Start-up meeting

MA8701 General Statistical Methods

Mette Langaas, Department of Mathematical Sciences, NTNU

08.01.2019

Last changes: (08.01: added link to slides from Thiago, and contact information for lecturers, removed student names)

Aim

- General information on
  - course content,
  - learning outcome,
  - prerequisites,
  - learning activities
  - evaluation
- A short introduction into the 5 thematic parts of the course, with presentation of lecturers [ML, BAD, TGM]
- The dataset to be used for the compulsory exercise: what, how and why? [TGM]
- TMA4268 Statistical learning - topics/lectures that you might want to attend?
- Presentation of students and setting up a reference group (in small courses all students may be the reference group)
- Questions/suggestions/ideas?

Course content

taken from
https://www.ntnu.edu/studies/courses/MA8701#tab=omEmnet

The course is usually given every second year, and only if a sufficient number of students register. It is given next time Spring 2019. If too few students register, the course is given as a guided self study.

~~The course provides a broad introduction to the basic principles and methods of statistical inference and prediction.~~

The course will give a detailed presentation of selected (contemporary?) advanced topics in statistical inference and learning.

Together with course MA8704 Probability theory and asymptotic techniques it provides a theoretical basis for PhD students in statistics.

Old content - change due to the new course TMA4268 Statistical learning.

The course includes: Introduction to supervised learning. ~~Linear methods for regression and classification.~~ Basic expansions and regularization. Kernel smoothing methods. ~~Likelihood inference and asymptotic methods.~~ ~~Model inference, assessment and selection.~~ ~~Empirical Bayes methods.~~

New content

Part 1: Regularized linear and generalized linear models [2w]
Part 2: Smoothing and splines [2w]
Part 3: Experimental design in statistical learning [1w]
Part 4: Deep neural nets [3w]
Part 5: Active learning [2w]

Learning outcome

1. Knowledge. (tentative new outline)

The course gives the student a through background in selected topic relevant for statistical inference and learning. Together with course MA8704 Probability theory and asymptotic techniques it provides a theoretical basis for PhD students in statistics, and together with MA8702 Anvanced computer intensive statistical methods it provides a computational basis.

2. Skills (tentative new outline)

After completing this course the students are able to use advanced techniques in statistical inference and learning for analysing complex and large amounts of data.

3. Competence

The students will be able to participate in scientific discussions and carry out research in statistics at high international level. They will be able to participate in applied projects involving statistical methods and to apply their knowledge to problems in theoretical statistics.

Important background knowledge

for the students.

"Compulsory" courses

TMA4267 Linear statistical models
TMA4295 Statistical inference
TMA4300 Computer intensive statistical methods
TMA4315 Generalized linear models

Helpful knowledge

TMA4268 Statistical learning
TMA4180 Optimization

Programming/IT-knowledge

In addition good programming skills in either R or Python, and it is also preferable if you have some knowledge of *commands in unix and the skills to be able to run a script on a computer cluster*.

Suggested help for unix shell:
http://swcarpentry.github.com/shell-novice/

Course learning activities

- Teaching (and supervision) activities are planned in calendar weeks 2-13, and presentation of project work in weeks 14+15.
- We start on Tuesday January 8, and continue until Tuesday April 9, 2018.
- All weeks we will have lectures Tuesdays at 12.15-15.00, but week 8 (experimental design) we have lectures Tuesday 12.15-14.15 and Thursday 10.15-12.00
- All lectures will be in 734, 7. etg, sentralbygg 2, unless we find that this room is not meeting our needs.

Evaluation

A larger project will count 30% of the grade, and is to be presented orally by the students (R or Python). The project work can be done in teams of 2 or 3 students.

There will be a final oral exam counting 70% of the grade (date not decided)

The grade for this course is pass/fail, and 70/100 score is required to pass (this is the standard rule for PhD courses at NTNU).

The course parts

Introduction to the course

Lecture: week 2 08.01 - this lecture!

Part 1: Statistical learning with sparsity (Benjamin Dunn)

We expand on what you have learned about the lasso and ridge in TMA4268, and marry with the GLM from TMA4315.

Lectures: week 3 15.01 and week 4 22.01.

Readings:

1. Hastie et al (2015): Statistical learning with sparsity (selected chapters) Book at https://web.stanford.edu/~hastie/StatLearnSparsity_files/SLS_corrected_1.4.16.pdf (free), see chapters below. NA

Benjamin elaborates!

Day 1 (15.01.2019)

- Intro to lasso - Chapters 2.1-2.6, 2.1, 5.4
  - Linear regression
  - Why sparsity?
  - Least absolute shrinkage and selection operator (lasso) and related approaches
  - Fitting the model
- Exercise (bring your computer)
  - Write script (in R or python) to implement the cyclic coordinate descent method for lasso
- GLMs with regularization Chapters 2.9, 3.1-3.2, 3.7, 5.4
  - Generalized linear models (GLM)
  - Logistic regession with l1 (example)
  - Fitting the model

Day 2 (22.01.2019)

- Generalizations of lasso – Chapter 4.1-4.3, 4.5-4.6
  - Elastic net
  - Relaxed lasso
  - Grouped lasso
  - Fused lasso
  - Non-convex penalties
- Exercise (bring your computer)
  - Analysis of breast cancer data using logistic regression with l1 and l2
- Inference for lasso – Chapter 6.1, 6.3 and Dezeure et al. (2015)
  - Bootstrap method
  - Multi sample-splitting

Performing computations and project work: (Erlend Aune)

Lecture: week 5 29.01

Erlend Aune will talk about how to perform computations (hopefully on a GPU cluster) for the compulsory project.

In addition: those of you who are not familiar with **Classification** can attend the lectures in TMA4268 Statistical learning Monday 28.01 at 08.15 in S4 and Thursday 31.01 at 14.15-16.00 in F6 with Mette Langaas.

Part 2: Smoothing and splines (Bo Lindqvist)

Lectures: week 6 05.02 and week 7 12.02

Readings: Friedman, Hastie and Tibshirani (2008): Elements of Statistical Learning. Chapter 5 and 6. Book at https://web.stanford.edu/~hastie/ElemStatLearn/ (free)

Slides from Bo: https://www.math.ntnu.no/emner/MA8701/2019v/Smoothingandsplines.pdf

Part 3: Experimental design in statistical learning (John Tyssedal)

Lectures: week 8 19.03 12.15-14.00 and 21.03 10.15-12.00

Readings: TBA

John elaborates:

Machine learning methods are often complex and they may resist formal analysis methods. Therefore to learn about and better understand their behavior we need to do empirical investigations which are best performed using controlled experiments. In these lectures we will mainly concentrate on three topics:

1. How to optimize the choice of hyperparameters
2. How to compare the performance of two or more algorithms on a single data set
3. How to compare the performance of two or more algorithms on several datasets

all based on methods within Design of Experiments.

Motivation: Expert Systems with Applications, Volume 109, 1 November 2018, Pages 195-205: "Design of experiments and response surface methodology to tune machine learning hyperparameters, with a random forest case-study"
Authors: Gustavo A. Lujan-Morenoa, Phillip R. Howard, Omar G. Rojasa, Douglas C. Montgomery

In this paper, we propose a design of experiments (DOE) methodology as the first step to screen the most significant hyperparameters (factors) of a ML algorithm.

Reducing the number of factors to a subset which has the greatest effect on model performance considerably reduces the number of model-fitting runs in the next round of hyperparameter tuning experiments.

The screening phase is done using fractional factorial designs, which are well-suited for scenarios in which we do not have the luxury of running many experiments; screening may also be done using other designs, as explained at the end of Section 2.1.

Once the main factors are identified, a full factorial experiment can be run on the factors as a confirmatory procedure.

The second phase of our method consists of applying response surface methodology (RSM) to model a first- or second-order polynomial which approximates the performance of the model given different hyperparameter configurations.

Part 4: Deep neural nets (Thiago Martins)

Lectures: week 9 26.02, week 10 05.03 and week 11 12.03

Readings:

NA or NA

You choose what you read, both built on the keras package. This book you have to buy - at Manning as ebook or Akademika on paper. The library should also have 2 copies of the R book.

Thiago's slides

Part 5: Active learning (Erlend Aune)

Lectures: week 12 19.03 and week 13 26.03

Readings: Review article to be announced, but this should be a good read http://burrsettles.com/pub/settles.activelearning.pdf

Erlend refers us to the Wiki page for Active learning: https://en.wikipedia.org/wiki/Active_learning_(machine_learning) and says that he will probably also cover the cutting edge *transfer learning*.

The dataset to be used for the compulsory exercise: what, how and why?

Thiago elaborates!

Link to data set:
https://www.kaggle.com/c/avito-demand-prediction/overview

TMA4268 Statistical learning

Some of you have not taken our first course on statistical learning, TMA4268. (Who are you?)

Lectures are Mondays at 08.15-10.00 in S4 and Thursdays 14.15-16.00 in F6/Smia.

Textbook is James et al: Introduction to Statistical Learning, with Applications in R, which is made as a soft version of the Elements book that we use in Part 2.

Here is an overview of the topics we cover and when: https://www.math.ntnu.no/emner/TMA4268/2019v/TMA4268overview.html

Contact information for course staff

- Benjamin Dunn: benjamin.dunn@ntnu.no
- Bo Lindqvist: bo.lindqvist@ntnu.no
- John Tyssedal: john.tyssedal@ntnu.no
- Thiago Martins: tgm@aiascience.com
- Erlend Aune: erlend.aune.1983@gmail.com
- Mette Langaas: Mette.Langaas@ntnu.no