

Module 1: Introduction

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

March xx, 2020

Acknowledgements

This course had been built up by Mette Langaas at NTNU in 2018 and 2019. I am using a lot of her material, and material from her TAs, throughout the course.

I would like to thank Mette for her great work and for the permission to use her material!

Introduction

Aims of the first module

- What is statistical learning?
- Types of problems we will look at
- Course overview and learning outcome, activities and team
- Course modules, practical details (Blackboard)
- Getting to know you – who are you? Background?
- Key concepts from your first course in statistics – that you will need now, mixed with notation for this course
- Introduction to R and RStudio

Learning material for this module

- Our textbook James et al (2013): An Introduction to Statistical Learning - with Applications in R (ISL). Chapter 1 and 2.3.
- [Rbeginner](#) and [Rintermediate](#)
- Link to background on Matrix Algebra: [Härdle and Simes \(2015\)](#) - [A short excursion into Matrix Algebra](#) (on the reading list for TMA4267 Linear statistical models).

What is statistical learning?

- Refers to *a vast set of tools to understanding data* (text book, p. 1).
- Focuses on the whole chain:

model \rightarrow method \rightarrow algorithm \rightarrow analysis \rightarrow interpretation

- Both **prediction** and **understanding** (inference \rightarrow drawing conclusions).
- Statistical learning is a statistical discipline.

Statistical Learning vs. “Machine Learning”

Machine learning is more focused on the algorithmic part of learning, and is a discipline in computer science.

But, many methods/algorithms are common to the fields of statistical learning and machine learning.

Statistical Learning vs. “Data Science”

In data science the aim is to

- extract knowledge and understanding from data, and
- requires a combination of statistics, mathematics, numerics, computer science and informatics.

This encompasses the whole process of data acquisition/scraping, going from unstructured to structured data, setting up a data model, performing data analysis, implementing tools and interpreting results.

In statistical learning we will not work on the two first above (scraping and unstructured to structured).

[R for Data Science](#) is an excellent read and relevant for this course!

Problems you will learn to solve

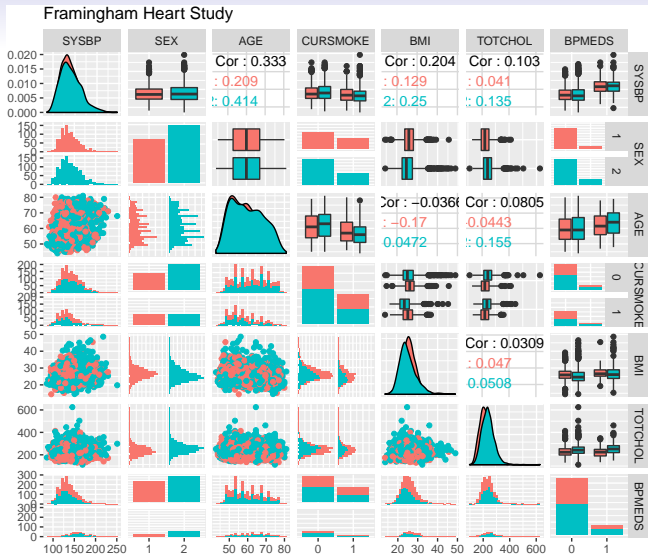
There are **three main types of problems** discussed in this course:

- Regression
- Classification
- Unsupervised methods

using data from science, technology, industry, economy/finance, ...

Example 1: Regression (Etiology of CVD)

- The Framingham Heart Study investigates the underlying causes of cardiovascular disease (CVD) (see <https://www.framinghamheartstudy.org/>).
- Aim: modelling systolic blood pressure (SYSBP) using data from $n = 2600$ persons.
- For each person in the data set we have measurements of the following seven variables.
 - SYSBP systolic blood pressure (mmHg),
 - SEX 1=male, 2=female,
 - AGE age (years),
 - CURSMOKE current cigarette smoking at examination: 0=not current smoker, 1=current smoker,
 - BMI body mass index,
 - TOTCHOL serum total cholesterol (mg/dl),
 - BPMEDS use of anti-hypertensive medication at examination: 0=not currently using, 1=currently using.



What does this plot show?

Red: male; turquoise: female

- Diagonal: density plot (generalization of histogram), or barplot.
- Lower diagonals: scatterplot, histograms
- Upper diagonals: correlations values or boxplots

Etiology of CVD - model

- A *multiple normal linear regression model* was fitted to the data set with

$$-\frac{1}{\sqrt{\text{SYSBP}}}$$

as response (output) and all the other variables as covariates (inputs).

- The results are used to formulate hypotheses about the etiology of CVD - to be studied in new trials.

```

modelB = lm(-1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL + BPMEDS,
  data = thisds)
summary(modelB)$coeff
summary(modelB)$r.squared
summary(modelB)$adj.r.squared

```

```

##              Estimate   Std. Error    t value    Pr(>|t|)
## (Intercept) -1.105693e-01 1.341653e-03 -82.4127584 0.000000e+00
## SEX2        -2.989392e-04 2.390278e-04  -1.2506465 2.111763e-01
## AGE         2.378224e-04 1.433876e-05  16.5859862 8.461545e-59
## CURSMOKE1   -2.504484e-04 2.526939e-04  -0.9911136 3.217226e-01
## BMI         3.087163e-04 2.954941e-05  10.4474583 4.696093e-25
## TOTCHOL     9.288023e-06 2.602433e-06   3.5689773 3.648807e-04
## BPMEDS1     5.469077e-03 3.265474e-04  16.7481874 7.297814e-60
## [1] 0.2493538
## [1] 0.2476169

```

```
summary(modelB)
```

```
##
## Call:
## lm(formula = ~1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL +
##     BPMEDS, data = thisds)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.0207366 -0.0039157 -0.0000304  0.0038293  0.0189747
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.106e-01  1.342e-03 -82.413 < 2e-16 ***
## SEX2        -2.989e-04  2.390e-04  -1.251 0.211176
## AGE         2.378e-04  1.434e-05  16.586 < 2e-16 ***
## CURSMOKE1   -2.504e-04  2.527e-04  -0.991 0.321723
## BMI         3.087e-04  2.955e-05  10.447 < 2e-16 ***
## TOTCHOL     9.288e-06  2.602e-06   3.569 0.000365 ***
## BPMEDS1     5.469e-03  3.265e-04  16.748 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.005819 on 2593 degrees of freedom
## Multiple R-squared:  0.2494, Adjusted R-squared:  0.2476
## F-statistic: 143.6 on 6 and 2593 DF,  p-value: < 2.2e-16
```

Example 2: Classification (iris plants)

The **iris** flower data set is a very famous multivariate data set introduced by the British statistician and biologist Ronald Fisher in 1936.

The data set contains **three plant species** {setosa, virginica, versicolor} and **four features measured** for each corresponding sample: **Sepal.Length**, **Sepal.Width**, **Petal.Length** and **Petal.Width**.

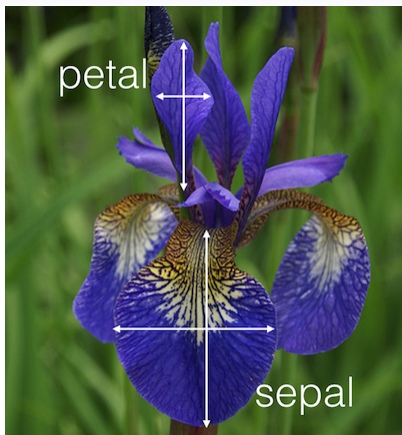


Figure 1: Iris plant with sepal and petal leaves

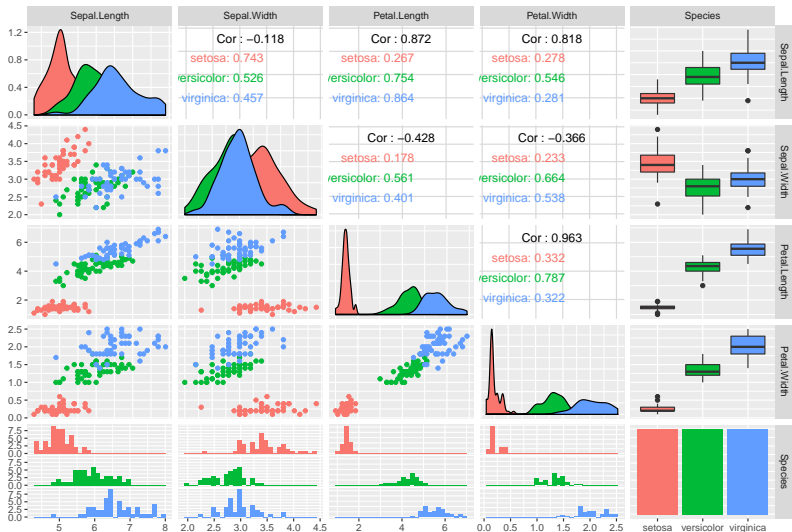
<http://blog.kaggle.com/2015/04/22/scikit-learn-video-3-machine-learning-first-steps-with-the-iris-dataset/>


```
head(iris)
```

```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1         5.1         3.5         1.4         0.2   setosa
## 2         4.9         3.0         1.4         0.2   setosa
## 3         4.7         3.2         1.3         0.2   setosa
## 4         4.6         3.1         1.5         0.2   setosa
## 5         5.0         3.6         1.4         0.2   setosa
## 6         5.4         3.9         1.7         0.4   setosa
```

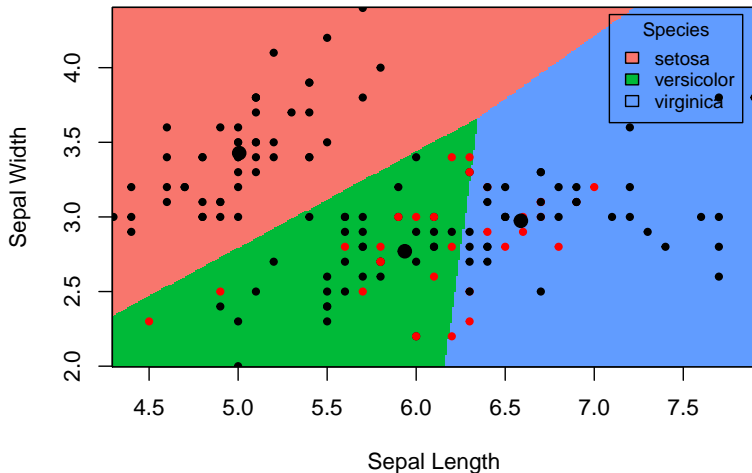
Aim: correctly classify the species of an iris plant from sepal length and sepal width.

Classification of Iris plants



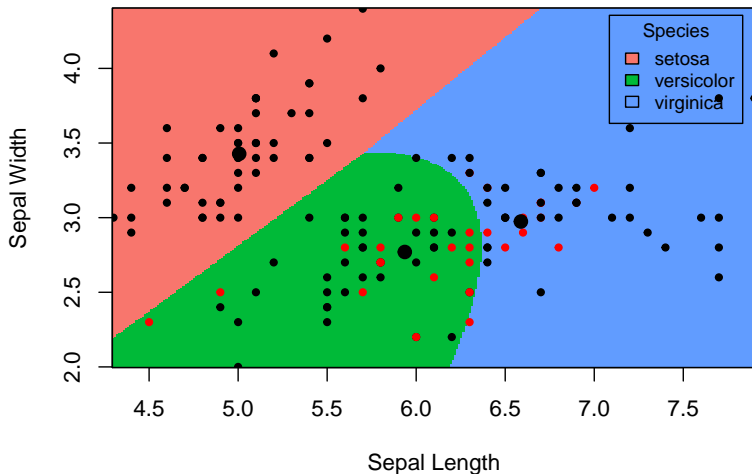
Linear boundaries

In this plot the small black dots represent correctly classified iris plants, while the red dots represent misclassifications. The big black dots represent the class means.



Non-linear boundaries

Sometimes a more suitable boundary is not linear.



Example 3: Unsupervised methods (Gene expression)

- In a collaboration with researchers the Faculty of Medicine and Health the relationship between inborn maximal oxygen uptake and skeletal muscle gene expression was studied.
- Rats were artificially selected for high- and low running capacity (HCR and LCR, respectively),
- Rats were either kept sedentary or trained.
- Transcripts significantly related to running capacity and training were identified
- To further present the findings heat map of the most significant transcripts were presented (high expression are shown in red and transcripts with a low expression are shown in yellow).
- This is hierarchical cluster analysis with pearson correlation distance measure.

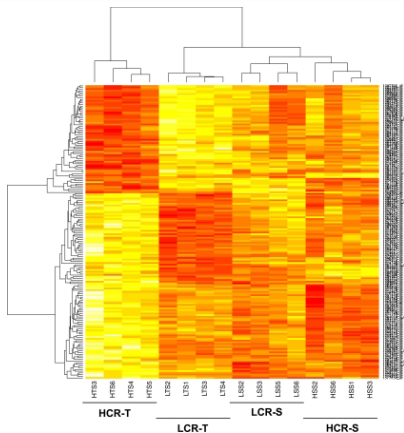


Figure 2: Heatmap with dendrograms

More: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2585023/>

Who is this course for?

Primary requirements

- Bachelor level: 3rd year student from Science or Technology programs, and master/PhD level students with interest in performing statistical analyses.
- Statistics background: TMA4240/45 Statistics, ST1101+ST1201, or equivalent.
- No background in statistical software needed: but we will use the R statistical software extensively in the course. Knowing python will make this easier for you!
- Not a prerequisite but a good thing with knowledge of computing - preferably an introductory course in informatics, like TDT4105 or TDT4110.

Overlap

- TDT4173 Machine learning and case based reasoning: courses differ in philosophy (computer science vs. statistics). See Bb under FAQ for more details.
- TMA4267 Linear Statistical Models: useful to know about multivariate random vectors, covariance matrices and the multivariate normal distribution. Overlap only for Multiple linear regression (M3).

About the course

Focus: Statistical theory **and** doing analyses

- The course has focus on **statistical theory**, but we apply all models and theory using (mostly) available function in R and real data sets.
- It is important that the student in the end of the course **can analyse all types of data** (covered in the course) - not just understand the theory.
- And vice versa - the student must also **understand** the model, methods and algorithms used.

Learning outcome

1. **Knowledge.** The student has knowledge about the most popular statistical learning models and methods that are used for *prediction* and *inference* in science and technology. Emphasis is on regression- and classification-type statistical models.
2. **Skills.** The student can, based on an existing data set, choose a suitable statistical model, apply sound statistical methods, and perform the analyses using statistical software. The student can present, interpret and communicate the results from the statistical analyses, and knows which conclusions can be drawn from the analyses, and what are the caveats.

Course content

- Statistical learning
- Multiple linear regression
- Classification
- Resampling methods
- Model selection/regularization
- Non-linearity
- Support vector machines
- Tree-based methods
- Unsupervised methods
- Neural nets

Learning methods, activities and grading

- Lectures, exercises and works (projects).
- Portfolio assessment is the basis for the grade awarded in the course. This portfolio comprises
 - a written final examination (80%).
 - works (projects) (20%).
- The results for the constituent parts are to be given in %-points, while the grade for the whole portfolio (course grade) is given by the letter grading system. Retake of examination may be given as an oral examination. The lectures may be given in English.

Literature for the course

Textbook: James, Witten, Hastie, Tibshirani (2013): “An Introduction to Statistical Learning”.

Tentative reading list: the whole book + the module pages+ two compulsory exercises.

Course team

Lecturers:

- Stefanie Muff (IMF and CBD at NTNU)
- Thiago G. Martins (AIAscience and IMF/NTNU)
- Andreas Strand (NTNU)

Teaching assistants (TA)

- Martina Hall
- Michail Spitieris

Teaching philosophy

The modules

- Divide the topics of the course into modular units with specific focus.
- This (hopefully) facilitates learning?
- 12 modules (weeks of teaching) =
 - **Module 1:** introduction (this module)
 - **Modules 2 - 11:** the 10 topics listed previously
 - **Module 12:** summing up
- Two weeks without lectures (but supervision of exercises)

The lectures

Mondays at 8.15-10 in S4 and Thursdays at 14.15-16.00 in F6 or Smia

- We have 2 · 2 hours of lectures every week (except when working with the compulsory exercises).
- Each week is a new topic and a new module
- Some weeks (modules 2, 5, 7 and 9) the second 2h are partly a lecture and partly an *interactive lecture* in Smia, where active learning is in focus (more below).
- The other weeks (modules 3, 4, 6, 8, 10 and 11) the second lecture is a plenary lecture in F6.
- The first week of the course the second lecture is replaced by an R workshop!

The weekly supervision sessions

Thursdays 16.15-18 in Smia

For each module *recommended exercises* are given. These are partly

- theoretical exercises (from book or not)
- computational tasks
- data analysis

These are supervised in the weekly exercise slots.

We have joint supervision sessions with the TMA4267 Linear statistical models course.

TAs from both courses will be present.

The compulsory exercises

- We will have **two compulsory exercises**, each with a maximal score of 15 points.
- These are supervised in the weekly exercise slots and there will be one week without lectures (only with supervision) for each compulsory exercise.
- Focus: theory, analysis in R, and interpretation.
- Work in **groups of maximum 3**; handed in on Blackboard and be written in R Markdown (both .Rmd and .pdf handed in).
- The TAs grade the exercises.
- This gives 20% of the final evaluation in the course, and the written exam the final 70%.

How to use the module pages?

- A slides version (output: `beamer__presentation`) of the pages used in the plenary lectures.
- A webpage version (output: `html__document`) used in the interactive lectures.
- A document version (output: `pdf__document`) used for student self study.
- The Rmd version - used as notebook to investigate changes to the R code.
- Additional class notes (written in class) linked in.

The module pages are the backbone of the course!

Student active learning

We (probably) all have different ways in which we learn - and we have different learning ambitions when attending a course.

Felder and Silverman learning styles

Back in 1988 Felder and Silverman published an article where they suggested that there was a mismatch between the way students learn and the way university courses were taught. They devised a taxonomy for learning styles - where four different axis are defined:

1. active - reflective: How do you process information: actively (through physical activities and discussions), or reflexively (through introspection)?
2. sensing-intuitive: What kind of information do you tend to receive: sensitive (external agents like places, sounds, physical sensation) or intuitive (internal agents like possibilities, ideas, through hunches)?
3. visual-verbal: Through which sensorial channels do you tend to receive information more effectively: visual (images, diagrams, graphics), or verbal (spoken words, sound)?
4. sequential - global: How do you make progress: sequentially (with continuous steps), or globally (through leaps and an integral approach)?

The idea:

- Acknowledging these different learning style axes
- Choose teaching styles that match the learning styles of the students.
- Many students (according to Felder and coauthors) have a visual way of learning, and then teachers should spend time devising visual aids (in addition to verbal aids - that were the prominent aids in 1988), and so on.

Student active learning in TMA4268

- Students have different background knowledge, learning ambitions and learning styles (verbal/visual, active/reflective, global/sequential, ...).
- Teachers aim to make diverse learning resources available in courses, and
- *promote student active teaching methods.*

Hypothesis: Students will learn better/deeper with a combination of different teaching methods.

Therefore we aim to:

- Provide learning environments, opportunities, interactions, tasks and instruction that foster deep learning.
- Provide guidance and support that challenges students based on their current ability.
- Students discover their current strengths and weaknesses and what they need to do to improve.

We will now focus on *active* and *reflective* learning styles and learning methods.

Active vs. reflective learning styles

What are student reflective learning methods/tasks?

- Plenary lectures
- Reading textbook
- Self study

Plenary lecture

1. The teacher is lecturer
2. Limited amount of interaction student-lecturer
3. Limited amount of interaction student-student
4. Lecturer active, student passive
5. Where: lecture theatre, auditorium
6. Resources: (black)board, screen (for projector), microphone for lecturer, desks in rows.

Positive: By using the traditional lecture method, lecturers can present a large amount of material in a relatively brief amount of time.

Negative: is it also efficient for the student? Maybe efficient for students with a *reflective learning style*?

What are student active learning methods/tasks?

- Pause in plenary lecture to ask questions and let students think and/or discuss.
- In-class quizzes (with the NTNU invention Kahoot!) - individual and team mode.
- Projects - individual or in groups.
- Group discussion
- Interactive lectures

Interactive lecture

1. The teacher is not a lecturer, but an advisor, and teaching assistant (TA) may also be present
2. Focus on interaction student-teacher
3. Focus on interaction student-student
4. Student and teacher/TA both active
5. Where: a room designed for interaction, collaboration and activity
6. Resources: (white)boards, screens, group tables, laptop/PC, possibly a podium for short announcements

Positive: Students are active. This is how work is done in many cooperative environments (your future workplace?).

Negative: Challenging group processes. Unfamiliar setting for students?

Challenge: requires dedicated space and suitable problems to work on in groups

Innovative learning space: Smia

A room where interaction and activity is in focus. Flat floor with group tables, whiteboard and screen - PC and electricity outlets.

<https://www.ntnu.no/laeringsarealer/smia>

Interactive lectures in Smia

(modules 2,5,7 and 9): Work alone, in pairs, or larger groups (lecturer help to form), introduction (5 min) - work on problem sets (35+, lecturer and TA supervise) - summing up (5 min), pause, repeat. Lecturer interrupts if common confusions or interesting observations by some of the tables.

The course modules

(PL=plenary lecture in S4 (Mondays) or F6 (Thursdays),
IL=interactive lecture in Smia, E=exercise in Smia)

A detailed overview is found on Bb, and outside Bb in this [table](#).

1. Introduction

[Ch 1, ML] 2019-w2 (PL: 07.01, R-workshop (IL/E): 10.01)

- About the course
- Key concepts in statistics.
- Intro to R and RStudio

2. Statistical Learning

[Ch 2, ML] 2019-w3 (PL: 14.01, IL: 17.01 and E: 17.01)

- Estimating f (regression, classification), prediction accuracy vs model interpretability.
- Supervised vs. unsupervised learning
- Bias-variance trade-off
- The Bayes classifier and the KNN - a flexible method for regression and classification

First hours of IL: bias-variance trade off. Second hour: we target students that doesn't plan to take TMA4267 Linear statistical models, and we work with random vectors, covariance matrices and the multivariate normal distribution (very useful before Modules 3 and 4).

3. Linear Regression

[Ch 3, ML] 2019-w4 (PL: 21.01, PL: 24.01 and E: 24.01)

- Simple and multiple linear regression: model assumptions and data sets
- Inference: Parameter estimation, CI, hypotheses, model fit
- Coding of qualitative predictors
- Problems and extensions
- Linear regression vs. KNN.

4. Classification

[Ch 4, ML] 2019-w5 (PL: 28.01, PL: 31.01 and E: 31.01)

- When to use classification (and not regression)?
- Logistic regression
- Linear discriminant analysis LDA and quadratic discriminant analysis QDA
- Comparison of classifiers

5. Resampling methods

[Ch 5, ML] 2019-w7 (PL: 04.02, IL: 07.02, E: 07.02)

Two of the most commonly used resampling methods are cross-validation and the bootstrap. Cross-validation is often used to choose appropriate values for tuning parameters. Bootstrap is often used to provide a measure of accuracy of a parameter estimate.

- Training, validation and test sets
- Cross-validation
- The bootstrap

Part 1: Modules 2-5

is finished with compulsory exercise 1. In week 7 we have supervision in Smia 11.02 (8.15-10.00), 14.02 (14.15-18.00).

To be decided: deadline for handing in Compulsory Ex 1. Suggestion: Friday 22.02 at 16?

6. Linear Model Selection and Regularization

[Ch 6, TGM] 2019-w8 (PL: 18.02, PL: 21.02, E: 21.02)

- Subset selection
- Shrinkage methods (ridge and lasso)
- Dimension reduction with principal components
- Issues when working in high dimensions

7. Moving Beyond Linearity

[Ch 7, AS/ML] 2019-w9 (PL: 25.02, IL: 28.02, E: 28.02)

- Polynomial regression
- Step functions
- Basis functions
- Regression and smoothing splines
- Local regression
- Generalized additive models

8. Tree-Based Methods

[Ch 8, ML/TR] 2019-w10 (PL: 04.03, PL: 07.03, E: 07.03)

- Classification and regression trees
- Trees vs linear models
- Bagging, boosting and random forests

9. Support Vector Machines

[Ch9, ML] 2019-w11 (PL: 11.03, IL: 14.03, E: 14.03)

- Maximal margin classifiers
- Support vector classifiers
- Support vector machines
- Two vs. many classes
- SVM vs. logistic regression

10. Unsupervised learning

[Ch 10, TGM] 2019-w12 (PL: 18.03, PL: 21.03, E:21.03)

- Principal component analysis
- Clustering methods

11. Neural Networks

[ML] (PL: 25.03, PL: 28.03, E: 28.03)

- Network design and connections to previous methods
- Fitting neural networks
- Issues in training neural networks

Part 2: Modules 6-11

is finished with compulsory exercise 2. In week 14 we have supervision in Smia 01.04 (8.15-10.00), 04.04 (14.15-18.00).

To be decided: deadline for handing in Compulsory Ex 2. Suggestion: Friday 12.04 at 16?

12. Summing up and exam preparation

[ML] 2019-w15 (PL: 08.04, E:11.04)

- Overview - common connections
- Exam and exam preparation.

Who are you - the students?

Todo: Find out about kahoot!!

In class - go to kahoot.it to answer these questions. Answers in class added, with 51 respondents.

Study programme

- MTFYMA FysMat ()
- BMAT ()
- MSMNFNA Master in Mathematical Sciences ()
- Other ()

Study year

- 1 or 2 ()
- 3 ()
- 4 ()
- 5, >5 or PhD ()

Have you/will you take TMA4267 Linear Statistical Models?

- Yes, previously= in 2019 or earlier ()
- Yes, now= in 2020 ()
- Yes, planned for 2021 or later ()
- Not planned ()

Do you know R?

- No ()
- No, and I do not want to learn R ()
- Yes, but only the basics ()
- Yes, in depth ()

Plenary lectures on Monday mornings at 8.15-10 - do you plan to attend?

- Yes ()
- No, this is too early ()
- No, since I rarely attend lectures. ()
- No, for some other reason ()

What do you think will be most fun in TMA4268?

- Learning new statistical theory ()
- Trying out statistical theory in R ()
- Analysing data ()
- Learn new hot topics ()

Practical details

- go to Blackboard

Guest access

Course information-Course modules-R resources-Compulsory exercises-Reading list and resources-Exam

Reference group

At least 3 members, one from ideall different programmes

- At least one from IndMat, year 3
- Any programme, year 4
- Not IndMat

Volunteers?

Key concepts (stats) and notation

Todo – compare to hand-written notes by Mette (M1L1notes.pdf);
Perhaps delete this slide then

Plan for this week: Workshop for R and RStudio

Thursday January 10 at 14.15-18.00 in Smia

Say here what the plan is

Please do this before the next lecture

- Install R (use the Norwegian CRAN mirror):
<https://www.r-project.org>
- Install Rstudio <https://www.rstudio.com/products/rstudio/>

If you need help on installing R and RStudio on you laptop computer, contact orakel@ntnu.no.

R, Rstudio, CRAN and GitHub - and R Markdown

1. What is R? <https://www.r-project.org/about.html>
2. What is RStudio? <https://www.rstudio.com/products/rstudio/>
3. What is CRAN? <https://cran.uib.no/>
4. What is GitHub and Bitbucket? Do we need GitHub or Bitbucket in our course?
<https://www.youtube.com/watch?v=w3jLJU7DT5E> and <https://techcrunch.com/2012/07/14/what-exactly-is-github-anyway/>
5. What is R Markdown?

1-minute introduction video:

<https://rmarkdown.rstudio.com/lesson-1.html>

Then, if more is needed also a chapter from the Data Science book:

<http://r4ds.had.co.nz/r-markdown.html>

We will use R Markdown for writing the Compulsory exercise reports in our course.

6. What is `knitr`? <https://yihui.name/knitr/>

A first look at R and R studio

- `Rbeginner.html`
- `Rbeginner.pdf`
- `Rbeginner.Rmd`

A second look at R and probability distributions

- [Rintermediate.html](#)
- [Rintermediate.pdf](#)
- [Rintermediate.Rmd](#)

To see solutions added to the files, add -sol to filename to get

- [Rintermediate-sol.html](#)

And resources about plots

- [Rplots.html](#)
- [Rplots.pdf](#)
- [Rplots.Rmd](#)

To see solutions added to the files, add -sol to filename to get

- [Rplots-sol.html](#)

Additional nice R resources

- P. Dalgaard: Introductory statistics with R, 2nd edition, Springer, which is also available freely to NTNU students as an ebook: [Introductory Statistics with R](#).
- Grolemund and Hadwick (2017): “R for Data Science”, <http://r4ds.had.co.nz>
- Hadwick (2009): “ggplot2: Elegant graphics for data analysis” textbook: <https://ggplot2-book.org/>
- [Overview of cheat sheets from RStudio](#)
- Questions on R: ask the course staff, colleagues, and [stackoverflow](#).

Acknowledgements

Thanks to Julia Debik for contributing to this module page.