11th February, 2022

STK2100 - Machine learning and statistical methods for prediction and classification

Mandatory assignment 1 of 2

Submission deadline

Thursday 10th March 2022, 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

Spesific requirements to this assignment:

In order to get the assignment **accepted** you need to fullfil the following requirements:

- Include plots and code you have used as appendices to your report.
- A real attempt on **all** (sub-)questions in Problem 1 and Problem 2.
- A satisfactory answer in at least 2/3 of the (sub-)questions in Problems 1 and 2.
- Problem 3 is not required for the compulsory exercise to be fullfiled but *do* give good exam training

Remember that it is allowed both to collaborate and to ask for help!

Within the exercises several commands that can be used in **R** are included. These are also available from the webpage. If some libraries are not available at your computer, you need to install them, for example by

```
install.packages("MASS")
```

All the command listed in the assignment are aslo available in a separate .R file on the course webpage.

It is allowed to use other programs, but there will then be extra requirements to good documentation on what you have done and you can not expect to obtain help with respect to implementional details.

GOOD LUCK!

Problem 1. We will in this exercise look at a dataset `nuclear` and see how we can use regression for prediction of costs of light-water reactor. The data is available in the file `nuclear.dat` while a description of the data is available at `nuclear.txt`, both available from the course data webpage https://www.uio.no/studier/emner/matnat/math/STK2100/data/.

Our interest will be in the `cost` variabele while the other variables will be explanatory variables. Since `cost` is always positive, we will model this variable at the log-scale.

(a) Make the data available through the commands

```
datadir = "http://www.uio.no/studier/emner/matnat/math/STK2100/data/"
nuclear = read.table(paste(datadir,"nuclear.dat",sep=""),header=T)
n = nrow(nuclear)
```

Make also different plots in order to get some understanding of the data.

(b) We will first look at a model

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

where Y_i is `cost` at log scale for observation i.

What are the standard assumptions about the noise terms ε_i? Discuss also which of these assumptions that are most important.

Fit this model including all the observations with `log(cost)` as response and all the other variables as covariates.

Discuss the results.

(c) Now remove the variable with the highest corresponding P-value and fit the new model.

Why is this a reasonable procedure?

Discuss potential changes of the P-values for the remaining variables. You can relate this to correlations between the explanatory variables.

(d) Continue to remove explanatory variables until all P-values are less than 0.05. What is your final model?

Make different plots in order to evaluate whether the model is reasonable.

(e) Use the final model to predict response and make a model based on the average quadratic error $(\frac{1}{n}\sum_{i=1}^n(y_i - \hat{y}_i)^2)$ in order to evaluate how good the model is. Discuss weaknesses with such a procedure.

Assume now we want to predict `cost` for a new data point. More specifically we are interested in $\theta = E[Y|\boldsymbol{x}^*]$ as well as $\eta = E[\exp(Y)|\boldsymbol{x}^*]$ where $\boldsymbol{x}^*$ is defined by

```
d.new = data.frame(date=70.0,t1=13,t2=50,cap=800,pr=1,
                   ne=0,ct=0,bw=1,cum.n=8,pt=1)
```

Prediction can be performed by the command `predict.lm`.

The command below gives predictions for a fitted model for θ [1]

```
predict(fit ,d.new,interval="confidence")
predict(fit ,d.new,interval="predict")
```

where `fit` is the fitted model.

(f) Run the two commands. Discuss the differences between the two `predict` commands.

(g) The intervals given in the previous point is related to `Cost` on log-scale. Try to construct intervals for `Cost` on the original scale.

(h) Also try out Lasso regression on this data set.

If you use cross-validation for selection of the penalty parameter, which variables are then included in the final model?

Also compare this with the model you obtained earlier.

Hint: Look at the `Hitters_lasso.R` script.

Problem 2. We will in this exercise look at linear regression with quantitative (categorical) explanatory variables. We will look at a dataset from Devore & Berk (2012), exercise 11.5. The dataset consists of measurements of iron content in four different iron formations (Form, 1=carbonate, 2=silicate, 3=magnetite, 4 =hematite). The table below shows the data, with 10 observations within each type of iron formation.

Form	1	2	3	4	5	6	7	8	9	10
1	20.50	28.10	27.80	27.00	28.00	25.20	25.30	27.10	20.50	31.30
2	26.30	24.00	26.20	20.20	23.70	34.00	17.10	26.80	23.70	24.90
3	29.50	34.00	27.50	29.40	27.90	26.20	29.90	29.50	30.00	35.60
4	36.50	44.20	34.10	30.30	31.40	33.10	34.10	32.90	36.30	25.50

We want to test whether there are differences in iron content between the different formations.

(a) Read the data by the command

```
datadir = "http://www.uio.no/studier/emner/matnat/math/STK2100/data/"
Fe <- read.table(paste(datadir,"fe.dat",sep=""),header=T,sep=",")
```

Further, give the command

```
options(contrasts=c("contr.treatment","contr.treatment"))
```

(we will come back to what this command actually do).

Try out the following command

[1] Note that it is enough to use the generic function `predict` here due to that **R** understands that it is the underlying command `predict.lm` which do predictions for linear models is to be used. If you, however, want to see how the function works, you need to use `help(predict.lm)`. There you also can see how to include uncertainties in the predictions.

```
fit1 <- lm(Fe~form,data=Fe)
summary(fit1)
```

Why do this go wrong?

Then give the command

```
Fe$form <- as.factor(Fe$form)
```

Now try out the same call to `lm` again and a following `summary` command. Why do you get a more reasonable result in this case?

The actual model used above is based on introducing *dummy* variables

$$x_{i,j} = \begin{cases} 1 & \text{if } c_i = j; \\ 0 & \text{otherwise.} \end{cases}$$

where c_i defines the category for which observation i belongs to. The model is then

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \cdots + \beta_K x_{i,K} + \varepsilon_i \tag{1}$$

where $K = 4$ in this case.

(b) From the summary command (after you used the `as.factor` command) you should get a regression table where there is no row correspondning to β_1. The specific `options` command given above actually include the constraint $\beta_1 = \hat{\beta}_1 = 0$.

Why is such a constraint necessary?

What interpretations do the other β_j parameters then have?

(c) An alternative constraint is to put $\beta_0 = 0$. This can be obtained by

```
fit2 <- lm(Fe~form+0,data=Fe)
summary(fit2)
```

What is the interpretation of the β_j's in this case?

(d) The constraints $\beta_1 = 0$ or $\beta_0 = 0$ are denoted by *contrasts* in the linear regression terminology. The `contr.treatment` used above correponds to putting the regression coefficient for the first category equal ot zero. An aternative is

```
options(contrasts=c("contr.sum","contr.sum"))
fit3 <- lm(Fe~form,data=Fe)
summary(fit3)
```

in which case a constraint/contrast $\sum_{j=1}^{K} \beta_j = 0$ is imposed. The `summary` command will still only give 4 rows in the regression table, not including the row corresponding to β_4 in this case. How can you obtain $\hat{\beta}_4$?

(e) Do the results indicate that there are differences between the formations? Which of the fitted models do you find most suitable for answering this question?

(f) Now try out the commands

```
newdata = data.frame(form=as.factor(c(1,2,3,4)))
pred1 = predict(fit1 ,newdata)
pred2 = predict(fit2 ,newdata)
pred3 = predict(fit3 ,newdata)
```

Compare the three predictions and comment on the results.

(g) Based on the summary outputs from the different models, is it possible to simplify the model in some way?

Hint: Not all the different outputs will tell the same story here.

Problem 3. This problem is a contiuation of Problem 2, but with more focus on mathematical derivations. These are **not** needed to be fulfilled in order to get the compulsory exercise accepted but do provide good exam training.

In the following we will use different symbols for the parameters in the different alternatives. In particular we write

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \beta_4 x_{i,4} + \varepsilon_i \qquad \beta_1 = 0 \tag{2}$$

$$Y_i = \alpha_0 + \alpha_1 x_{i,1} + \alpha_2 x_{i,2} + \alpha_3 x_{i,3} + \alpha_4 x_{i,4} + \varepsilon_i \qquad \alpha_0 = 0 \tag{3}$$

$$Y_i = \gamma_0 + \gamma_1 x_{i,1} + \gamma_2 x_{i,2} + \gamma_3 x_{i,3} + \gamma_4 x_{i,4} + \varepsilon_i \qquad \sum_{j=1}^{4} \gamma_j = 0 \tag{4}$$

(a) Show that the three models are equivalent. Write an explicit relationship between $\boldsymbol{\beta} = (\beta_0, \beta_2, \beta_3, \beta_4)$, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3, \alpha_4)$ and $\boldsymbol{\gamma} = (\gamma_0, \gamma_1, \gamma_2, \gamma_3, \gamma_4)$.

We will in the following concentrate on model (3) since this version is somewhat simpler mathematically.

(b) Let $\boldsymbol{X}$ be the design matrix for model (3), that is the ith row of $\boldsymbol{X}$ contains the values $x_{i,j}, j = 1, ..., K$. Show that $\boldsymbol{X}^T\boldsymbol{X}$ will be a diagonal matrix with diagonal elements n_j where n_j is the number of observations with $c_i = j$.

Also show that $\boldsymbol{X}^T\boldsymbol{y}$ is a vector where the j-th element is equal to element $\sum_{i:c_i=j} y_i$ (the sum of all responses corresponding to category j).

Based on this, derive formulas for the least squares estimates for $\alpha_1, ..., \alpha_K$. Discuss whether these estimates are reasonable.

(c) Based on the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, construct formulas for the estimates for $\boldsymbol{\beta}$.

Argue why also these estimates are least squares estimates for $\boldsymbol{\beta}$.

(d) Based on the relationship between $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$, construct formulas for the estimates for $\boldsymbol{\gamma}$.