2nd October, 2019

STK-IN4300 / STK-IN9300
Statistical learning methods in Data Science

Mandatory assignment 1 of 2

Submission deadline

Thursday 17th October 2019, 14:30 at Canvas.

Instructions

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. Students who fail the assignment, but have made a genuine effort at solving the exercises, are given a second attempt at revising their answers. 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) well before 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

GOOD LUCK!

Problem 1. Consider the data from Sinnaeve et al. (2009), available at https://www.ebi.ac.uk/arrayexpress/experiments/E-GEOD-12288/. The data consist of 22283 gene expressions measured for 222 people, some of them (110) has atherosclerotic coronary artery disease (CAD).

Consider only people with the disease (cases), and contrast the ability of a lasso and a ridge regression model to predict the Duke CAD index (`CADi`) from the gene expressions. Report the code and the results.

NB: You can use your favourite programming language. A possible R script to download the data is

```
library(BiocManager)
library(ArrayExpress)
system('mkdir E-GEOD-12288')
tmp <- getAE('E-GEOD-12288', type = 'processed',
       path = 'E-GEOD-12288')
tmpProc <- getcolproc(tmp)
tmpE <- procset(tmp, tmpProc[2])
info <- pData(tmpE)
# extract input matrix (gene expressions)
X <- t(exprs(tmpE))[info[, 40] == 'case', ]
# extract response (CADi)
CADi <- info[info[ , 40] == 'case', 37]
# remove temporary files
rm(info, tmp, tmpE, tmpProc)
```

Alternatively, you can download the data at https://www.uio.no/studier/emner/matnat/math/STK-IN4300/h19/data_o1.rdata.

Problem 2. Consider projection pursuit. Derive the expressions for w which minimizes the linearised expression for the object function

$$\sum_{i=1}^{N} g'(w_{\text{old}}^T x_i)^2 \left(\frac{y_i - g(w_{\text{old}}^T x_i)}{g'(w_{\text{old}}^T x_i)} + w_{\text{old}}^T x_i - w^T x_i \right)^2 .$$

Bibliography

Sinnaeve, P. R., Donahue, M. P., Grass, P., Seo, D., Vonderscher, J., Chibout, S.-D., Kraus, W. E., Sketch Jr, M., Nelson, C. & Ginsburg, G. S. (2009). Gene expression patterns in peripheral blood correlate with the extent of coronary artery disease. *PLoS ONE* **4**, e7037.