar with it

Writing a practice essay (not obligatory but recommended)

More careful thinking about methods and implementing them

Planning of tests/experiments, writing technical part of thesis

Implementing the tests/experiments

Analyzing and summarizing the results, drawing conclusions

Writing the maturity test

Finalizing the writing of the thesis

Topic

Basic idea of statistical and computational data analytics: explore an interesting set of data, explain its regularities using hypotheses and models; use computational methods to extract the best performing models and hypotheses out of several alternatives.

Data oriented thesis: find an interesting data set and a phenomenon/question to be investigated about it, choose methods etc. for this purpose

Methods oriented thesis: explore e.g. abilities of a recently presented statistical machine learning method, choose data etc. for this purpose

Theory oriented thesis: explore properties of a method/model/data analytics concept or principle by formal proofs and/or experiments

Data Sets?

When the basic nature of the data (e.g. ”temperature time series”) has been determined, one can search for openly available data about that topic online (e.g. a web search for ”temperature time series data set”)

General sources of data sets are e.g.

- UCI Machine Learning Repository http://archive.ics.uci.edu/ml/
- MLData.org http://mldata.org/

Particular types of data have their own repositories,

- for example in bioinformatics one repository is ArrayExpress https://www.ebi.ac.uk/arrayexpress/

In certain kinds of prediction tasks, competitions are regularly held, and the data of those competitions become available

Several governmental and other public organizations publish open data sets, e.g. ”vaalikone” data sets about political candidate opinions

Some companies occasionally publish open data sets, e.g. NetFlix held a prediction competition for movie ratings

Methods?

When the basic characteristics of the task (e.g. ”predict class labels of samples” or ”predict future values of a time series”) have been determined, one can seek methods appropriate for the task from the literature.

The newest methods are not yet in textbooks

Follow important journals and conferences related to the field/task

For published methods, openly available implementations may be available in various programming languages; the older the method the more likely it is to find an existing implementation, but for newest methods it is possible one may have to implement the method oneself.

Reporting

Like writing a scientific article

- Motivate the importance of the topic being studies
- Tell what has been done earlier on the topic
- Tell how the work in the thesis differs from earlier work
- Motivate why the differences (novelties) are important and interesting
- Justify the choices you have made
- Set the questions being investigaged in the thesis work
- Explain how the experiments to be performed will answer the questions
- Explain how the answers can be seen from the experiment results
- Draw final conclusions based on the answers
- Describe what follow-up questions could be studied

Supervision

Supervision form on the webpage
https://coursepages2.tuni.fi/mttts11/

Direct link:

https://content-webapi.tuni.fi/proxy/public/2019-06/opas_ykk_ohjaussuunnitelma_fineng.rtf

Thesis evaluators are often the same as the thesis supervisors, but in some cases there may be other evaluators instead of or in addition to supervisors. A form for setting evaluators of the thesis is available in the student's guide. Direct link:

https://content-webapi.tuni.fi/proxy/public/2019-04/opas_ykk_tarkastajat_fineng.rtf

Grading Principles

There used to be specific grading principles for a CBDA programme master's thesis, but these are currently unavailable. For the moment, you can consider

- the grading principles of computer science master's theses
  https://content-webapi.tuni.fi/proxy/public/2020-11/opas_cs_grading_2020.pdf
- the statement form of computer science master's theses (Finnish language version)
  https://content-webapi.tuni.fi/proxy/restricted/2019-09/intra_tie_pro_gradu_lausuntolomakepohja_10_09_2019.pdf
- the statement form of HTI/SDE master's theses
  https://content-webapi.tuni.fi/proxy/restricted/2019-10/intra_hti_sde_statement_form_masters_thesis_10_09_2019_1.pdf
- the grading principles of mathematics and statistics master's theses (Finnish language version).
  https://content-webapi.tuni.fi/proxy/public/2019-09/mtt_pro_gradu_tutkielman-arviointiperusteet-2019_0.pdf
- the statement form of mathematics and statistics master's theses (Finnish language version)
  https://content-webapi.tuni.fi/proxy/restricted/2020-04/intra_mtt_gradun_lausuntolomake_2020_30042020.pdf

Earlier Master's Theses

http://trepo.tuni.fi/

Note: in TREPO it is not yet possible to search for CBDA, DS, SDA theses specifically. We provide a list of example CBDA, DS, SDA theses on the seminar webpage.

Templates for the thesis

Mathematics and statistics template in LaTeX:
https://moodle.tuni.fi/course/view.php?id=1084 , download available under the "Tilastolliset ohjelmistot" section

Computer science word templates:
Finnish
https://content-webapi.tuni.fi/proxy/public/2020-01/itc_tie_tutkielma_2019.docx
English
https://content-webapi.tuni.fi/proxy/public/2019-10/itc_cs_thesis_09_10_2019.docx

Overleaf template for technology theses:
https://www.overleaf.com/project/5bf56fb665f762287eef13fb

Instructions for Writing the Thesis

The computer science word templates also contain useful instruction about writing the thesis:
Finnish
https://content-webapi.tuni.fi/proxy/public/2020-01/itc_tie_tutkielma_2019.docx
English
https://content-webapi.tuni.fi/proxy/public/2019-10/itc_cs_thesis_09_10_2019.docx

Writing a "practice essay":

- A separate long-form "practice essay" (different from the **maturity test**) is no longer an obligatory part of the thesis process.
- However, it is recommended to write up a part of the thesis contents early on so that supervisors can check the quality of the text and give feedback. It can also help clarify the central motivations and research setup of the thesis.

Maturity test

See https://www.tuni.fi/studentsguide/handbook/uni?page=5358

- In the maturity test, students demonstrate familiarity with the field of the thesis and scientific and professional maturity needed in expert positions in their field of study.
- The maturity test language is determined by the language of the student's school education: Finnish, Swedish, or other.
- Master’s degree students write an abstract for their thesis, which serves as the maturity test. (sometimes called a ”practice research paper”)

Demonstration of Finnish/Swedish proficiency

If the school education was in in Finnish/Swedish in Finland, the student must demonstrate proficiency in Finnish/Swedish through the test (maturity essay), unless they already did a maturity essay in a lower university degree in same language.

Until 31.7.2021: the proficiency is regardless done by the abstract; if the thesis language is not the language (Finnish/Swedish) above, the abstract must be written in that language too.

Good Scientific Practice

Like other scientific work, the thesis work should be carried out following academic ethics and good scientific practice, e.g. avoiding plagiarism or any falsification of data or results. See the student's handbook page

https://www.tuni.fi/studentsguide/handbook/uni?page=2255

Originality of the thesis will be checked using the **Turnitin tool**.

For theses that involve personal data, complying with the needs of **data protection** is part of responsible conduct of research. A resource to learn about data protection has recently been created and is available through Moodle.

Other Material

https://coursepages2.tuni.fi/mttts11/

(e.g. LaTeX instructions, instructions from the university library about how to store the thesis...)

Next Meetings

Meetings roughly every other Tuesday, sometimes more often: planned schedule available at https://coursepages2.tuni.fi/mttts11/

Progress reports by online thesis journal:

- Create a shareable online document, e.g. a OneDrive, Google Doc or Overleaf project, where you journal ongoing development of the thesis
- Share the ability to view it with the seminar lecturer and your supervisor. The document should only detail the work to the level you are willing to share with them.
- The document helps keep track of current issues, ideas & resources, and also acts as a history to remind about unexplored options or reasons for the choices made so far.

In each meeting:

- discussion of some common topic regarding the thesis work in general or regarding the ongoing challenges in the progress reports
- each participant presents their work topic, status and challenges to the others (about 10min/person).