MSA220/MVE440 Final exam

Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Spring semester 2019
Deadline: 14th June 2019

1 General info

Examinator: Rebecka Jörnsten

Course coordinator: Felix Held

Official start date: 27th May 2019

Hard Deadline: 14th June 2019

How to submit: There are two activities on PingPong where you upload

1. your **final report in PDF format**. Upon submission, the document will be automatically sent to Urkund[1]. Plagiarism is not allowed.
2. your **code as a ZIP file**.

Remember that you have to **submit both!**

Grading:

- Project reports determine pass or fail, i.e. if you pass this part you are guaranteed a 3 at Chalmers and a G at University of Gothenburg (GU)
- The additional data analysis tasks determine if you get a higher grade
  - For a 4 at Chalmers or a VG at GU you need to give the equivalent of one well worked-out answer
  - For a 5 at Chalmers you need to give two well worked-out answers

  By "equivalent of one well worked-out answer" we mean that you either do really well on only one additional task, or reasonably well on both tasks. See below for what is included in a well worked-out answer.

2 Report on the projects

Observe

- This part of the exam determines **pass or fail**.
- Reports are **individual**
- **One A4 page text** per project. Figures and tables extra
- Data, real or simulated, can be described on a separate page.

The purpose of this first exercise is that you **individually** revisit the four projects that you worked on during the course. Each report should be written in standard academic style and should be **conceptually structured into four sections**:

[1] `https://www.urkund.com/`, a plagiarism checker that makes it very easy for us to see if your reports overlap, and to which extent

1. **Introduction:** Clearly formulate the task/research question **your group** tackled in the project.

   Answer the question: "What is the key question/problem?"

2. **Methods:** Clearly describe your approach, e.g. simulation setup, data set used, methods and models, and motivate why you chose this approach. If you are using a method that was discussed in the lecture, it is enough to state which method (and potentially in which setup/with with parameters) was used and why.

   Answer the question: "What was done to answer the key question **and why did we do it like this**?"

3. **Results:** Describe your findings short and concisely. Focus on model/method critique and insight into performance (e.g. comparing methods, for different data, for different simulation settings, ...).

4. **Discussion:** State the main take-home message of your project (**important!**) and, if applicable,

   - discuss your result in the context of what the other groups did,
   - or give an outlook of what you could do in the future, including expected results and motivation

Important:

- Write your report as clear as possible! No fluffy babbling to fill space.
- Motivate, motivate, motivate! Give clear motivations for why you use methods/do the analysis/approached the problem as you did.
- If your group separated the work, you still have to focus on the whole of the project for the report.
- Correct any errors in your analysis that you found during the presentations or while re-visiting the project.
- Do not contradict your findings. Just because you expect to observe something does not mean that you actually do. In this situation, reason about why you get results different from your expectations.
- You do not have to agree with your group. If you come to a different conclusion than your group did, then you have the chance to clarify this now in the report.
- You also have the chance to update and improve on your analysis/investigation, even if it was correct, since you have acquired more knowledge about the topics as the semester progressed.

3 Additional data analysis tasks

Observe

- Once you pass the reports, these questions determine if you get a higher grade
- Each question has to be answered in the same style as the reports (Introduction, Methods, Results, Discussion) and will be judged by the same measures
- A **well worked-out answer** is one where
  1. you chose and use methods appropriately,
  2. do not contradict your numerical results,
  3. and give clear and concise answers in each required section
- The only difference to the report section is that there is no strict page limit and that there is no need for an outlook/comparison to the other groups

3.1 Regression in big-n

In this task we want you to explore variable select and variable importance in case of big-n. You are provided with four datasets; (A) two with $p = 10$ variables and samples sizes $n = 100$ and $n = 10^5$, respectively; (B) two with $p = 800$ variables and sample sizes $n = 1000$ and $n = 10^5$, respectively. The same non-zero betas were used to generate data for each of (A) $\boldsymbol{\beta}_A$ and (B) $\boldsymbol{\beta}_B$ for the two different sample sizes. In addition, you are given the numeric response vectors $\mathbf{y} \in \mathbb{R}^n$ for each dataset. (Data is on PingPong)

Your task is to build a model for each dataset, i.e. determine which variables should be included in the model and how important they are. Put special emphasis on the differences in necessary/possible approaches in the small-n *vs* the big-n case.

Guidance: You could look into

- Linear modelling framework with penalisation (Lasso (normal, adaptive, SCAD, ...), elastic net, group lasso, ...)
- Standard forward or backward stepwise selection techniques
- R^2 measures of variable importance[2]
- Random Forest regression variable importance[3]

3.2 High-dimensional clustering

Suppose you are approached by the marketing department of a large company. They have collected a lot of different variables on some of their customers. You are provided with a dataset with $n = 229$ and $p = 598$ in a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. (Data is on PingPong)

The company wants you to find clusters in the data, so that they can focus their marketing on fewer collective groups instead of having to personalise their advertisement for each customer. In addition, they would like to know which variables explain the final clustering best.

Throughout this task, it is crucial that you motivate your decisions with clear numerical and/or graphical results. Otherwise management will not believe you.

[2] see e.g. the techniques described in Groemping (2006) Relative Importance for Linear Regression in R: The Package relaimpo. Journal of Statistical Software 17 DOI 10.18637/jss.v017.i01

[3] see e.g. the techniques described in the lecture and Genuer et al. (2017) Random Forests for Big Data. Big Data Research 9:28-46 DOI 10.1016/j.bdr.2017.07.003