STK2100 – Machine Learning and Statistical Methods for Prediction and Classificatio

Mandatory assignment 2 of 2

Submission deadline

Thursday April 24 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 person responsible for the course, Ingrid Hobæk Haff (e-mail: ingrihaf@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.

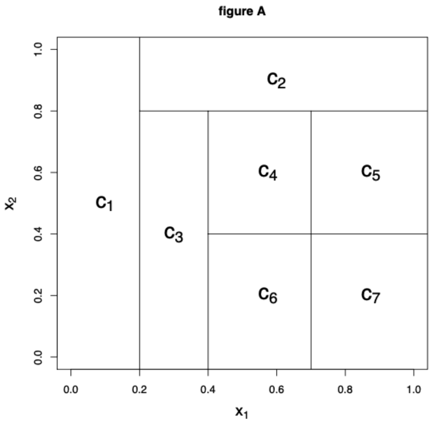
Complete guidelines about delivery of mandatory assignments:

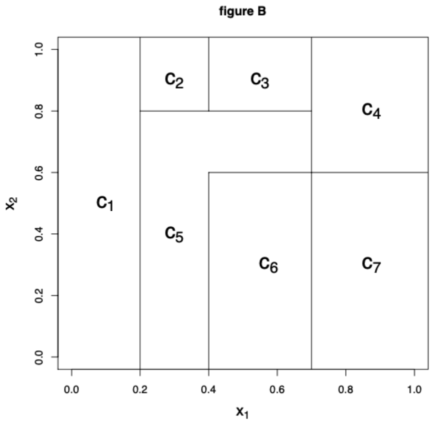
`www.uio.no/english/studies/admin/compulsary-activities/mn-math-mandatory.html`

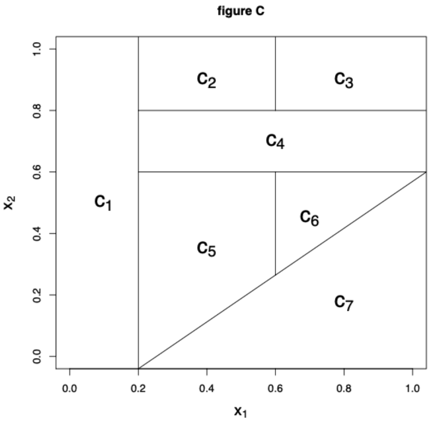
GOOD LUCK!

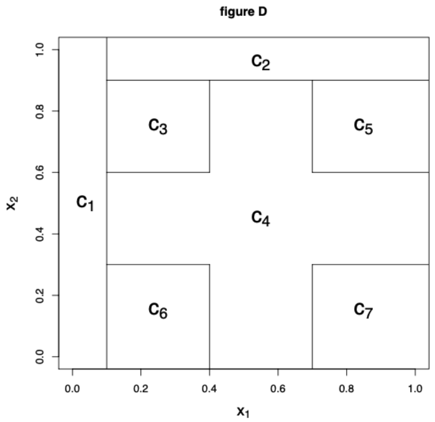
Problem 1

Consider the following figures:









(a) Which of the above divisions of the covariate space into regions R_m can have been produced by a regression tree? Explain why the one/ones you did not select cannot have been produced by a regression tree.

(b) For the divisions that could have been produced by a regression tree, draw the corresponding tree, including the formulas in the nodes (of the type $x_j \leq t_k$) and the values c_m in the leaves.

(c) Using the tree(s) selected in (b), provide an estimate for the following values:

i) $f(x_1 = 0.6, x_2 = 0.6)$

ii) $f(x_1 = 0.1, x_2 = 0.6)$

iii) $f(x_1 = 0.6, x_2 = 0.1)$

Problem 2

In this problem, we will consider a slightly modified version of the the data set `Wage` from the book ISLR (see the weekly exercises), where the aim is to predict the wage of a group of men from the Atlantic Region in the U.S., based on a set of covariates, using different models. The data are available in the file `wage.csv` under `https://www.uio.no/studier/emner/matnat/math/STK2100/v25/oblig/`, and can be be obtained using the following R commands:

```
wage <-
read.csv("https://www.uio.no/studier/emner/matnat/math/STK2100/v25/oblig/wage.csv",
header=TRUE)
```

The data set contains the following columns:

- `year`: the year when the wage was registered
- `age`: age
- `maritl`: marital status (Married/Unmarried)
- `race`: race (White/Black/Other)
- `education`: highest degree of finished education (¡ High scool/High school grad/Some college/College grad/Advanced Degree)
- `jobclass`: type of job (Industrial/Information)
- `health`: state of health ($\leq$ Good/$\geq$ Very good)
- `health_ins`: has health insurance (Yes/No)
- `wage`

Note that the two covariates `year` and `age` are quantitative, and the remaining ones are qualitative. R-code that is particularly relevant are the ones from the lectures on January 28 and March 25.

(a) (i) Start by randomly dividing the data set into a training and a test set. Make sure that both the training and the test set contain all the categories of qualitative covariates. Use the same training and test sets in all the subproblems of Problem 2.
(ii) Fit a linear regression model to the training set, and comment on the fit. Do all covariates seem to have an effect on the wage? Justify your answer.
(iii) Evaluate the prediction performance on the resulting model on the test set, using the mean squared error loss $\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ as the performance measure. This will be compared to the performance of models later in the problem.

(b) (i) Fit an additive model to the same data, using natural cublic splines for the two quantitative covariates `year` and `age`.
(ii) Why should you not model the qualitative covariates with splines?

(iii) Should both quantitative covariates be modelled with splines, or can the effect of one of them or both be just as well captured by a simple linear term? Justify your answer.
(iv) Do all the covariates seem to have an effect on the wage? Justify your answer. (iv) Evaluate your model of choice on the test data, and compare to the linear regression model from (a).

(c) (i) Fit a regression tree to the data, and plot it. Which covariates contribute to the model? Comment on potential differences from the models fitted in a) and b) with respect to this.
(ii) Prune the tree using 10-fold cross-validation. Plot the resulting tree, and comment on potential differences from the tree fitted in (i).
(iii) Evaluate your regression tree of choice on the test data, and compare the results to the ones you got in (a) and (b). Which model would you choose for predicting the wage? Justify your answer.

Problem 3

In this problem, we will consider a slightly modified version of the `Vertebral Column` dataset from (`https://archive.ics.uci.edu/dataset/212/vertebral+column`). It contains information about 310 orthopaedic patients, which belong to 2 classes healthy and unhealthy (these patients have either disk hernia or spondilolysthesis, but are put in the same class). Each patient is represented in the data set by 6 biomechanical attributes derived from the shape and orientation of the pelvis and lumbar spine, as well as the class

- `pelvInc`: pelvic incidence
- `pelvTilt`: pelvic tilt
- `lumbLord`: lumbar lordosis angle
- `SacrSl`: sacral slope
- `pelvRad`: pelvic radius
- `degrS`: grade of spondylolisthesis
- `class`: healthy=0 / unhealthy=1

The data are available in the file `vertebral-column.csv` under `https://www.uio.no/studier/emner/matnat/math/STK2100/v25/oblig/`, and can be be obtained using the following R commands:

```
vert <-
read.csv("https://www.uio.no/studier/emner/matnat/math/STK2100/v25/oblig/vertebral-column.cs
header=TRUE)
```

The aim is to predict whether a patient is healthy or not based on the 6 biomechanical attributes, using different classification methods. R-code that is particularly relevant are the ones from the lectures on February 25 and April 2.

(a) (i) Start by randomly dividing the data set into a training (2/3) and a test set (1/3), such that the fraction of healthy and unhealthy patients is approximately the same in the two sets. Use the same training and test sets in all the sub-problems of Problem 3.
(ii) Fit a logistic regression model to the training set, and comment on the fit. Do all covariates seem to have an effect on the risk of being unhealthy? Justify your answer.
(iii) Evaluate the prediction performance on the resulting model on the test set, using the mean misclassification error $\frac{1}{N}\sum_{i=1}^{N} I(y_i \neq \hat{y}_i)$ as performance measure. This will be compared to the performance of models later in the problem.

(b) (i) Fit a linear discriminant analysis (LDA) model to the same data.
(ii) Evaluate the model on the test set, and compare to the logistic regression model.

(c) (i) Fit a quadratic discriminant analysis (QDA) model to the same data.
(ii) Evaluate the model on the test set, and compare to the logistic regression and the LDA model.

(d) (i) Fit a one-layer neural network to the data, trying different sizes (number of hidden neurons) and decay parameters λ (remember to standardise the covariates).
(ii) Evaluate the fitted models on the test set, and compare the best performing one to the models fitted earlier in the problem.