4th September, 2024

STK-IN4300/STK-IN9300: Statistical Learning Methods in Data Science

Mandatory assignment 1 of 2

Submission deadline

Thursday 19th SEPTEMBER 2024, 14:30 in Canvas (canvas.uio.no).

Instructions

Note that you have **one attempt** to pass the assignment. This means that there are no second attempts.

You can choose between scanning handwritten notes or typing the solution directly on a computer (for instance with LaTeX). The assignment must be submitted as a single PDF file. Scanned pages must be clearly legible. The submission must contain your name, course and assignment number.

It is expected that you give a clear presentation with all necessary explanations. Remember to include all relevant plots and figures. All aids, including collaboration, are allowed, but the submission must be written by you and reflect your understanding of the subject. If we doubt that you have understood the content you have handed in, we may request that you give an oral account.

In exercises where you are asked to write a computer program, you need to hand in the code along with the rest of the assignment. It is important that the submitted program contains a trial run, so that it is easy to see the result of the code.

Application for postponed delivery

If you need to apply for a postponement of the submission deadline due to illness or other reasons, you have to contact the Student Administration at the Department of Mathematics (e-mail: studieinfo@math.uio.no) no later than the same day as the deadline.

All mandatory assignments in this course must be approved in the same semester, before you are allowed to take the final examination.

Complete guidelines about delivery of mandatory assignments:

uio.no/english/studies/admin/compulsory-activities/mn-math-mandatory.html

Summarizing and visualizing data

In addition to performing statistical analyses and building prediction models, an important part of the data science pipeline is to summarize and present data in an understandable and concise way. This assignment will focus on summarization, presentation and visualization of data.

Find a dataset of your own choice, with the requirements that

- it contains at least five variables;
- at least one variable is continuous and at least one variable is categorical;
- there is one numerical response variable;
- it is different from any dataset used in the weekly exercises so far.

You may, for example, use a dataset from the UCI machine learning repository.

For writing the report, it is highly recommended to use R Markdown or some similar file format. For those that want to use R Markdown but prefer Python over R, the reticulate package can be a useful. However, most importantly, when writing the report, make sure to include the code you used to produce the results along with the resulting tables and figures, and of course any relevant comments and discussion.

Problem 1. Summary Statistics Table

Write a short description of your data and mention a possible use case for the data. In addition, make a summary statistics table (manually or by using any of the existing packages) that summarizes the information in the data on a variable level.

Problem 2. Bad Data Visualization

While a table of summary statistics provides a concise overview of the data, well-designed plots (or charts) can often be more effective in conveying a quick overview of the key characteristics of some data or results to a human. Badly designed plots, on the other hand, may not only be useless, but can also be directly misleading in terms of the information they provide.

For at least one categorical and one continuous variable in your data, make a bad plot and explain why it is bad and possibly even misleading. Here you may either make individual plots for each variable or make a single plot that involves several variables.

Problem 3. Good Data Visualization

Provide a good version of the bad plot(s) from Problem 2 and explain how the plot has been improved. For one out of many resources on data visualization, see https://clauswilke.com/dataviz/.

Problem 4. Simple analysis

Perform linear regression on the data, including some method for performing model selection.
Include some measure of performance on the final model.
Based both on the analysis and your previous visualizations, evaluate whether a linear model is sufficient.

GOOD LUCK!