Module 1: Introduction

TMA4268 Statistical Learning V2020

Stefanie Muff, Department of Mathematical Sciences, NTNU

6th January, 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!

Learning outcomes of TMA4268

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.

Learning material

1. **The main learning source is the textbook by James, Witten, Hastie, Tibshirani (2013)**: "An Introduction to Statistical Learning". The textbook can be downloaded here: https://www-bcf.usc.edu/~gareth/ISL/
    - The ebook an also be downloaded from Springer: https://www.springer.com/gp/book/9781461471370 (NB, need to be on NTNU network or via vpn.)
    - There are 15 hours of youtube videos by two of the authors of the book, Trevor Hastie an Rob Tibshirani -the inventors of statistical learning - all links here
2. All the lecture notes, **including any classnotes** made on the board (not necessarily available online).
3. **Additional reading material will be clearly indicated in the modules and on the course page.**

Course page

All the relevant information for the course can be found here:

https://wiki.math.ntnu.no/tma4268/2020v/start

On each module page, all the relevant learning material and exercises (incl. solutions) will be provided in due time.

The Statistical Learning Team 2020

The TAs:

- Martina Hall; PhD student
- Michail Spitieris; PhD student

The Lecturers

- Stefanie Muff; Associate Professor
- Thiago Guerrera Martins; NTNU/AIAscience (Modules 6 & 10)
- Andreas Strand; PhD student (Module 7)

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 (probability theory and statistical methods), 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 prerequisist 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).
- 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.

Teaching philosophy

- Divide the topics of the course into modular units with specific focus.
- This (hopefully) facilitates learning?
- Two weeks without lectures (but supervision of exercises)

Course content: The 12 Modules

- **Module 1**: Introduction (this module)
- **Modules 2 - 11**:
    2. Statistical learning
    3. Multiple linear regression
    4. Classification
    5. Resampling methods
    6. Model selection/regularization
    7. Non-linearity
    8. Support vector machines
    9. Tree-based methods
    10. Unsupervised methods
    11. Neural nets
- **Module 12**: Summing up

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) ($2 \times 10\% = 20\%$).
- 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 are given in English.

The lectures

Mondays at 10.15-12.00 in S1 and Fridays at 14.15-16.00 in S4

- We have $2 \cdot 2$ hours of lectures every week (except when working with the compulsory exercises).
- **Note**: The **first lecture of each module will be on Fridays**, the second lecture on **Mondays**, and the exercise that corresponds to this module on Thursdays. See here for a tentative schedule.
- Some weeks the Monday lecture will be *interactive* or with some self-study / exercise component, where *active learning* is in focus.
- The other weeks (modules 3, 4, 6, 8, 10 and 11) the Monday lecture is a plenary lecture in S1.
- **I suggest that you always bring your laptop to the Monday lecture.**
- In the **first week of the course** we need time for an R workshop!

The weekly supervision sessions

Thursdays 08.15-10:00 in Smia

- For each module *recommended exercises* are uploaded. These are partly
  - theoretical exercises (from book or not)
  - computational tasks
  - data analysis
- These are supervised in the weekly exercise slots.
- Solutions will be provided to check yourself (no grading).

The compulsory exercises

- We will have **two compulsory exercises**, each with a maximal score of 50 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, the written exam the remaining 80%.

- The **first compulsory exercises** will be held after Modules 1-5.

  Suggested submission deadline:

  Thursday, 20. Februar 2020, 12:00h.

- The **second compulsory exercise** will be held after Modules 6-11.

  Suggested submission deadline:

  Monday, 13. April 2020, 12:00h.

Tentative schedule

A tentative schedule (i.e., with continous updates) can be found under the following link (als available from our course page):

https://github.com/stefaniemuff/statlearning/blob/master/TMA4268_schedule2020.pdf

The lecture material

- All the material presented in class will be available on our course webpage (https://wiki.math.ntnu.no/tma4268/2020v/start).
- There will be both a .pdf and an .Rmd version of the lecture notes. This will allow you to check and use the code that I use to generate the slides.
- For exercises, we provide a .pdf, .Rmd and an .html version.

Student active learning

Student's learning styles are different! Felder and Silverman (1988) suggested the following learning style axes:

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)? Many students have a visual learning style!
4. **sequential - global**: How do you make progress: sequentially (with continuous steps), or globally (through leaps and an integral approach)?

We try!

- ... to acknowledging these different learning style axes.
- ... to choose teaching styles that match the learning styles of as many students as possible.
- ... to provide learning environments, opportunities, interactions, and tasks that help to learn deeper.
- ... to provide guidance and support that challenges students based on their current ability.

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

Active vs. reflective learning styles

Reflective learning methods

- Plenary lectures
- Reading textbook
- Self study

Active learing methods

- Pause in plenary lecture to ask questions and let students think and/or discuss.
- In-class quizzes
- Projects - individual or in groups
- Group discussion
- Interactive lectures

Test your learning style

If you are interested in your learning style, we are very grateful if you can fill out this questionnaire:

https://innsida.ntnu.no/forms/view.php?id=221738

- The intro is in Norwegian.
- The questionnaire is in English.
- If you have questions, contact Mettee Langaas (mette.langaas@ntnu.no).
- You will get an email with your results.
- We have been given permission to collect these data for research.

Who are you - and what are your expectations?

In class - go to https://app.klicker.uzh.ch/join/bkx to answer these questions.

Reference group

At least 3 members, ideally one from different programmes

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

Volunteers?

Module 1

Aims of the first module

- An introduction to statistical learning. What is it?
- Types of problems we will look at
- **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 (Introduction) and 2.3 (Lab: Introduction to R).
- `Rbeginner`, `Rintermediate`, and `Rplots`

Recommended:

- Watch the video lecture for Chapter 1 by Hastie and Tibshirani here.
- Background on Matrix Algebra: Härdle and Simes (2015) - Chapter 2: 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).
- Main distinction: *Supervised* versus *unsupervised learning*.
- Both **prediction** and **understanding** (inference $\rightarrow$ drawing conclusions).
- Statistical learning is **a statistical discipline**, but the boarders are becoming more blurred.

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 both fields.

Statistical Learning vs. "Data Science"

Data science

- The aim is to extract knowledge and understanding from data.
- 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 (acquisition 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:

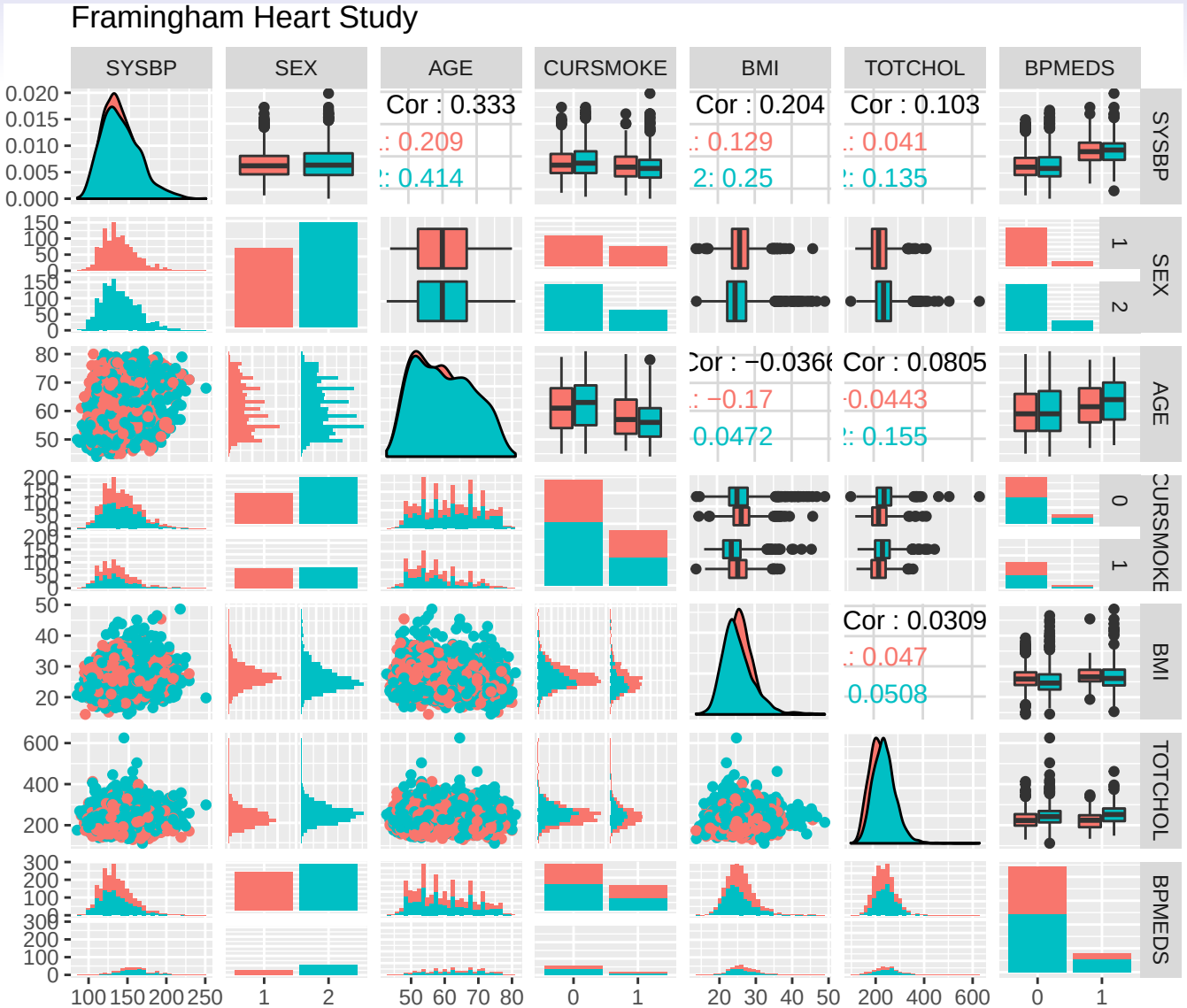
- Regression (supervised)
- Classification (supervised)
- Unsupervised methods

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, boxplots, barplots

Etiology of CVD

The question: **What are the factors that cause high CVD?**

So we are interested in *inference* (explanation) and not predction!

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

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

  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.

```r
modelB = lm(-1/sqrt(SYSBP) ~ SEX + AGE + CURSMOKE + BMI + TOTCHOL + BPMEDS,
    data = thisds)

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}
- **four features measured** for each corresponding sample:
  - `Sepal.Length`
  - `Sepal.Width`
  - `Petal.Length`
  - `Petal.Width`.

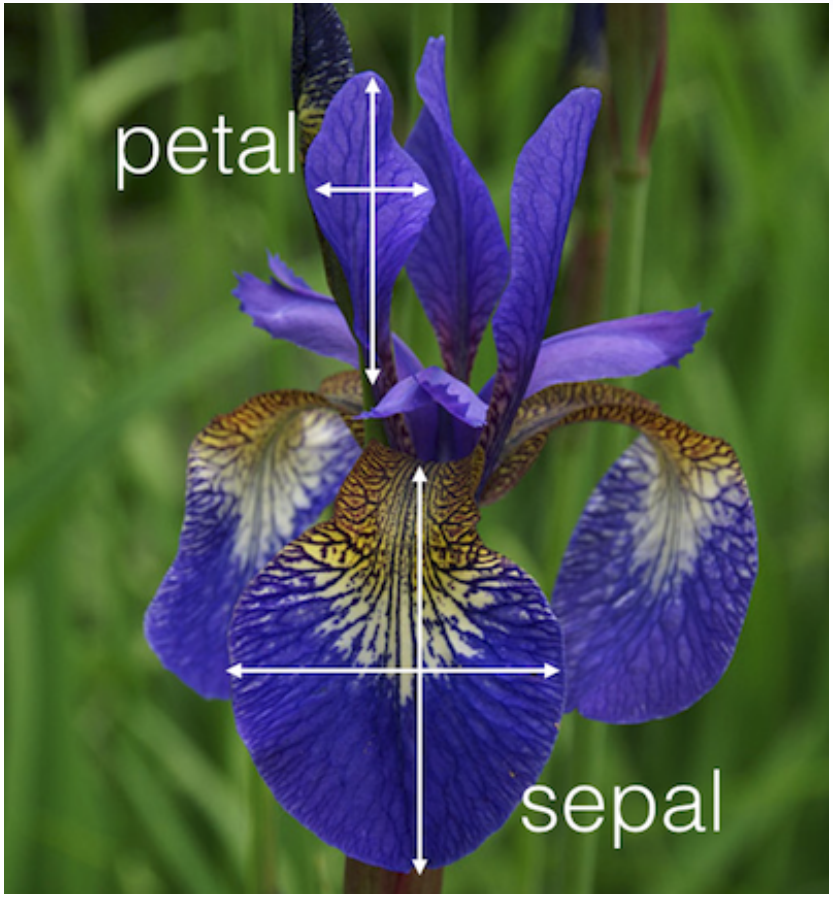


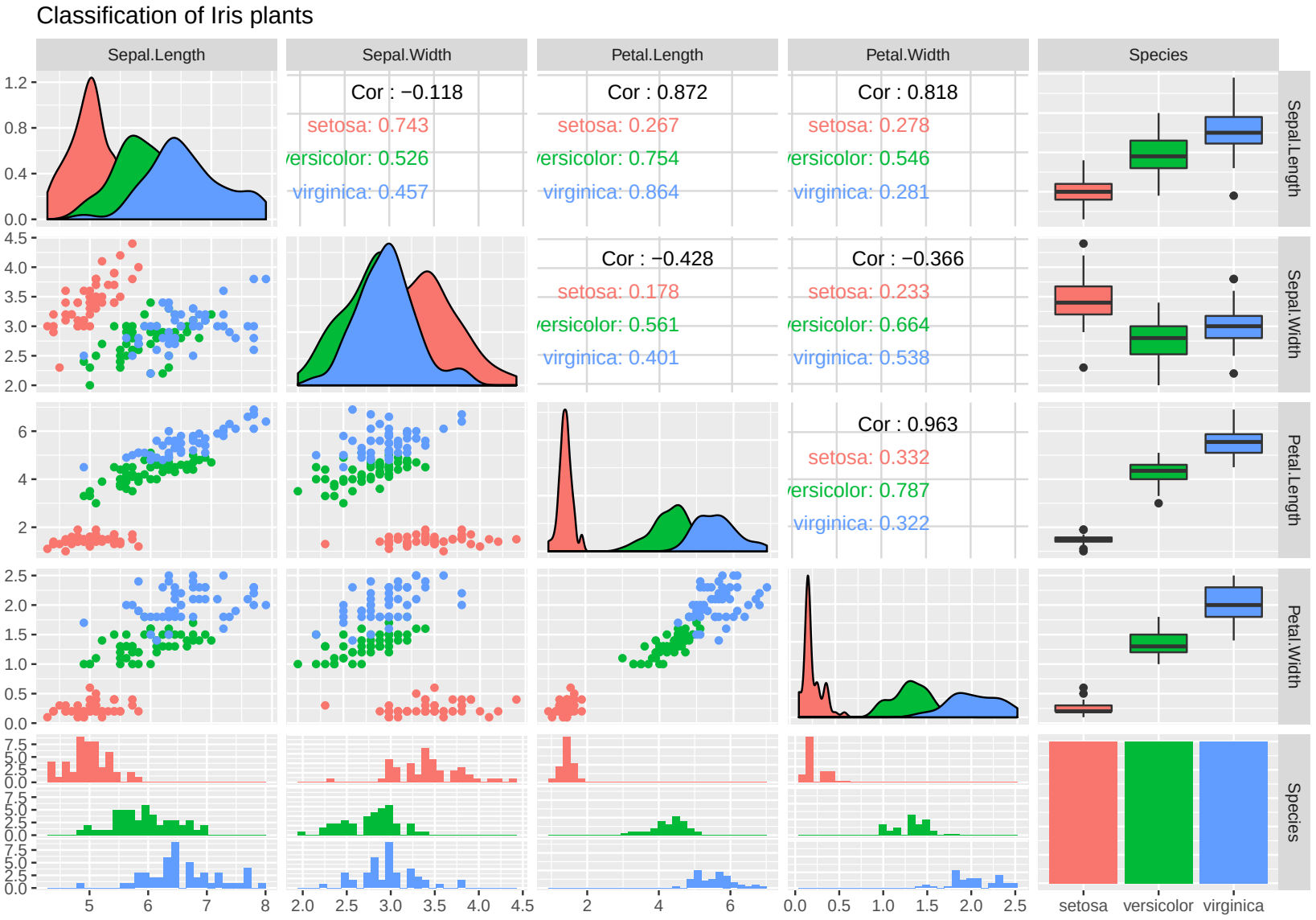
Figure 1: Iris plant with sepal and petal leaves

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

```r
head(iris)
```

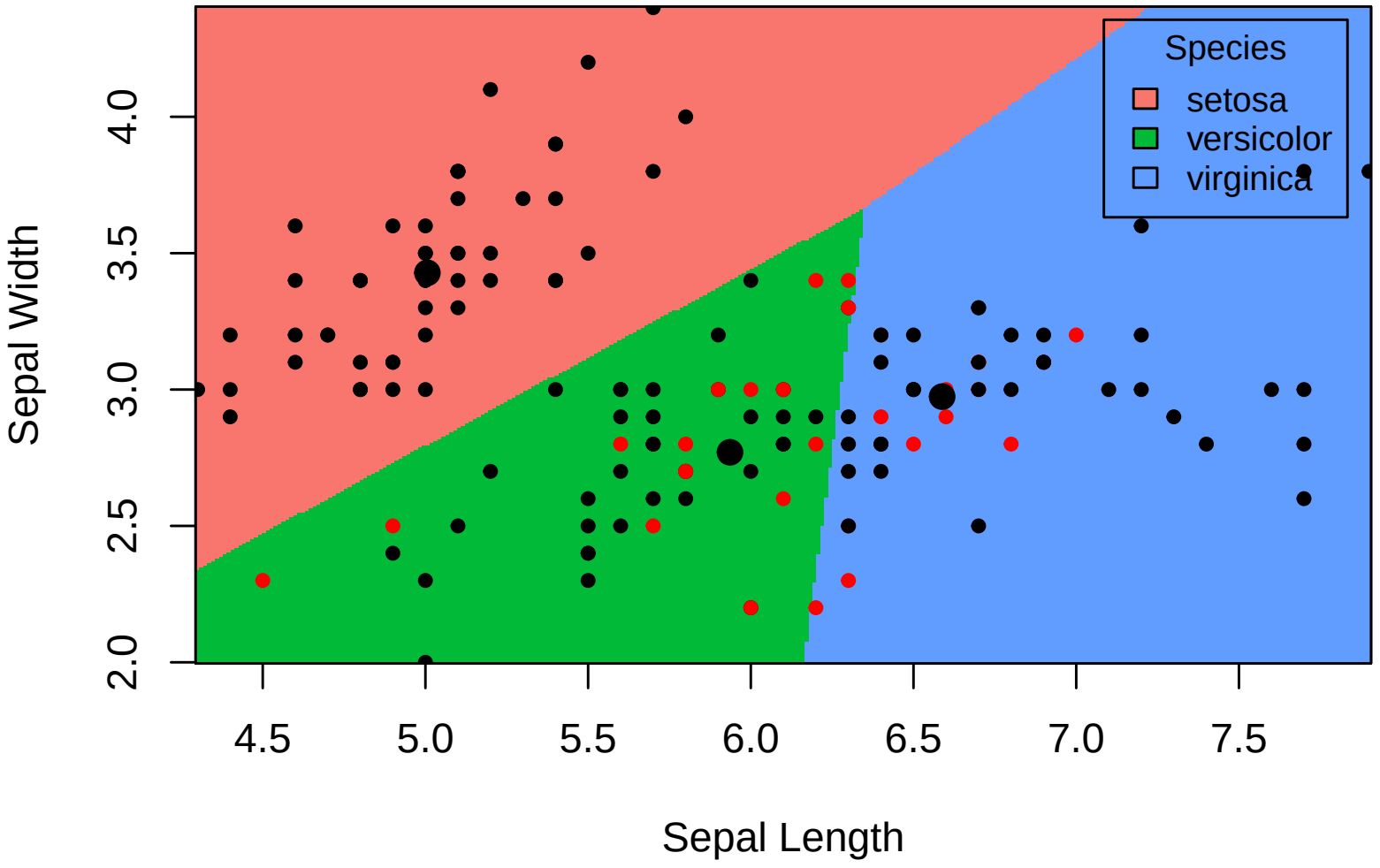
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5         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.



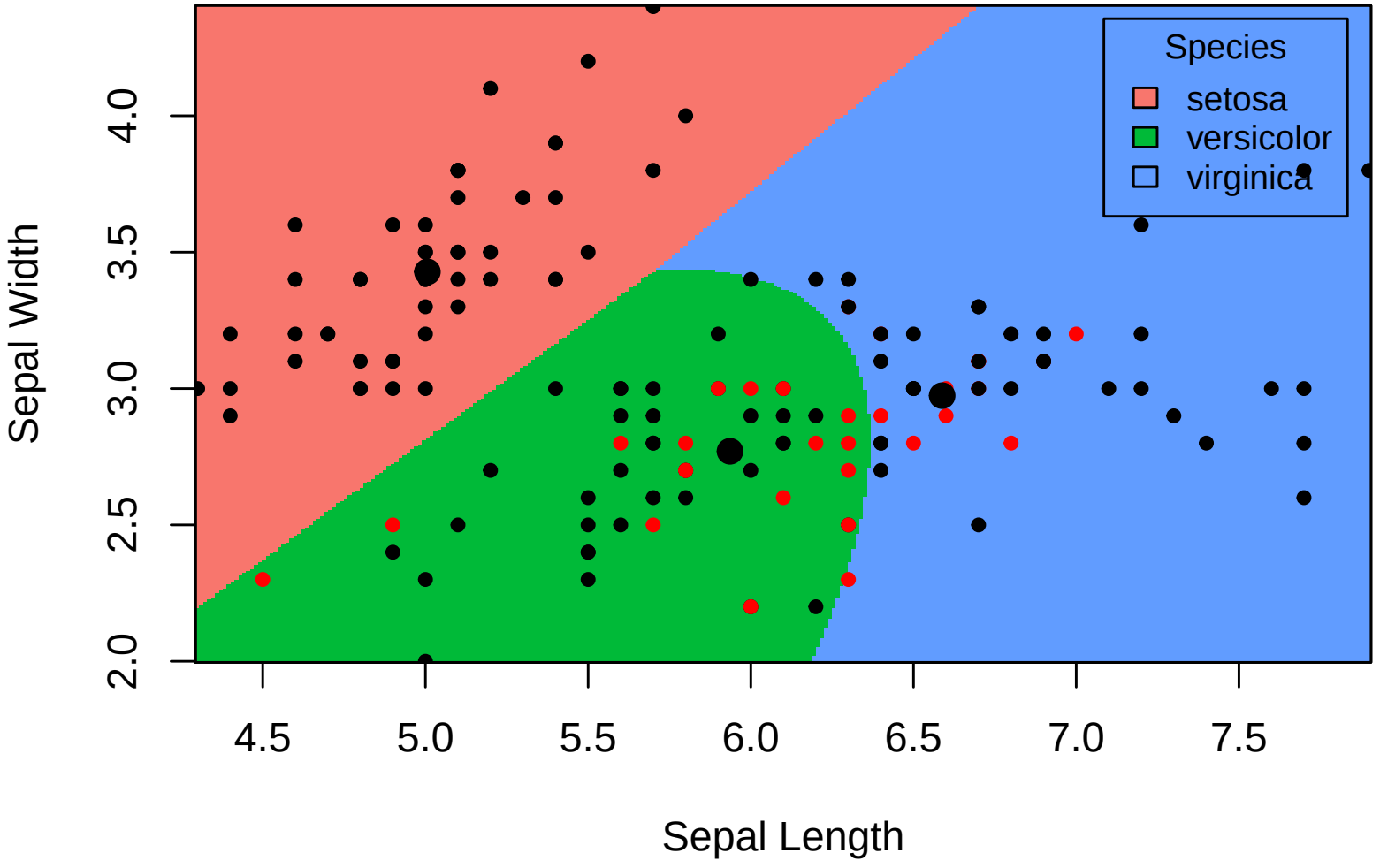
Linear boundaries

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 seditary 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.

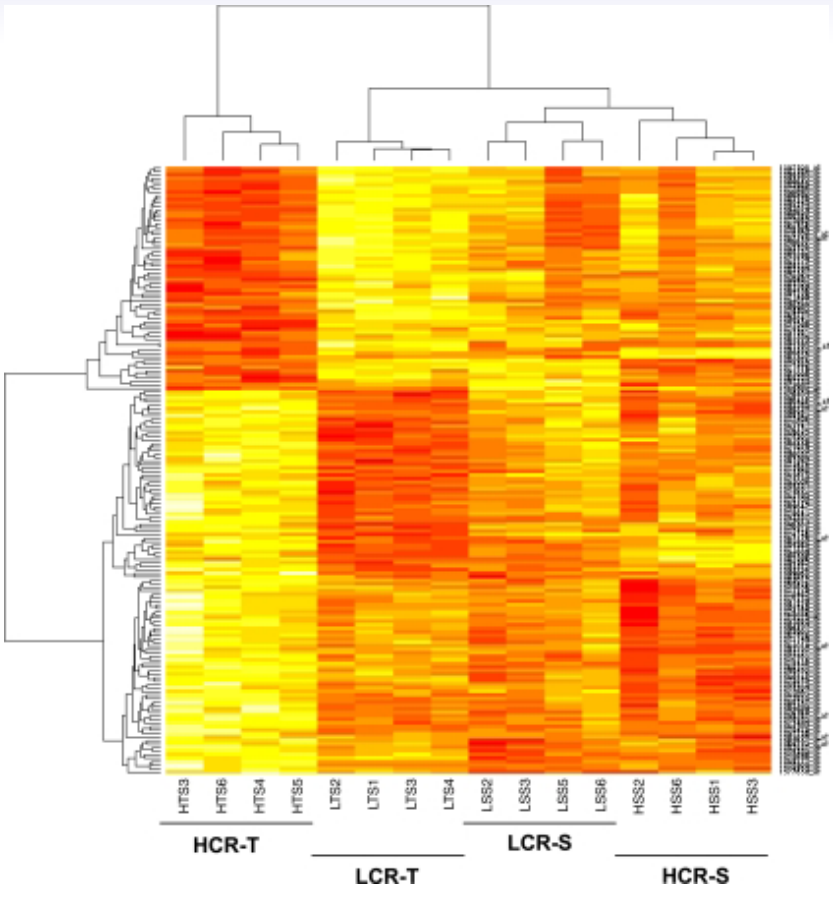


Figure 2: Heat map of the most significant transcripts. Transcripts with a high expression are shown in red and transcripts with a low expression are shown in yellow.

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

Example 4: Unsupervised methods (Network clustering)

Finding clusters in protein-protein-interaction networks.

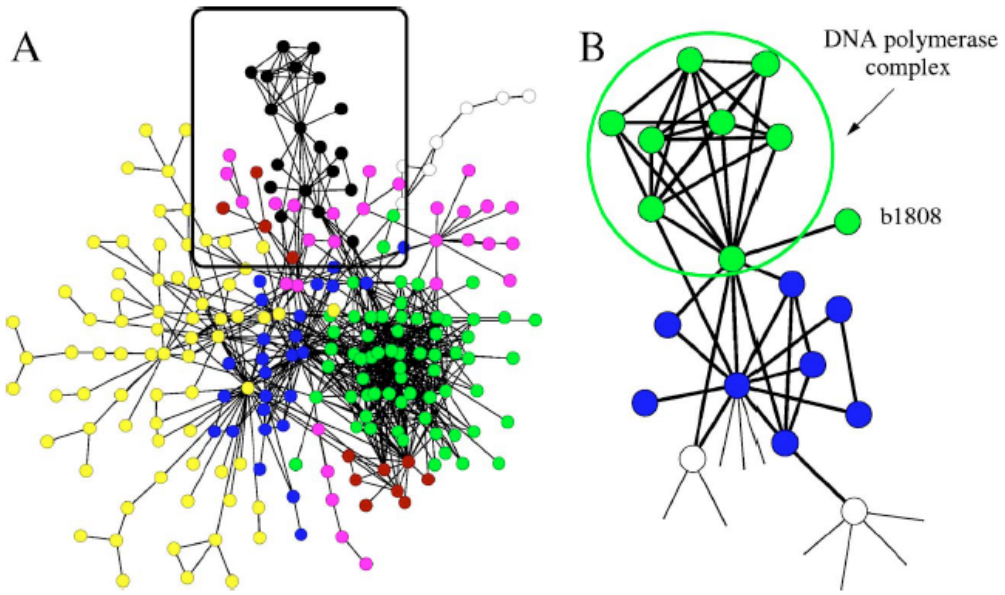


FIG. 4. (Color online) (a) Largest connected component of the PPI of *E. coli*. The colors represent the clusterization found by optimizing modularity. (b) *LQ* clusterization of the black *Q* cluster. The green circle contains proteins belonging to the DNA polymerase complex. The unknown protein b1808 is assigned to this complex according to *LQ* while the complete *Q* cluster is heterogeneous.

Plan for this week

Thursday January 9, 8.15-10.00 in Smia

- Workshop for R and RStudio
- Stefanie, Martina and Michail present

Friday January 10, 14.15-16.00 in S4

- Lecture 1 of Module 2

Getting started with R

- 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

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/

Getting to work with RStudio

→ Short demo in class.

A first look at R and RStudio

- Rbeginner.pdf
- Rbeginner.Rmd

A second look at R and probability distributions

- 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.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.

Supplement: Exam from 2019

This is the digital exam held in Inspera.

- Problem set
- Tentative solutions
- Grading document

However, please note that this was the exam created by Mette Langaas and her team, and not by us. The recommended and compulsory exercises will give you hints abou the exam, so we highly recommend you will work on them.