STK2100 – Machine Learning and Statistical Methods for Prediction and Classificatio

Mandatory assignment 1 of 2

Submission deadline

Thursday March 6 2025, 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.

Specifically about this assignment

Include the code that you have used in the report.

Complete guidelines about delivery of mandatory assignments:

`www.uio.no/english/studies/admin/compulsary-activities/mn-math-mandatory.html`

GOOD LUCK!

Problem 1

A list of regression models for the response Y and a single covariate x is given below:

1. $Y = \frac{\beta_0}{1+\beta_1 x} + \beta_2 x^{1/2} + \varepsilon$
2. $Y = \beta_0 + \frac{\beta_1}{x} + \beta_2 x^2 + \varepsilon$
3. $Y = \frac{1}{\beta_0+\beta_1 x+\beta_2 x^2+\varepsilon}$
4. $Y = \beta_0 + \beta_1 x^{\beta_2} + \varepsilon$
5. $Y = \beta_0 x^{\beta_1} \varepsilon$

a) Which of these models can be written as a linear regression model

   i) as they are?

   ii) when a parameter is kept fixed? If so, which parameter?

   iii) after a suitable transformation of Y? If so, which transformation?

b) For each of the models, specify the form of the model matrix $\mathbf{X}$ and the vector of regression coefficients.

Problem 2

We will in this exercise look at a dataset `nuclear`, and see how we can use linear regression to predict the cost of building light-water reactors. The data is available in the file `nuclear.dat` while a description of the data is given in `nuclear.txt`, both availble on the webpage `https://www.uio.no/studier/emner/matnat/math/STK2100/data/`. Since `cost` is always positive, we let the reponse variable of the linear regression model be $y = \log(\texttt{cost})$, and the remaining 10 variables are treated as covariates. Relevant R code for solving this problem can be found both in `r-code-week5.r`, `r-code-week6.r` and `r-code-week7.r`, which you will find in the time table of the course, under January 28, February 4 and February 19, respectively.

a) Start by fitting a linear regression model with $y = \log(\texttt{cost})$ as the response, including all 10 covariates. Construct 95% confidence intervals for the βs corresponding to the covariates `t1`, `t2` and `bw`.

b) We know that for Y with corresponding covariate vector $\mathbf{x}$ and prediciton $\widehat{Y} = \mathbf{x}^T\hat{\boldsymbol{\beta}}$, we have

$$\frac{Y - \widehat{Y}}{\hat{\sigma}\sqrt{1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}} \sim t_{N-p-1},$$

so that

$$Pr\left(\widehat{Y} - t_{1-\alpha/2,N-p-1}\hat{\sigma}\sqrt{1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}} \leq Y \leq \widehat{Y} + t_{1-\alpha/2,N-p-1}\hat{\sigma}\sqrt{1 + \mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}\right) = 1-\alpha.$$

Thus, a $(1-\alpha)\cdot 100\%$ prediction interval for Y is given by

$$\widehat{y} \pm t_{1-\alpha/2;N-p-1}\hat{\sigma}\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}.$$

Now, let Z be the cost, and recall that $Y = \log(Z)$. Show that a $(1-\alpha)\cdot 100\%$ prediction interval for Z is given by

$$\exp\left(\widehat{y} \pm t_{1-\alpha/2;N-p-1}\hat{\sigma}\sqrt{1+\mathbf{x}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{x}}\right) = (\exp(L), \exp(U)),$$

where (L, U) is the $(1-\alpha)\cdot 100\%$ prediction interval for Y (you can use the results concerning the prediction interval for Y without showing them). Then compute the 95% prediction interval for the cost Z with corresponding covariate vector $\mathbf{x} = (70.0, 13, 50, 800, 1, 0, 0, 1, 8, 1)^T$ (note that a prediction interval for Y can be obtained with the R command `predict` with `interval="predict"`).

c) Make individual hypothesis tests of the type $H_0 : \beta_j = 0$ against $H_A : \beta_j \neq 0$ for the covariates `t1`, `t2` and `bw`. Also make a joint test for whether all the three βs corresponding to these covariates are 0. What can you conclude from these hypothesis tests? Looking at these and other results from the fitting in a), what could be the reasons to exclude some of the covariates?

d) Perform a forward selection of covariates, for instance using the `regfitsubsets` command from the `leaps` package in R, with `method="forward"`. In which order are the covariates included in the model? Use both the AIC and the BIC to choose the best subset. Which subsets do you end up with? Discuss the differences.

e) Perform a backward selection of covariates, for instance using the `regfitsubsets` command from the `leaps` package in R, with `method="backward"`. In which order are the covariates excluded from the model? Use both the AIC and the BIC to choose the best subset. Which subsets do you end up with? Discuss the differences. Comparing all the subsets selected with forward and backward selection, which is the best according to the AIC and BIC, respectively?

f) From d) and e), you have potentially 4 different "best" models, depending on whether you have used forward or backward selection, and whether you have used the AIC or the BIC as a criterion. You should now assess which of these models is best when estimating the prediction error with K-fold cross-validation with $K = 10$. Which of these models would you now prefer?

g) Repeat f), but now using the bootstrap to estimate the prediction error. More specifically, you should use the "0.632 estimator", with $B = 1000$ bootstrap samples.

h) Instead of doing subset selection, we now want to use shrinkage. Perform ridge regression on the data set, using 10-fold cross-validation to choose the

penalty parameter λ, for instance using the `cv.glmnet` function from the R package `glmnet`, with `alpha=0`. Assess the prediction error of the chosen model using 10-fold cross-validation, as in f). How does this model compare to the best subset models?

i) Perform Lasso regression on the data set, using 10-fold cross-validation to choose the penalty parameter λ, for instance using the `cv.glmnet` function from the R package `glmnet`, with `alpha=1`. Assess the prediction error of the chosen model using 10-fold cross-validation, as in f). How does this model compare to the best subset models and the ridge regression model from h)?