MSA220/MVE441 Exam June 2021

Mathematical Sciences
Chalmers University of Technology and University of Gothenburg

Spring semester 2021
Deadline: 11th June 2021

General info

Examiner: Rebecka Jörnsten

Course coordinator: Felix Held

Official start date: 27th May 2021

Hard Deadline: 11th June 2021

Setup: Since this is a take-home exam you are allowed to freely use the course material and similar resources available to you. However, in contrast to the projects, your submissions must be **individual** and we will judge your performance **individually**.

How to submit: There are separate assignments on Canvas where you upload

1. your answers as a report in **PDF format**. Please upload your answer to each question separately at the corresponding assignment. Note that this has to be a **written report** and not a presentation as in the projects. Upon submission, the document will be automatically sent to Ouriginal[1], a plagiarism checker. Plagiarism is not allowed and will be reported[2].

2. your **code as a ZIP file**. It is allowed to submit Jupyter notebooks containing the code. Code has to be your own and submissions with suspiciously similar/identical results will be compared for code similarity.

You have to **submit both** for a valid submission! This means, submitting answers to at least one question and the corresponding code is the bare minimum for a valid submission.

Grading:

- During correcting, the three exam questions will be weighted equally.
- Grades are determined as follows
  - To pass the exam (3 at Chalmers, G at GU) your need to achieve the equivalent of 1.5 complete answers (see definition on next page).
  - If you only give partial answers or parts of your answer are incorrect, your performance across answers will be averaged weighted by how well you performed on each question. **To maximise your chances of passing and getting a high grade make sure to submit answers to all exam questions.** In our experience submitting answers to only two exam question is a risky game.
  - At Chalmers, the equivalent of two complete answers results in a four, and three complete answers result in a five.
  - At GU, to get a VG, you need to give complete answers to two questions and perform at least acceptable on the third question.

[1] https://www.ouriginal.com/

[2] Chalmers has a document that demonstrates how to avoid plagiarism (https://student.portal.chalmers.se/en/chalmersstudies/policy-documents/Documents/20090920_Academic_Honesty.pdf)

Formalities

Observe

- Answers should be given as short reports, about two pages. Figures and tables are considered separate.
- A **complete answer** is one where
  1. you choose and use methods appropriately,
  2. you discuss your results correctly,
  3. and give clear, concise, and correct answers to the questions asked.
- Write your answers as clearly as possible!
- Motivate, motivate, motivate! Give clear motivations for why you use methods/do a certain analysis/approach the problem the way you do.
- Do not contradict your findings. Always consider your discussions in light of your results.

Structure

The answers to the exercises have to be documented in a written report and should contain the following aspects.

1. **Methods:** Clearly describe your approach, simulation setup, chosen methods and models, ..., and motivate your choices. Also describe how you selected model parameters.

   Answer the questions: "**What** was done to answer the question(s) and **why** was it done this way?"

2. **Results:** Describe your findings short and concisely. Focus on results that are related to answering the questions. If you attach figures and tables you have to refer to them in your text. **Figure and table captions** must explain all elements of a figure or table.

3. **Discussion:** Interpret your results in light of the questions and argue why and how your results support your answer.

We do not accept Jupyter notebooks that have simply been converted to PDF without further consideration. Please also note that there should **not be any code in the report** (this includes copy-pasted tables from Jupyter/REPL output) and the report should be **understandable and contain all relevant information** (e.g. hyperparameter selection, special settings for methods, ...) **without us having to look at the code**.

Questions

Note: On Canvas you will find a link where you can download all datasets required for the questions below.

Question 1: Clustering

You are provided with a clustering dataset (`Q1_X.csv`) containing 560 samples and 974 features.

Tasks Determine the number of clusters in the dataset and find a way to visualise the best clustering in a sensible format.

Hints

- Make sure that you justify each decision you make.
- Think about whether dimension reduction is necessary.
- Think about the assumptions and limitations of the clustering algorithms you use.

Question 2: Feature selection

You are provided with a regression dataset containing a response vector $\mathbf{y} \in \mathbb{R}^n$ (`Q2_y.csv`) and two sets of features $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$ (`Q2_X1.csv`) and $\mathbf{X}_2 \in \mathbb{R}^{n \times p_2}$ (`Q2_X2.csv`).

Tasks Proceed in two steps

1. Determine the most important predictors for the response $\mathbf{y}$ using (A) only the features in $\mathbf{X}_1$ and then (B) both sets of features in $\mathbf{X}_1$ and $\mathbf{X}_2$ together (i.e. form the feature matrix $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \in \mathbb{R}^{n \times (p_1 + p_2)}$).
2. Describe the changes between your results from (A) and (B). Find explanations for the differences you observe.

Hints

- You should be able to justify your feature selection with confidence.

Question 3: Correcting mislabelled observations

You explored the impact of mislabelling on the performance of a classification algorithm in Project 1. Here, you will develop an algorithm to detect mislabelled observations.

Tasks Come up with an algorithm that detects mislabelled observations in a binary classification dataset. The method should work for a classification algorithm of your choice and for a moderate amount of random mislabelling, say, 5–10% of all observations independently of the original class.

Proceed in two steps

1. Simulate data for a binary classification problem, mislabel some of the observations and develop an algorithm that can detect the mislabelled observations.

   Show that your algorithm works by comparing the ground truth (since you simulated the data you know which observations are mislabelled) to the observations that were detected to be mislabelled. Make sure that your algorithm maximises true positives and minimises false positives.

2. Apply your algorithm to the supplied binary classification dataset (`Q3_labels_mislabelled.csv` and `Q3_X.csv`) and detect the mislabelled observations. Compare those found by your algorithm to the supplied ground-truth (`Q3_labels_correct.csv`). Determine the error rate as well as the number of true and false positives. Does your algorithm perform equally well for both classes?

Hints

- The correct labels should only be used as a test set and are not allowed to be used in the training stage of your model/algorithm.
- You are allowed to use ensembles of classification methods.
- Pay attention to how your algorithm handles observations close to the decision boundary.
- Observations that are predicted to be of the opposite class on repeated sub-samples are probably mislabelled.
- There is no point in cheating by manually tweaking the output of your algorithm in Step 2. It is important that you justify *what* and *why* you do, not if you get 70% or 90% of mislabelled observations correctly. Of course, a high rate of true positives while getting few false positives is desirable.