Compulsory Exercise 2

TMA4268 Statistical Learning V2021

Emma Skarstein, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU

Hand out date: March 22, 2021

The submission deadline is Monday, April 19th 2021, 23:59h using Blackboard

Introduction

Maximal score is 50 points. Your score will make up 20% points of your final grade.

Supervision

Supervisions will be via Whereby by the teaching assistants and the lecturer during the usual lecture and exercise hours on April 12th and 13th. Ideally, only one member of a group comes with the questions that the group compiled. This ensures that all groups have the opportunity to ask questions. More information is as always available from the course website.

General supervision:

- Monday, April 12, 14.15-16.00
- Tuesday April 13, 10.15-12.00
- Tuesday April 13, 16.15-18.00

Remember that there is also the Piazza forum, and we strongly encourage you to use it for your questions - this ensures that all other students benefit from the answers.

Practical issues

- Group size is 2 or 3 – please use the same groups as for Compulsory 1.
- Remember to write your names and group number on top of your submission.
- The exercise should be handed in as *one R Markdown file and a pdf-compiled version* of the R Markdown file (if you are not able to produce a pdf-file directly please make an html-file, open it in your browser and save as pdf - no, not landscape - but portrait please). We will read the pdf-file and use the Rmd file in case we need to check details in your submission.
- In the R-chunks please use both `echo=TRUE` and `eval=TRUE` to make it simpler for us to read and grade.
- Please do not include all the text from this file (that you are reading now) - we want your R code, plots and written solutions - use the template from the course page (https://wiki.math.ntnu.no/tma4268/2021v/subpage6).
- Please **not more than 15 pages** in your pdf-file! (**This is a request**!)
- Please save us time and do NOT submit word or zip, and do not submit only the Rmd. This only results in extra work for us!

R packages

You probably need to install most of the following packages in R for this exercise:

```
install.packages("knitr")  #probably already installed
install.packages("rmarkdown")  #probably already installed
install.packages("ggplot2")  #plotting with ggplot
install.packages("ggfortify")
install.packages("leaps")
install.packages("glmnet")
install.packages("tree")
install.packages("caret")
install.packages("randomForest")
install.packages("readr")
install.packages("e1071")
install.packages("dplyr")
install.packages("gbm")
```

Multiple single choice problems

Some of the problems are *multiple choice questions*. This is how these will be graded:

There are four choices, and each of them can be TRUE or FALSE. If you make one mistake (either wrongly mark an option as TRUE/FALSE) you get 1P, if you have two or more mistakes, you get 0P. Your answer should be given as a list of answers, like TRUE, TRUE, FALSE, FALSE, for example.

Problem 1 (10P)

a) (2P) - Multiple choice

Which of the following statements are true, which false?

(i) In general, regularization will reduce test error, but not training error.

(ii) In best subset selection, we cannot use RSS as a criterion to choose between models with different numbers of predictors, only to select between models with the same number of predictors.

(iii) The tuning parameter λ in ridge and lasso regression should be chosen through cross-validation.

(iv) Principal component regression can be used for variable selection.

For the following tasks we will use a data set showing the number of birds killed by cats during a year. The number of killed birds is recorded along with 16 other variables. There were 452 cats participating in the study. The variables are

- `birds`: number of birds killed by the cat throughout a year
- `sex`: the sex of the cat
- `weight`: the cat's weight
- `dryfood`: daily amount of dry food (g)
- `wetfood`: daily amount of wet food (g)
- `age`: the cat's age
- `owner.income`: household yearly income (NOK)
- `daily.playtime`: daily time of owners spent playing with cat (minutes)

- `fellow.cats`: number of additional cats in household
- `owner.age`: average age of humans in household
- `house.area`: house area (sq. meters)
- `children.13`: number of children under the age of 13 in the household
- `urban`: whether the cat's home is in a urban location or not
- `bell`: does the cat wear a bell?
- `dogs`: number of household dogs
- `daily.outdoortime`: amount of time spent outside daily (minutes)
- `daily.catnip`: self-reported daily amount of catnip (g)
- `neutered`: whether or not the cat has been neutered/spayed

```
id <- "1iI6YaqgG0QJW5onZ_GTBsCvpKPExF30G"  # google file ID
catdat <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download", id),
    header = T)
```

We let `birds` be our response, and begin by splitting into training and testing sets (50% of the data in each) using the following code:

```
set.seed(4268)
train.ind = sample(1:nrow(catdat), 0.5 * nrow(catdat))
catdat.train = catdat[train.ind, ]
catdat.test = catdat[-train.ind, ]
```

b) (2P) - Best subset selection

Use best subset selection to identify a satisfactory model that uses a subset of the variables. Justify any choices you make. Report the selected variables and the test MSE.

c) (2P) - Lasso regression

Now use lasso regression on the same data set. Explain how you choose λ. Report the non-zero coefficients, and the test MSE.

R-hints:

```
x.train <- model.matrix(birds ~ ., data = catdat.train)[, -1]
y.train <- catdat.train$birds
x.test = model.matrix(birds ~ ., data = catdat.test)[, -1]
y.test = catdat.test$birds
```

d) (1P)

For the lasso regression, what happens when $\lambda \to \infty$? What happens when $\lambda = 0$?

e) (2P)

Now check whether the test MSE is actually better for the models in b) and c), compared to

i) a model with only intercept, and
ii) a multiple linear regression using all covariates.

f) (1P)

Present all the MSE values from the best subset selection, lasso regression, intercept-only and ordinary linear regression in a table. Explain what you see. Does it fit with what you would have expected?

Problem 2 (6P)

a) (2P) - Multiple choice

Which of the following statements are true, which false?

(i) A *natural cubic spline* has fewer degrees of freedom than a *cubic spline*.
(ii) Smoothing splines attempt to reduce the variance of a fit by penalizing the second-order derivative of the fit.
(iii) A regression spline with polynomials of degree $M-1$ has continous derivatives up to order $M-2$, but not at the knots.
(iv) A regression spline of order 3 with 4 knots has 8 basis functions (not counting the intercept).

b) (2P)

Write down the basis functions for a cubic spline with knots at the quartiles q_1, q_2 of variable X.

c) (2P)

When you look at the plot of `birds` against `daily.outdoortime` in the cat-dataset we used in problem 1, it may look like there is a slight non-linearity in the relationship.

(i) Fit polynomial regression model for `birds` with `daily.outdoortime` as the only covariate for a range of polynomial degrees ($d = 1, \ldots, 10$) and plot the results. Use the training data for this task.

(ii) Report the *training* MSE for (i), and explain what is happening.

Problem 3 (10P)

a) (2P) - Multiple choice

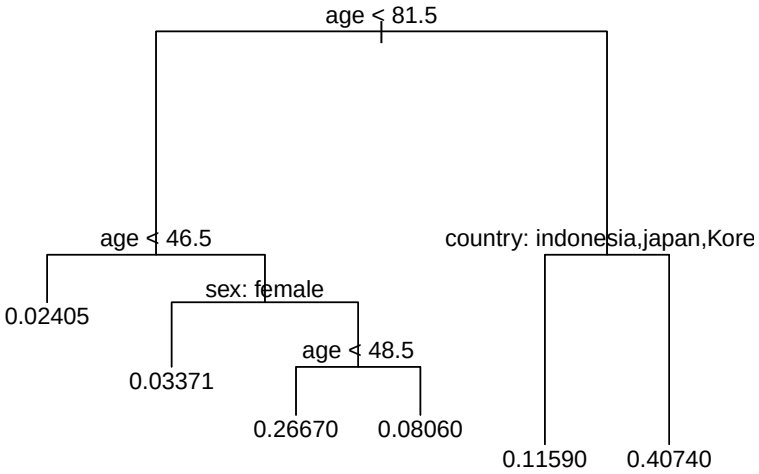
We are using again the Covid-19 dataset that we analyzed in compulsory exercise 1. Remember that the outcome was a binary variable that indicated if a patient died (`deceased=1`) or not (`deceased=0`). This time we are building a regression tree where we use the response as a numeric value (why?), and then apply cost-complexity pruning (figures and code below). Which of the following statements are true, which false?

(i) The probability of dying (`deceased = 1`) is about 40.7% for a French person with age above 81.5.

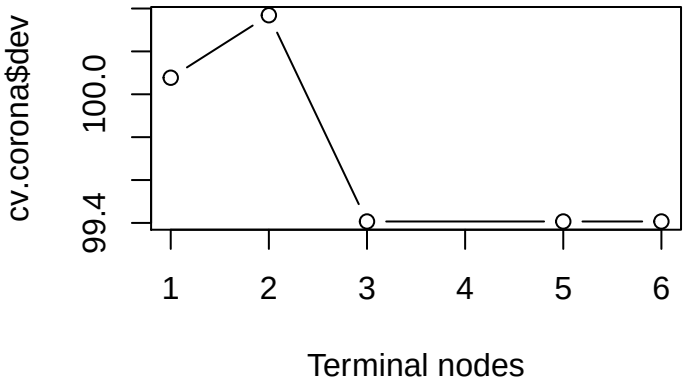
(ii) Age seems to be a more important predictor for mortality than sex.

(iii) Cost-complexity pruning was done using 10-fold CV.

(iv) It looks like the tree with 6 terminal nodes was the best choice, so there is no need to prune the original tree.




```
set.seed(1)
cv.corona = cv.tree(t.corona, K = 5)
plot(cv.corona$size, cv.corona$dev, type = "b", xlab = "Terminal nodes")
```



b) (2P)

Imagine that you decide to prune the tree from a) down to three leaves. From looking at the tree above, which three leaves are they (1P) and why (1P)? You do not need to carry out an analysis, only look at the tree above and argue. You may choose to draw a figure, or to only describe your tree with words.

c) (6P)

We will use the classical data set of *diabetes* from a population of women of Pima Indian heritage in the US, available in the R `MASS` package. The following information is available for each woman:

- diabetes: `0`= not present, `1`= present
- npreg: number of pregnancies
- glu: plasma glucose concentration in an oral glucose tolerance test
- bp: diastolic blood pressure (mmHg)
- skin: triceps skin fold thickness (mm)
- bmi: body mass index (weight in kg/(height in m)^2)
- ped: diabetes pedigree function.
- age: age in years

We will use a training set (called `d.train`) with 300 observations (200 non-diabetes and 100 diabetes cases) and a test set (called `d.test`) with 232 observations (155 non-diabetes and 77 diabetes cases). Our aim is to make a classification rule for the presence of diabetes (yes/no) based on the available data. You can load the data as follows:

```
id <- "1Fv6xwKLSZHldRAC1MrcK2mzdOYnbgv0E"  # google file ID
d.diabetes <- dget(sprintf("https://docs.google.com/uc?id=%s&export=download", id))
d.train = d.diabetes$ctrain
d.test = d.diabetes$ctest
```

We are interested in a tree-based method in order to build a model to predict the presence for `diabetes`. In addition, we want to understand which factors are most relevant in predicting diabetes.

(i) (3P) Start by using a simple classification tree. Apply cost-complexity pruning using 10-fold CV on the training data set. Report the misclassification error on the test set.

(ii) (3P) Now construct a classification tree based on a more advanced method. Train the model using the training data and report the misclassification error for the test data. Explain your choice of the (tuning) parameters. Which two variables are the most influential ones in the prediction of diabetes?

R-hint: Please use `set.seed(1)` before your run cross-validation in task (i), so that it is easier to reproduce your results.

Problem 4 (12P)

a) (2P) - Multiple choice

Imagine you have gene expression data for leukemia patients, with $p = 4387$ genes measured on blood samples for a total of $n = 92$ patients, of wich 42 have leukemia and 50 patients are healthy. Which statements are true?

(i) In this dataset we are guaranteed to find a separating hyperplane, unless there are exact feature ties (two patients with the exact same gene data, but different outcome).
(ii) In the analysis of this dataset we should prefer a soft-margin classifier over a separating hyperplane to omit overfitting.
(iii) Logistic regression is the preferred method for this data set, because it gives nice interpretable parameter estimates.
(iv) By choosing a large budget parameter C we are making the model more robust, but introduce more bias.

b) (7P)

We are looking at a (subset of) a dataset that contains gene expression data for 60 patients and more than 22'000 genes per patient. It is generated from the U133A platform and collected from The Children's Hospital at Westmead and was extracted and modified using the original publication by Anaissi et al (2016; https://doi.org/10.1371/journal.pone.0157330). Here we only use a subset of 10'000 genes for computational efficiency:

```
id <- "1x_E8xnmz9CMHh_tMwIsWP94czPa1Fpsj"  # google file ID
d.leukemia <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = T)
```

We are splitting the dataset into a training set and a test set of size 45 and 15, respectively:

```
set.seed(2399)
t.samples <- sample(1:60, 15, replace = F)
d.leukemia$Category <- as.factor(d.leukemia$Category)
d.leukemia.test <- d.leukemia[t.samples, ]
d.leukemia.train <- d.leukemia[-t.samples, ]
```

(i) (1P) Why is a support vector machine (SVM) more suitable here than a logistic regression approach? What other approach could be used instead of a SVM?

(ii) (1P) Say in 1-2 sentences what the paper where we have taken the data from (https://doi.org/10.1371/journal.pone.0157330) intends to do, that is, what kind of method do they suggest and what is the purpose?

(iii) (3P) Fit a support vector classifier (linear boundary) to find a function that predicts whether a child that has been successfully treated against leukemia later relapsed or not. Fix the tuning parameter to $C = 1$ (1P). Report the confusion tables and misclassification error rates for both the training and test sets (0.5P each).

- (0.5P) Is it surprising to see the training error rate? Why?
- (0.5P) Which is the most common type of error that you see in the test set? Do you think the classification method is successful?

(iv) (2P) Repeat the analysis with a radial kernel using $C = 1$. Do the analysis twice, one with $\gamma = 10^{-2}$ and once with $\gamma = 10^{-5}$. Interpret the training and test error rates that you find now, and compare to the results in (iii).

R-hints: To run cross-validation over a grid of two tuning parameters, we would usually use the `tune()` function. However this is not an efficient way of studying different tuning parameters for this large dataset.

c) (3P)

The SVM is an extension of the support vector classifier, by enlarging the feature space using kernels. The polynomial kernel is a popular choice and has the following form

$$K(\boldsymbol{x}_i, \boldsymbol{x}_i') = (1 + \sum_{j=1}^{p} x_{ij} x_{i'j'})^d$$

Show that for a feature space with inputs X_1 and X_2 and for degree $d = 2$, the above kernel can be represented as the inner product

$$K(\boldsymbol{x}_i, \boldsymbol{x}_i') = \langle h(x), h(x') \rangle \ ,$$

where $h(x)$ is a 6-dimensional transformation function in an enlarged space. Explicitly derive $h(x) = (h_1(x), \ldots, h_6(x))$.

Problem 5 (12P)

a) (2P) - Multiple choice

Which of the following statements are true, and which are false?

(i) In principal component analysis, the second principal component is the linear combination of the predictors that has the largest variance of all the linear combinations that are uncorrelated with the first principal component.

(ii) It makes no difference for the results of PCA if the variables are standardized beforehand or not.

(iii) K-means clustering is robust against differing initial choices of cluster assignments.

(iv) PCA is most helpful when all the variables are uncorrelated.

b) (3P)

In this problem, we use a simple data set of six observations to preform K-means clustering manually, as specified in algorithm 10.1 on page 388 of the ISLR book. The observations are given below. In this problem we will use $K = 2$.

```
x1 <- c(1, 2, 0, 4, 5, 6)
x2 <- c(5, 4, 3, 1, 1, 2)
```

(i) (1P) Randomly assign a cluster to each of the observations, and visualize the resulting observations with the color indicating the assigned cluster.
R-hint: You can use the function `sample()` for this. Please use `set.seed(1)` before you sample the random clusters, so it is easier to reproduce.

(ii) (1P) Manually calculate the centroids for each cluster (that is, you are encouraged to do the calculations in R, but don't use any non-trivial functions that you haven't written yourself). Plot the centroids on the plot from (i).

(iii) (1P) Now assign each observation to the centroid which it is closest to, in Euclidean distance. Display the new cluster assignments as in (i).

The following dataset consists of 40 tissue samples with measurements of 1,000 genes. The first 20 tissues come from healthy patients and the remaining 20 come from a diseased patient group. The following code loads the dataset into your session with colum names decribing if the tissue comes from a diseased or healthy person.

```
id <- "1VfVCQvWt121UN39NXZ4aR9Dmsbj-p9OU"  # google file ID
GeneData <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = F)
colnames(GeneData)[1:20] = paste(rep("H", 20), c(1:20), sep = "")
colnames(GeneData)[21:40] = paste(rep("D", 20), c(1:20), sep = "")
row.names(GeneData) = paste(rep("G", 1000), c(1:1000), sep = "")
GeneData = t(GeneData)
GeneData <- scale(GeneData)
```

c) (2P)

Perform hierarchical clustering with complete, single and average linkage using **both** Euclidean distance and correlation-based distance on the dataset. Plot the dendograms. Hint: You can use `par(mfrow=c(1,3))` to plot all three dendograms on one line or `par(mfrow=c(2,3))` to plot all six together.

d) (2P)

Use these dendograms to cluster the tissues into two groups. Compare the groups with respect to the patient group the tissue comes from. Which linkage and distance measure performs best when we know the true state of the tissue?

e) (2P)

(i) (1P) Use PCA to plot the samples in two dimensions. Color the samples based on the tissues group of patients.

(ii) (1P) How much variance is explained by the first 5 PCs?

f) (1P)

Use your results from PCA to find which genes that vary the most accross the two groups.