# MA8701 Advanced methods in statistical inference and learning

## L1: Introduction

Mette Langaas IMF/NTNU

28 December, 2020

# Learning outcome

1. **Knowledge**
   - Understand and explain the central theoretical aspects in statistical inference and learning.
   - Understand and explain how to use methods from statistical inference and learning to perform a sound data analysis.
   - Be able to evaluate strengths and weaknesses for the methods and choose between different methods in a given data analysis situation.

## 2. Skills

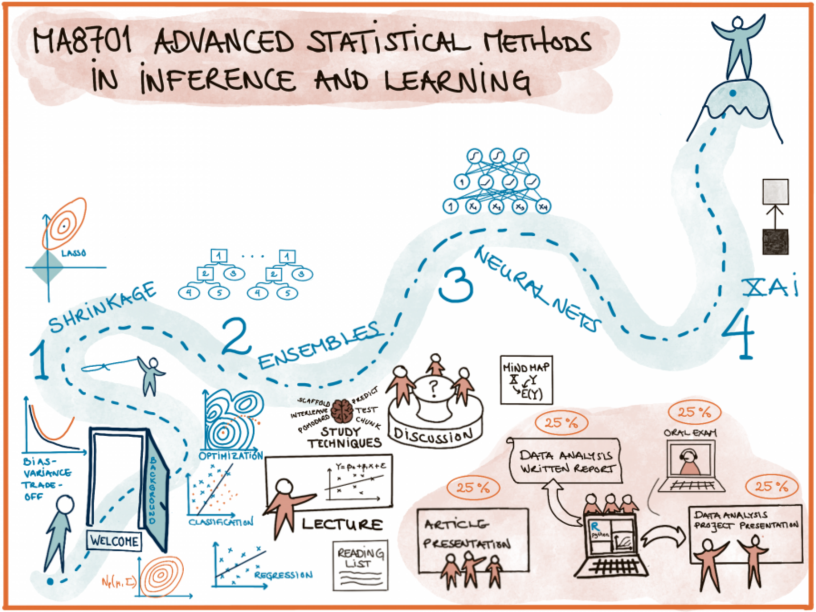
Be able to analyse a dataset using methods from statistical inference and learning in practice (using R or Python), and give a good presentation and discussion of the choices done and the results found.

## 3. Competence

- The students will be able to participate in scientific discussions, read research presented in statistical journals, and carry out research in statistics at high international level.
- They will be able to participate in applied projects, and analyse data using methods from statistical inference and learning.

# Useful/required previous knowledge

- TMA4267 Linear Statistical Methods,
- TMA4268 Statistical learning,
- TMA4295 Statistical inference,
- TMA4300 Computer intensive statistical methods,
- TMA4315 Generalized linear models
- Good understanding and experience with R, or with Python, for statistical data analysis.
- Knowledge of RMarkdown for writing reports and presentations


MA8701 ADVANCED STATISTICAL METHODS
IN INFERENCE AND LEARNING
LASSO
SHRINKAGE
1
2
ENSEMBLES
3
NEURAL NETS
4
XAi
BIAS-VARIANCE TRADE-OFF
WELCOME
BACKGROUND
$N_p(\mu,\Sigma)$
OPTIMIZATION
CLASSIFICATION
REGRESSION
SCAFFOLD
INTERLEAVE
POMODORO
PREDICT
TEST
CHUNK
STUDY TECHNIQUES
$Y=\beta_0+\beta_1 x+\epsilon$
LECTURE
READING LIST
?
DISCUSSION
MIND MAP
X Y
E(Y)
25 %
DATA ANALYSIS WRITTEN REPORT
25 %
ORAL EXAM
25 %
ARTICLE PRESENTATION
R
Python
25 %
DATA ANALYSIS PROJECT PRESENTATION

# Course topics

The starting point is that we cover important parts of

The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics, 2009) by Trevor Hastie, Robert Tibshirani, and Jerome Friedman.

but, since the book is from 2008 this means that for many topic we need (to be up to date) additional selected material in the form of book chapters and research articles.

## Introduction [this part, one week]

Sort out assumed background knowledge, and learn something new

- Regression (what do we know)
- Classification (ditto)
- Bias-variance trade-off (what and why important)
- Model selection and model assessment (new)

## Part 1: Shrinkage [3 weeks]

or “Regularized linear and generalized linear models”.

- ELS 3.2.3,3.4, 3.8, 4.4.4.
- Hastie, Tibshirani, Wainwright (HTW): “Statistical Learning with Sparsity: The Lasso and Generalizations”. Selected chapters.
- Post-selective inference (articles)
- An introduction to analysing text

Includes one data analysis project with short report.

## Part 2: Ensembles [4 (5) weeks]

- trees, bagging and forests
- general ensembles (similar to super learner)
- boosting
- hyper-parameter tuning

Selected chapters in ELS (8.7, 8.8, 9.2, parts of 10, 15, 16) and several articles.

## Part 3: Neural nets [(2) 3 weeks]

- Goodfellow, Bengio, Courville: Deep learning (2016). MIT press. https://www.deeplearningbook.org/. Selected chapters.
- Evaluating uncertainty

## Part 4: XAI [2 weeks]

Lectured by Kjersti Aas.

Articles on

- LIME,
- partial dependence plots,
- Shapley values,
- relative weights and
- counterfactuals.

## Closing [1 week]

- w/oral presentations of second data project from Parts 2-4.

# Learning methods and activities

- Lectures will be on 14 Mondays 9.15-12 in S21 (and zoom). We will not record the teaching because we will try to include student activities in groups (on tables or break-out rooms).
- Exploring different study techniques (one or more each lecture).
- Problem sets to work on between lectures.
- Final individual oral exam (25% of pass/fail grade) in May.

- One practical compulsory group project in data analysis (application of course theory using R or Python) with short report. Topic: Part 1 on Shrinkage, chosen data set discussed with lecturer before start. Due mid February. First given comments by one other group, then evaluated by course responsible. (25% of pass/fail grade)

- One article group presentation, orally (15 minutes+questions). Material from Parts 2 and 3 preferred, and must be decided on with lecturer (might also be parts of your own master thesis if applicable). Due before Easter. (25% of pass/fail)

- Practical compulsory project in data analysis (application of course theory using R or Python) with oral presentation (15 minutes+questions). Topic: Part 2-4, data set and methods discussed with lecturer before start. Due after Part 4 is finished. (25% of pass/fail grade)

- For the two data analysis projects: one should be with a data set requiring regression and one with classification type analysis.

# Course wiki

https://wiki.math.ntnu.no/ma8701/2021v/start

**Questions?**

# Class activity: Cat or dog?

Aim: get to know each other - to improve on subsequent group work!

```
while (at least one student not presented)
   lecturer give two alternatives, you choose one.
   lecturer choose a few students to present their view
   together with giving their name and study programme
   (and say if they are looking for group members)
```

- Dog person or cat person?
- When performing logistic regression - do you then say you do statistical learning or machine learning?
- I will show you the result of a descriptive analysis: summary or graphical display?
- Learning something new: read a book or watch a video?
- Analysing data: R or python?
- Analysing data: report p-values and or confidence intervals
- In class: taking notes or not?
- Use camel case or snake case for programming?

camel: writing compound words such that each word in the middle of the phrase begins with a capital letter, with no intervening spaces or punctuation. "camelCase" or "CamelCase".

snake: writing compound words where the elements are separated with one underscore character (_) and no spaces, with each element's initial letter usually lower cased within the compound and the first letter either upper- or lower case as in "foo_bar"

# Part 0: Introduction

finally - we start on the fun stuff!

## Plan

Sort out assumed background knowledge,

- Regression (ELS ch 3, except 3.2.3, 3.2.4, 3.4, 3.7, 3.8)
- Classification (ELS ch 4.1-4.5, except 4.4.4)

and the cover new aspects for

- Model selection and assessment (ELS Ch 7.1-7.6, 7.10-7.12), including statistical learning and the bias-variance trade-off (ELS ch 2)

## Regression

1) What do we know about multiple linear regression? We discuss main aspects.
2) How will we build on MLR and use regression in Parts 1-4?

## Resources

(mostly what we learned in TMA4267, or ELS ch 3, except 3.2.3, 3.2.4, 3.4, 3.7, 3.8)

- From TMA4268: https://www.math.ntnu.no/emner/TMA4268/2019v/TMA4268overview.html and in particular https://www.math.ntnu.no/emner/TMA4268/2019v/3LinReg/3LinReg.html
- From TMA4315: https://www.math.ntnu.no/emner/TMA4315/2018h/TMA4315overviewH2018.html and in particular https://www.math.ntnu.no/emner/TMA4315/2018h/2MLR.html

## Classification

What do we know about classification? (TMA4268 and TMA4315 mainly, or ELS ch 4.1-4.5, except 4.4.4)

- Sampling vs diagnostic paradigm
- Parametric vs non-parametric methods
- LDA
- Logistic and multinomial regression

## Resources

(mostly what we learned in TMA4267, or ELS ch 4.1-4.5, except 4.4.4)

- From TMA4268: https://www.math.ntnu.no/emner/TMA4268/2019v/TMA4268overview.html and in particular https://www.math.ntnu.no/emner/TMA4268/2019v/4Classif/4Classif.html
- From TMA4315: https://www.math.ntnu.no/emner/TMA4315/2018h/TMA4315overviewH2018.html and in particular https://www.math.ntnu.no/emner/TMA4315/2018h/3BinReg.html and https://www.math.ntnu.no/emner/TMA4315/2018h/6Categorical.html.

## Model selection and assessment (including Bias-variance trade-off)

1) What do we know about the bias-variance trade-off? We discuss main aspects.
2) Main part: We look into ELS ch 7.
3) How will we build on this in Parts 1-4?

## Resources

(what we learned in TMA4268, and ELS ch 2 and 7.1-7.6+7.10-7.12)

From TMA4268: https://www.math.ntnu.no/emner/TMA4268/2019v/TMA4268overview.html and in particular https://www.math.ntnu.no/emner/TMA4268/2019v/2StatLearn/2StatLearn.html

# Exercises

# References

- R Markdown Cookbook: https://bookdown.org/yihui/rmarkdown-cookbook/
- R Markdown cheat sheet: >https://rstudio.com/wp-content/uploads/2015/03/rmarkdown-reference.pdf>
- https://en.wikipedia.org/wiki/Camel_case
- https://en.wikipedia.org/wiki/Snake_case
- https://machinelearningmastery.com/mcnemars-test-for-machine-learning/

When relevant? https://www.math.ntnu.no/emner/TMA4315/2017h/qq.html