Module 1: Introduction

TMA4268 Statistical Learning V2023

Sara Martino, Department of Mathematical Sciences, NTNU

11th January, 2024

Acknowledgements

- This course had been built up by Mette Langaas at NTNU in 2018 and 2019. I am using a some 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!

- Thanks to Julia Debik for contributing to this module page.

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 the caveats are.

Learning material

1) **The main learning source** is the textbook by James, Witten, Hastie, Tibshirani (2021, 2nd edition): "An Introduction to Statistical Learning". The textbook can be downloaded here: https://www.statlearning.com/
    - There are 15 hours of youtube videos by two of the authors of the book, Trevor Hastie an Rob Tibshirani. Links will be added to the module subpages.
2) All the lecture notes (iPad notes will be made available online).
3) The R course here: https://digit.ntnu.no/courses/course-v1:NTNU+IE-IMF+2023_AUG/about
4) **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/tma4268/2024v/start

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

The Statistical Learning Team 2024

The TAs:

- Kenneth Aaase; PhD student
- Daesoo Lee; PhD student

The Lecturers

- Sara Martino; Associate Professor
- Stefanie Muff; Associate Professor

Who is this course for?

Primary requirements

- Bachelor level: 3rd year students 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!
- Advantage with knowledge of computing - for example 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, time to work on the compulsory exercises.

Course content: The 12 Modules

- **Module 1**: Introduction & R course (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) Tree-based methods 1
  9) Tree-based methods 2
  10) Unsupervised methods
  11) Neural networks (new in the course book edition 2)
- **Module 12**: Summing up

Learning methods, activities and grading

- Lectures, exercises and works (projects).
- The assessment is a 100% final school exam.
- To be allowed to the exam, you need to reach **at least 60% in both compulsory exercises**.
- Retake of examination may be given as an oral examination. The lectures are given in English.

The lectures

Thursdays at 8.15-10.00 in EL6 and Fridays at 12.15-14.00 in EL6

- We have 2×2 hours of lectures every week (except when working with the compulsory exercises).
- See here https://github.com/stefaniemuff/statlearning2/blob/main/TMA4268_schedule2024.pdf for a tentative schedule.
- **I suggest that you always have your laptop handy for the lecture.** So you can run code or do an exercise in class.

The first week: The R course

- In this **first week of the course** you will have to work through parts 1-6 of the R course on the openEdX page here: https://digit.ntnu.no/courses/course-v1:NTNU+IMF001+2020/course/
- Log in with your Feide account (scroll a bit down).
- There is a discussion forum that you can use for the R course.

Recommended exercises

Thursdays at 16.15-19.00 in EL6

- 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).
- Starting next week.

The compulsory exercises

- We will have **two compulsory exercises/projects**.
- Both projects need to be completed, where at least **60% of the points must be reached in both of them** to be admitted to the exam.
- The exercises/projects 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**; groups are formed in Blackboard (Bb). Also hand-in is via Blackboard.
- Written in R Markdown (both .Rmd and .pdf handed in).
- The TAs grade the exercises (pass/fail).

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

  Suggested submission deadline:

  ** **.

- The **second compulsory exercise** will be held after Modules 6-10.

  Suggested submission deadline:

  ** **.

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/statlearning2/blob/main/TMA4268_schedule2024.pdf

The lecture material

- All the material presented in class will be available on our course webpage (https://wiki.math.ntnu.no/tma4268/2024v/start).
- There will be .pdf, .html and .Rmd versions of the lecture notes and exercises. This will allow you to check and use the code that is used therein.

The discussion forum on Mattelab

- Use our discussion forum on Mattelab for all course-relevant questions: https://mattelab2024v.math.ntnu.no/c/tma4268/6

- Avoid writing emails to the course staff. By posting your question on Mattelab we get the chance to answer the questions to everyone.

Who are you - and what are your expectations?

Log into www.menti.com and use the code 6390 7810.

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?

-
-
-

Thanks to the three people that volunteer.

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

Required:

- Our textbook James et al (2021): An Introduction to Statistical Learning - with Applications in R (ISL)[1].
  - Chapter 1 (Introduction)
  - 2.3 (Lab: Introduction to R)
- Go through parts 1 to 6 in the online R course: https://digit.ntnu.no/courses/course-v1:NTNU+IMF001+2020/course/

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).

[1] I do expect you to read the text book yourself

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 **inference** (understanding $\rightarrow$ 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

- Aim: to extract knowledge and understanding from data.
- Requires a combination of statistics, mathematics, numerics, computer science and informatics.

This encompasses the whole process of

1. data acquisition/scraping
2. going from unstructured to structured data
3. setting up a data model
4. implementing and performing data analysis
5. interpreting and communicating 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:

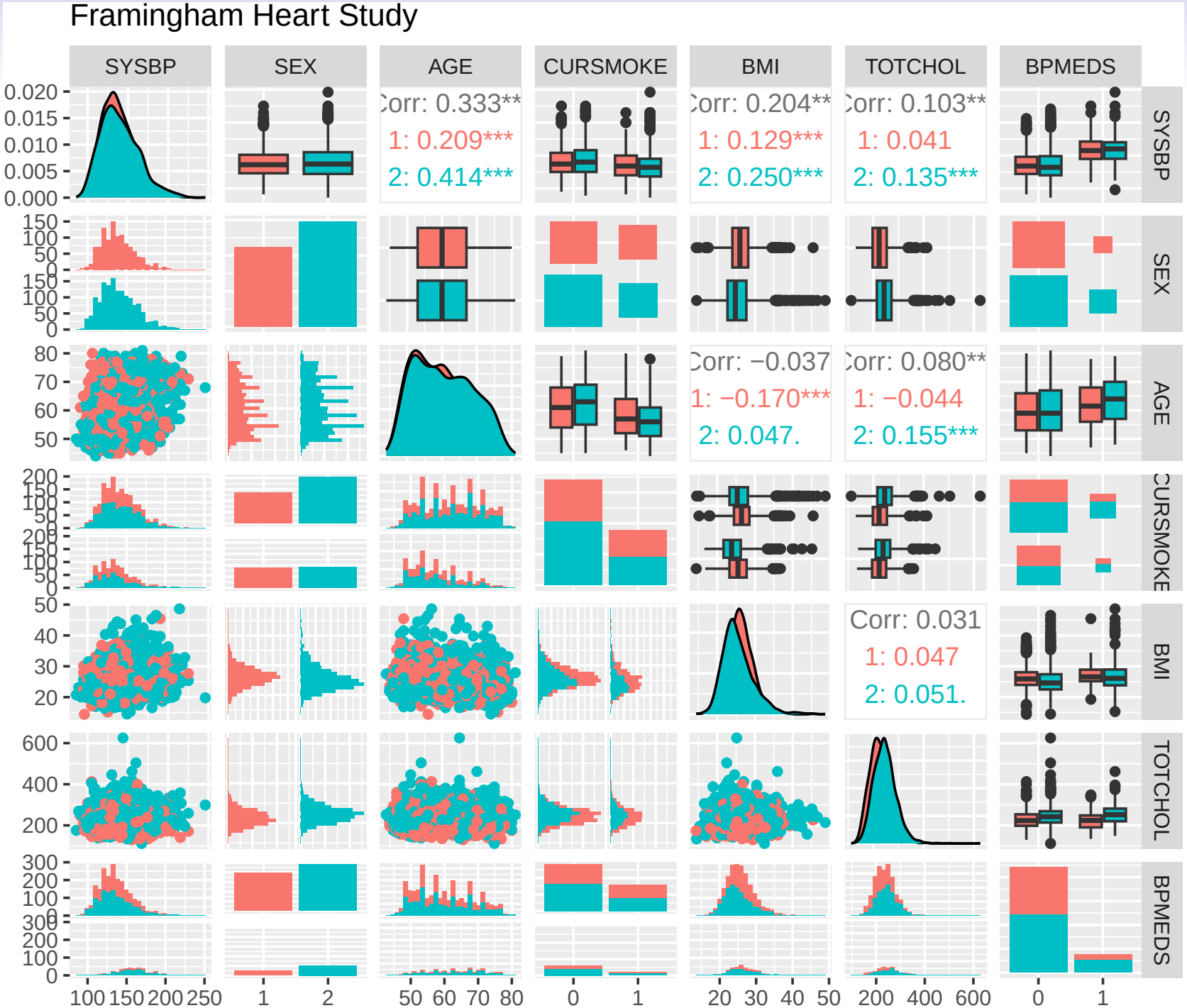
- Regression (supervised)
- Classification (supervised)
- Unsupervised methods

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.


Framingham Heart Study
SYSBP
SEX
AGE
CURSMOKE
BMI
TOTCHOL
BPMEDS
Corr: 0.333**
1: 0.209***
2: 0.414***
Corr: 0.204**
1: 0.129***
2: 0.250***
Corr: 0.103**
1: 0.041
2: 0.135***
Corr: −0.037
1: −0.170***
2: 0.047.
Corr: 0.080**
1: −0.044
2: 0.155***
Corr: 0.031
1: 0.047
2: 0.051.

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

We use `sex` to color the graph.

Etiology of CVD

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

$\rightarrow$ we are interested in *inference* (explanation), not prediction!

- 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`.

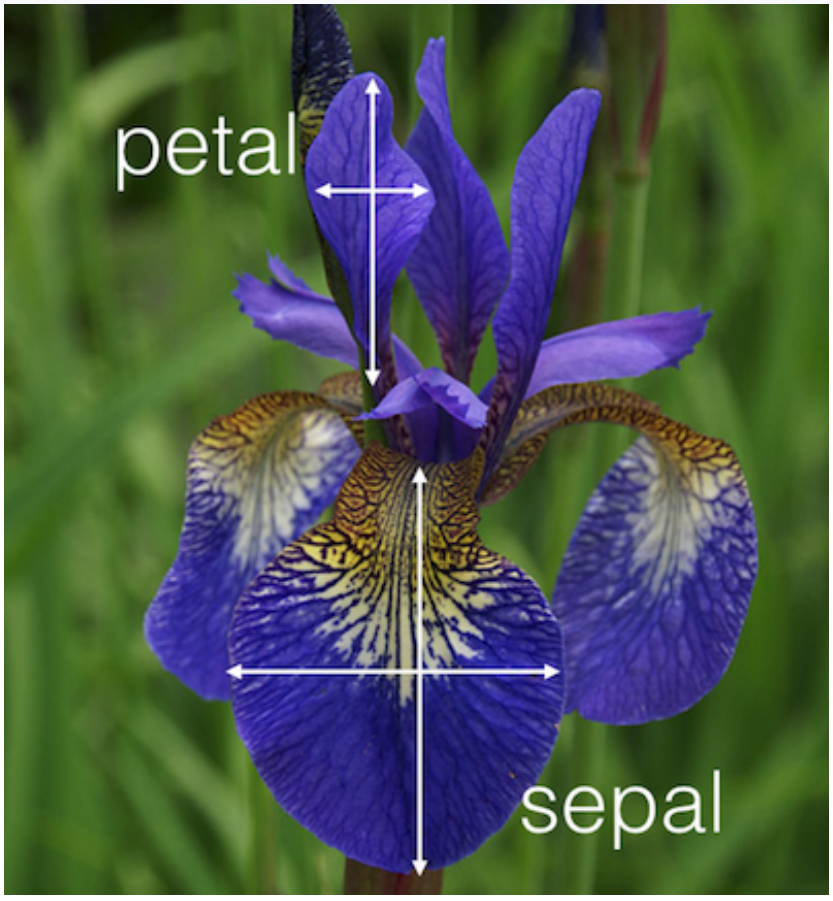


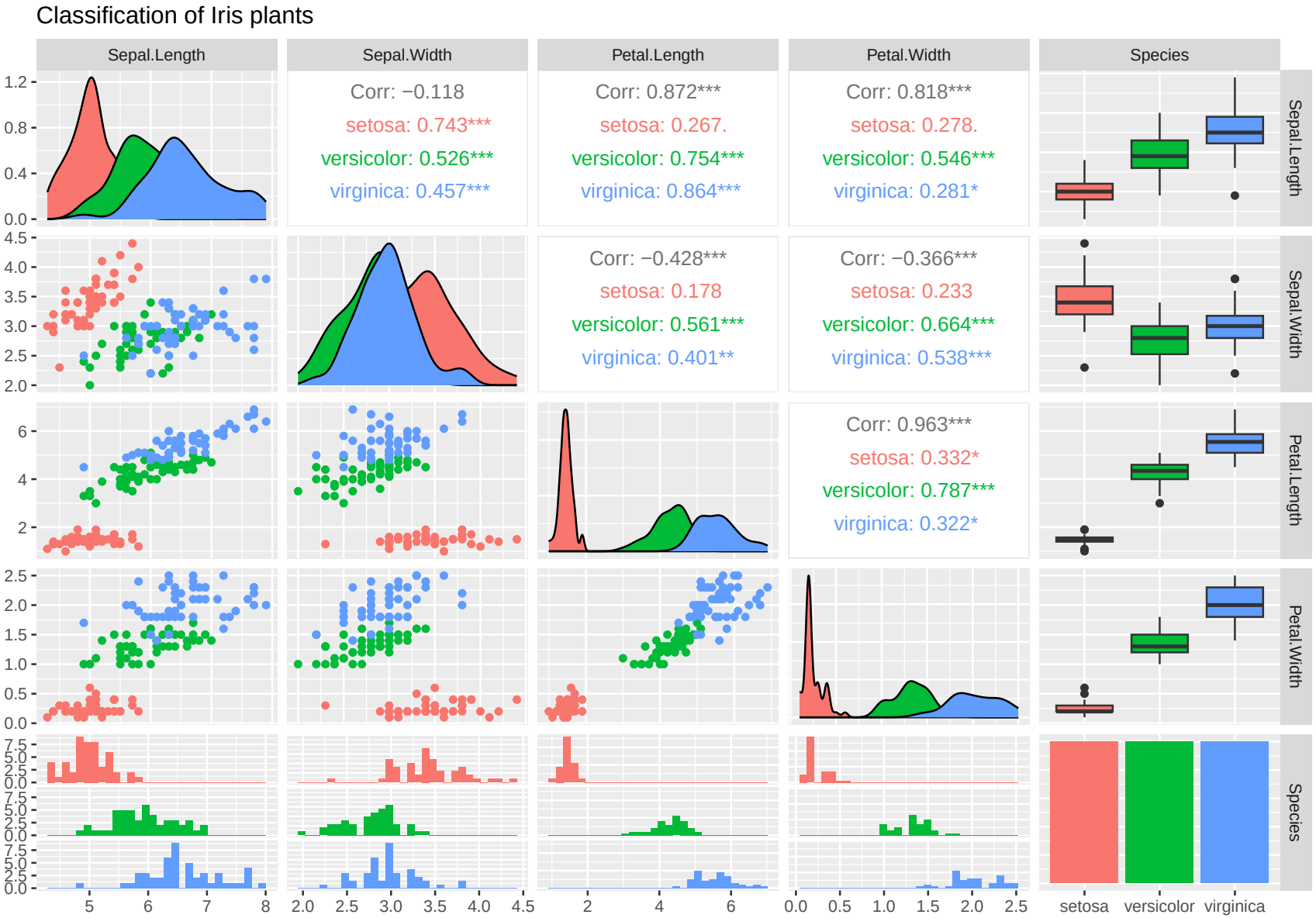
Figure 1: Iris plant with sepal and petal leaves

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

```r
head(iris)
```

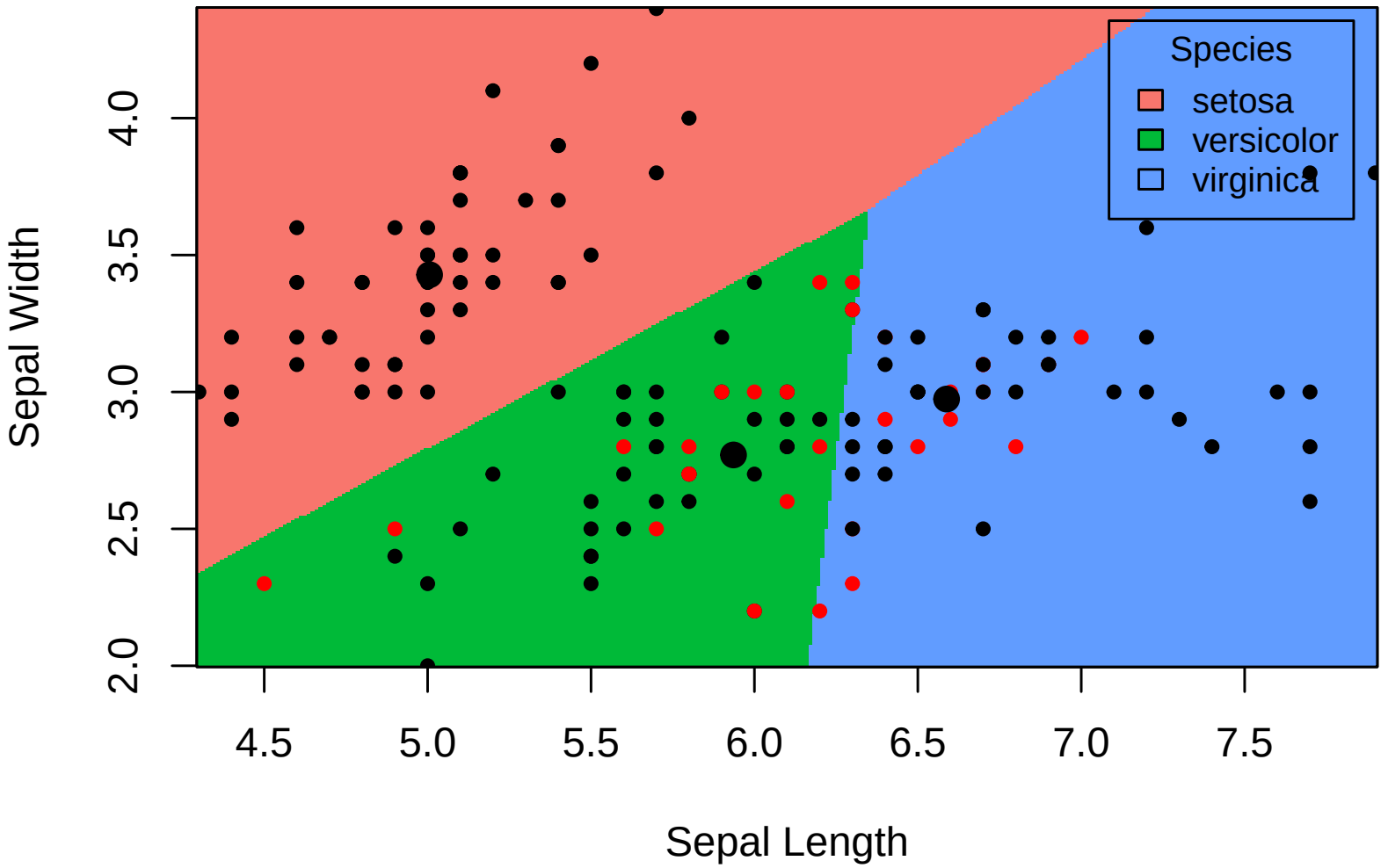
```
##   Sepal.Length Sepal.Width Petal.Length Petal.Width Species
## 1          5.1         3.5          1.4         0.2  setosa
## 2          4.9         3.0          1.4         0.2  setosa
## 3          4.7         3.2          1.3         0.2  setosa
## 4          4.6         3.1          1.5  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



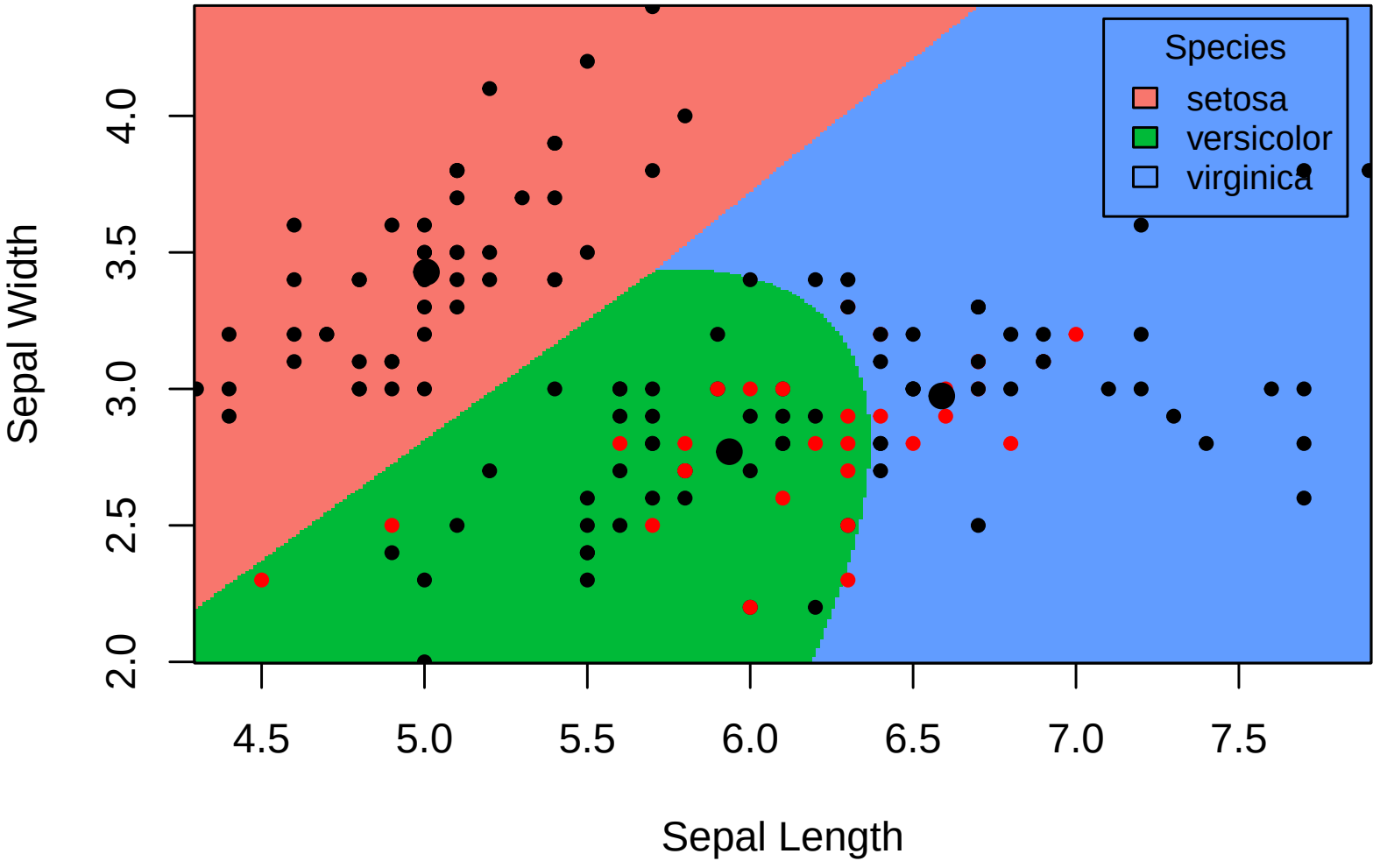
Linear boundaries

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 non-linear boundary is more suitable.



Example 3: Unsupervised methods (Gene expression)

(Check also the gene expression example in chapter 1 of the course book)

- 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.
- Heat maps showing the expression level for the most significant transcripts were presented graphically.
- This is *hierarchical cluster analysis* with pearson correlation distance measure (module 10).

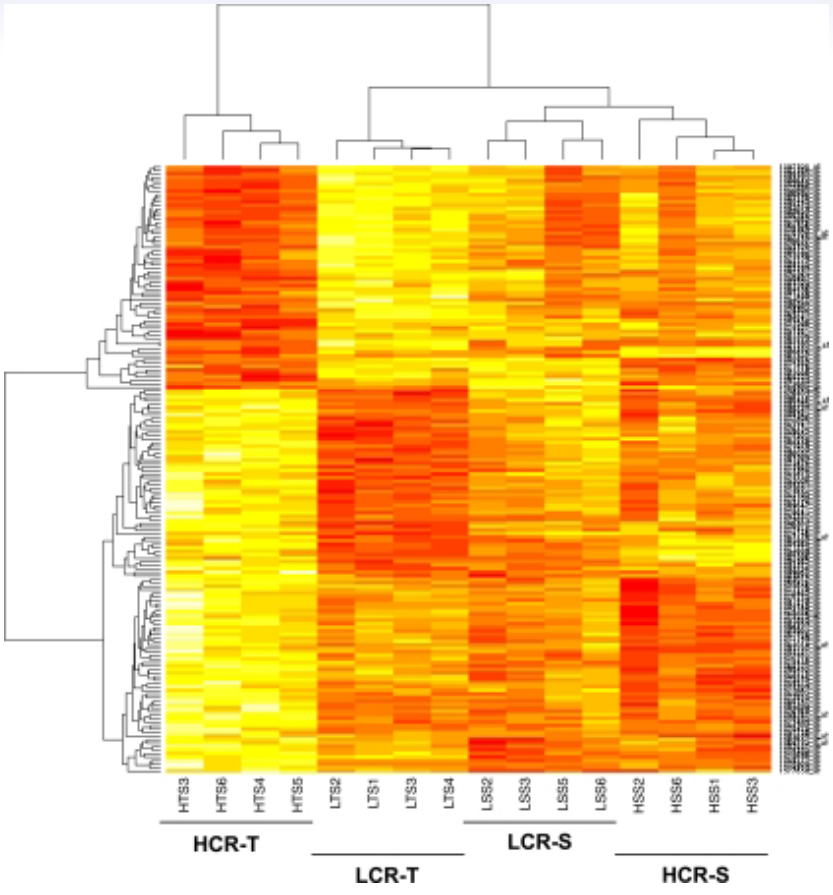


Figure 2: Heat map of the most significant transcripts. Transcripts with a high expression are shown in red and transcripts with a low expression are shown in yellow.

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

Example 4: Unsupervised methods (Network clustering)

Finding clusters in protein-protein-interaction networks.

Plan for rest of the week

- You can work through parts 1 - 6 in the R course https://digit.ntnu.no/courses/course-v1:NTNU+IE-IMF+2023_AUG/about
- Ideally use your Feide account to log in.
- There is a discussion forum (click on the “Discussion” tab).

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.

Some additional links

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/

Additional nice R resources

- 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.