Compulsory Exercise 1

TMA4268 Statistical Learning V2020

Martina Hall, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU

Hand out date: February 3, 2020

Last changes: 02.02.2020

The submission deadline is Thursday, February 20th 2020, 12:00h using Blackboard

Introduction

Maximal score is 50 points. You need a score of 20/50 for the exercise to be approved. Your score will make up 10% points of your final grade.

Supervision

Supervisions will be in the usual lecture rooms (S1, S4) and Smia

- Monday, February 3, during interactive lecture: Introduction to R Markdown and the template you should use for this exercise.

General supervision:

- Friday, February 7, 14.15-16.00 (S4)
- Monday February 10, 10.15-12.00 (S1)
- Thursday February 13, 8.15-10.00 (Smia)

Practical issues

- Maximal group size is 3 - join a group (self enroll) before handing in on Bb.
- If you did not find a group, you can email Stefanie (stefanie.muff@ntnu.no) and I will try to match you with others that are alone.
- 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/tma4268/2020v/subpage6).
- Please not more than 12 pages in your pdf-file! (This is a request, not a requirement.)
- Please save us time and NOT submit word or zip, and do not submit only the Rmd. This only results in extra work for us!

R packages

You need to install the following packages in R to run the code in this file.

```
install.packages("knitr")  #probably already installed
install.packages("rmarkdown")  #probably already installed
install.packages("ggplot2")  #plotting with ggplot
install.packages("ggfortify")
install.packages("MASS")
install.packages("dplyr")
```

Multiple/single choice problems

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

- **Multiple choice questions (2P)**: 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.
- **Single choice questions (1P)**: There are four or five choices, and only *one* of the alternatives is the correct one. You will receive 1P if you choose the correct alternative and 0P if you choose wrong. Only say which option is true (for example (ii)).

Problem 1 - 10P

We have a univariate continuous random variable Y and a covariate x. Further, we have observed a training set of independent observation pairs $\{x_i, y_i\}$ for $i = 1, \ldots, n$. Assume a regression model

$$Y_i = f(x_i) + \varepsilon_i \ ,$$

where f is the true regression function, and ε_i is an unobserved random variable with mean zero and constant variance σ^2 (not dependent on the covariate). Using the training set we can find an estimate of the regression function f, and we denote this by $\hat{f}$. We want to use $\hat{f}$ to make a prediction for a new observation (not dependent on the observations in the training set) at a covariate value x_0. The predicted response value is then $\hat{f}(x_0)$. We are interested in the error associated with this prediction.

a) (1P)

Write down the definition of the expected test mean squared error (MSE) at x_0.

b) (2P)

Derive the decomposition of the expected test MSE into three terms. (Todo: Think about how to split the points)

c) (1P)

Explain with words how we can interpret the three terms.

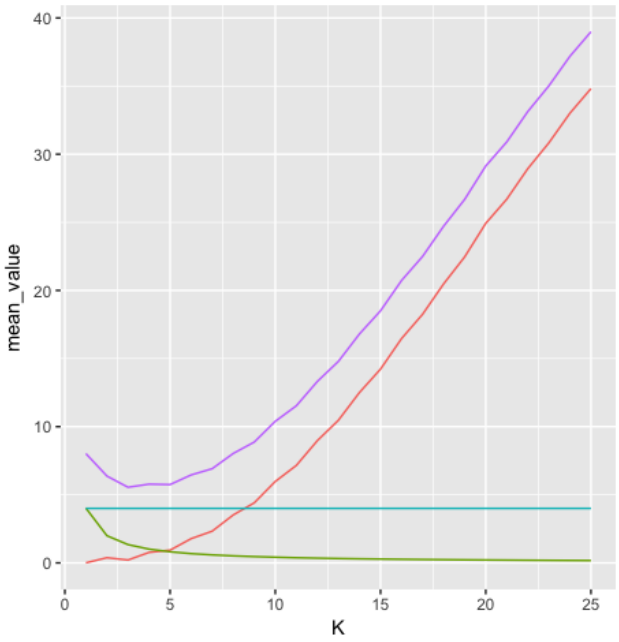

Figure 1: Squared bias, variance, irreducible error and total error for increasing values of K in KNN

d) (2P) - Multiple choice

Which of the following statements are true and which are false? Say for *each* of them if it is true or false.

(i) If the relationship between the predictors and response is highly non-linear, a flexible method will generally perform **better** than an inflexible method.
(ii) If the sample size n is extremely large and the number of predictors p is small, a flexible method will generally perform **worse** than an inflexible method. method.
(iii) If the number of predictors p is extremely large and the number of observations n is small, a flexible method will generally perform **better** than an inflexible method.
(iv) If the variance of the error terms, $\sigma^2 = Var(\epsilon)$ is extremely high, a flexible method will generally perform **worse** than an inflexible method.

List of answers:

e) (2P) - Multiple choice

Figure 1 shows the squared bias, variance, irreducible error and total error for increasing values of K in KNN. Which of the following statements are true and which are false? Say for *each* of them if it is true or false.

(i) The blue line corresponds to the irriducible error.
(ii) Increased K corresponds to increased flexibility of the model.
(iii) The squared bias increases with increased value of K.
(iv) The variance increases with increased value of K.

List of answers:

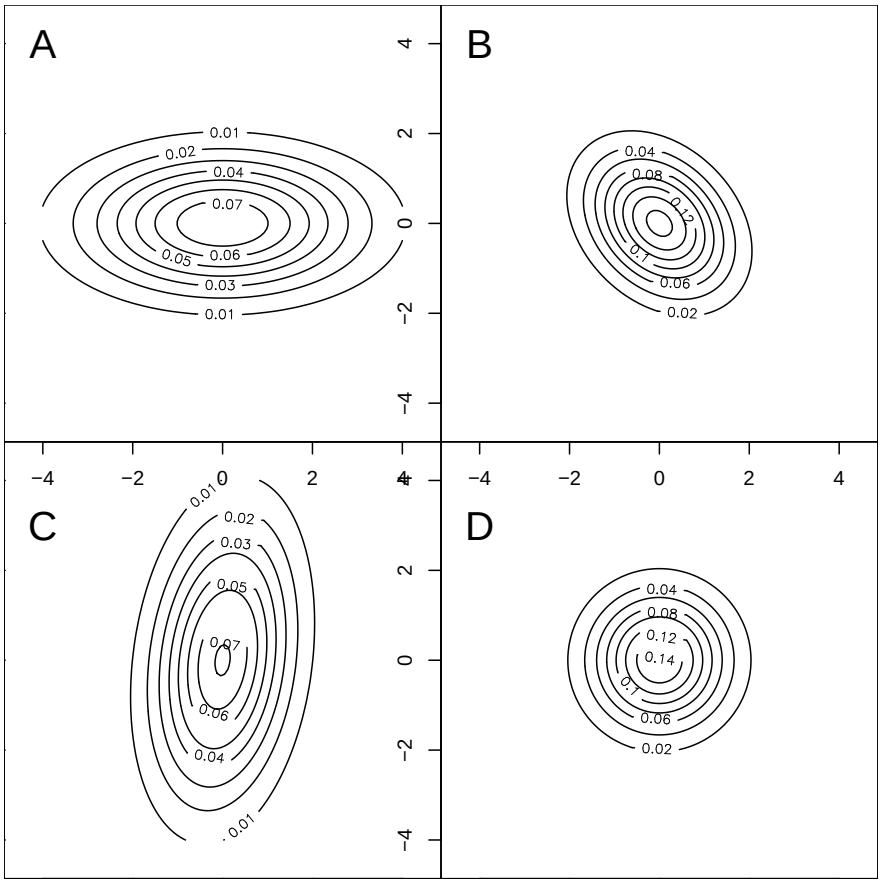

Figure 2: Contour plots

f) (1P) - Single choice

X is a 2-dimensional random vector with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 3 & 0.6 \\ 0.6 & 4 \end{bmatrix}$$

The correlation between the two elements of **X** is: (i) 0.05 (ii) 0.17 (iii) 0.03 (iv) 0.60 (v) 0.10

Answer:

g) (1P) - Single choice

Which of the plots (A-D) in Figure 2 corresponds to the following covariance matrix?

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.2 \\ 0.2 & 4 \end{bmatrix}$$

Answer:

Problem 2 - 15P

We are looking at a linear regression problem. Biologists studied badgers in Switzerland, and these animals mainly eat earthworms. In their exctrements one can find an undigestable part of the earthworm (the muscular stomach), and in order to find out how much energy a badger took in by eating earthworms, the biologists wanted to figure out how the circumference of the muscular stomach can be used to predict to the weight of the earthworm that the badger ate. So they went to collect 143 earthworm, which they weighted and measured.

The earthworm dataset can be loaded as follows:

```
id <- "1nLen1ckdnX4P9n8ShZeU7zbXpLc7qiwt"  # google file ID
d.worm <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id))
```

Look at the data by using both `head(d.worm)` and `str(d.worm)`. The dataset contains the following variables:

- `Gattung`: The species of the worm (L=Lumbricus; N=Nicodrilus; Oc=Octolasion)
- `Nummer`: A worm-specific ID
- `GEWICHT`: The weight of the earthworm
- `FANGDATUM`: The date when the data were collected
- `MAGENUMF`: The circumference of the muscular stomach

a) (2P)

What is the dimension of the dataset (number of rows and columns)? Which of the variables are qualitative, which are quantitative?

b) (2P)

An important step before fitting a model is to look at the data to understand if the modelling assumptions are reasonable. In a linear regression setup, it is for example recommended to look at the relation of the variables to see if the linearity assumption makes sense. If this is not the case, you can try to transform the variables.

Make a scatterplot of `GEWICHT` against `MAGENUMF`, where you color the points according to the three species (variable `Gattung`).

Does this relationship look linear? If not, try out some transformations of `GEWICHT` and `MAGENUMF` until you are happy.

R-hint:

```
ggplot(d.worm, aes(x = ..., y = ..., colour = ...)) + geom_point() + theme_bw()
```

c) (3P)

Fit a regression model that predicts the weight (`GEWICHT`) given the circumference of the stomach (`MAGENUMF`) and the species (`Gattung`). **Use the transformed version of the variable(s) from b)**. Use only linear terms that you combine with `+` (no interactions) (1P). After you fitted the models, write down the separate regression models with the estimated parameters for the three species as three separate equations (1P). Is `Gattung` a relevant predictor? (1P)

R-hints : `lm()`, `summary()`, `anova()`

d) (2P)

In question c) it was assumed that there is no interaction between the species and `MAGENUMF` to predict the weight of a worm. Test whether an interaction term would be relevant by fitting an appropriate model.

e) (2P)

Carry out a residual analysis using the `autoplot()` function from the `ggfortify` package.

- Do you think the assumptions are fulfilled? Explain why or why not.
- Compare to the residual plot that you would obtain when you would not use any variable transformations to fit the regression model.

f) (2P)

- Why is it important that we carry out a residual analysis, i.e., what happens if our assumptions are not fulfilled?
- Mention at least one thing that you could do with your data / model in case of violated assumptions.

g) (2P) - Multiple choice

Given a null hypothesis (H_0), an alternative hypothesis (H_1) and an observed result with an associated p-value. Think of the example where H_0 is that a slope parameter in a regression model is $\beta = 0$. Which of the following statements is correct?

(i) The p-value is the probability that H_0 is true.
(ii) The p-value is the probability that H_1 is true.
(iii) $(1 - p)$ is the probability that H_1 is true.
(iv) p is the probability to observe a data symmary under the null hypothesis (H_0) that is at least as extreme as the one observed.

List of answers:

Problem 3 - 16P

In this problem, we will use a dataset from the Wimbledon tennis tournament for Women in 2013. We will predict the result for player 1 (win=1 or loose=0) based on the number of aces won by each player and the number of unforced errors commited by both players. The data set is a subset of a data set from https://archive.ics.uci.edu/ml/datasets/Tennis+Major+Tournament+Match+Statistics, see that page for information of the source of the data.

The files can be read using the following code.

```
# read file
id <- "1GNbIhjdhuwPOBr0Qz82JMkdjUVBuSoZd"
tennis <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = T)
```

We will first create a logistic regression model where the probability to win for player 1 has the form

$$P(Y_i = 1|\mathbf{X} = \boldsymbol{x}_i) = p_i = \frac{e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}}}{1 + e^{\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+\beta_3 x_{i3}+\beta_4 x_{i4}}} ,$$

where x_{i1} is the number of aces for player 1 in match i, x_{i2} is the number of aces for player 2 in match i, and x_{i3} and x_{i4} are the number of unforced errors commited by player 1 and 2 in match i. $Y_i = 0$ represents player 1 winning match i, $Y_i = 0$ represents player 1 loosing match i.

a) (1P)

Use the above expression to show that $\text{logit}(p_i) = \log(\frac{p_1}{1-p_i})$ is a linear function of the covariates.

b) (1P)

The model above has been fitted and gives the following output. Interpret the effect of β_1, i.e. how will one more ace for player 1 affect the result of the tennis match?

```
##
## Call:
## glm(formula = Result ~ ACE.1 + ACE.2 + UFE.1 + UFE.2, family = "binomial",
##     data = tennis)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0517  -0.8454   0.3725   0.8773   2.0959
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -0.02438    0.59302  -0.041 0.967211
## ACE.1        0.36338    0.10136   3.585 0.000337 ***
## ACE.2       -0.22388    0.07369  -3.038 0.002381 **
## UFE.1       -0.09847    0.02840  -3.467 0.000527 ***
## UFE.2        0.09010    0.02479   3.635 0.000278 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 163.04  on 117  degrees of freedom
## Residual deviance: 124.96  on 113  degrees of freedom
## AIC: 134.96
##
## Number of Fisher Scoring iterations: 4
```

c) (4P)

We will now reduce the number of covariates in our model by looking at the difference between aces performed by player 1 and 2 and the difference in unforced errors made by the players. Use the following code to create these variables and divide the data into a train set and a test set.

```
# make variables for difference
tennis$ACEdiff = tennis$ACE.1 - tennis$ACE.2
tennis$UFEdiff = tennis$UFE.1 - tennis$UFE.2

# divide into test and train set
n = dim(tennis)[1]
```

```
n2 = n/2
set.seed(1234)  # to reproduce the same test and train sets each time you run the code
train = sample(c(1:n), replace = F)[1:n2]
tennisTest = tennis[-train, ]
tennisTrain = tennis[train, ]
```

- Use these variables to fit a logistic regression model on the form `Result` $\sim$ `ACEdiff` + `UFEdiff` on your training set.
- Using a 0.5 cutoff as decison rule, we classify an observation with covariates $\boldsymbol{x}$ as "Player 1 wins" if $\hat{P}(Y = 1|\boldsymbol{x}) > 0.5$. Write down the formula for the class boundary between the classes (results) using this decision rule. The boundary should be of the form $x_2 = bx_1 + a$.
- Make a plot with the training observations and and then add a line that represents the class boundary. Hint: in `ggplot` points are added with `geom_point` and a line with `geom_abline(slope=b, intercept=a)`, where a and b comes from your class boundary.
- Use your fitted model to predict the result of the test set. Make a confusion table using a 0.5 cutoff and calculate the sensitivity and specificity.

d) (1P)

Next, we will use LDA and QDA to classify the result using the same covariates (`ACEdiff` and `UFEdiff`) from the tennis data. In linear discriminant analysis with K classes, we assign a class to a new observation based on the posterior probability

$$P(Y = k|\mathbf{X} = \boldsymbol{x}) = \frac{\pi_k f_k(\boldsymbol{x})}{\sum_{l=1}^{K} \pi_l f_l(\boldsymbol{x})},$$

where

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_k})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_k})}.$$

- Explain what is π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and $f_k(x)$ in our `tennis` problem with the two covariates.

e) (3P)

In a two class problem ($K = 2$) the decision boundary for LDA between class 0 and class 1 is where x satisfies

$$P(Y = 0|\mathbf{X} = \boldsymbol{x}) = P(Y = 1|\mathbf{X} = \boldsymbol{x}).$$

- (1P) Show that we can express this as

$$\delta_0(\boldsymbol{x}) = \delta_1(\boldsymbol{x}), \tag{1}$$

  where

$$\delta_k(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k - \frac{1}{2} \boldsymbol{\mu}_k^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_k + \log \pi_k; \quad k \in \{0, 1\}. \tag{2}$$

- (1P) We use the rule to classify to class 1 for an observation with covariates $\boldsymbol{x}$ if $\hat{P}(Y = 1 \mid \boldsymbol{x}) > 0.5$. Write down the formula for the class boundary between the classes. Hint: formulate it as $ax_1 + bx_2 + c = 0$ and solve for x_2. Use R for the calculations.
- (1P) Make a plot with the training observations and the class boundary. Add the test observations to the plot (different markings). Hint: in `ggplot` points are added with `geom_points` and a line with `geom_abline(slope=b, intercept=a)` where a and b comes from your class boundary.

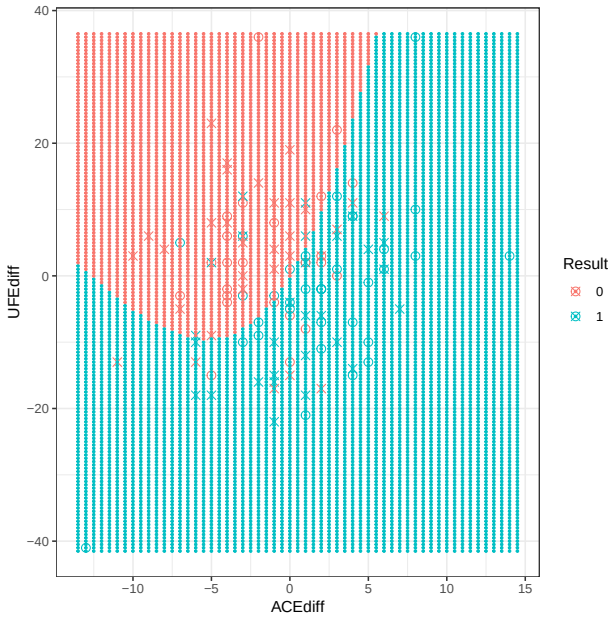

Figure 3: QDA decision boundary

f) (3P)

- (1P) Perform LDA on the training data (using the `lda()` function in R).
- (1P) Use your model to classify the results of the test set. Make the confusion table for the test set when using 0.5 as cut-off.
- (1P) Calculate the sensitivity and specificity on the test set.

g) (2P)

- Perform QDA on the training set. What is the difference between LDA and QDA?
- Make the confusion table for the test set when using 0.5 as cut-off. Calculate the sensitivity and specificity on the test set.

h) (1P)

Figure 3 shows the decision boundary for QDA, where observations falling into the red area will be classified as 0 (loose), and observations in the blue area will be classified as 1 (win). Circles represents observations from the train set, while crosses represents observations from the test set.

- Compare this plot with your corresponding plots for glm and LDA. Would you prefer glm, LDA or QDA for these data? Justify your answer this based on the results from the confusion matrices and in light of the decision boundaries meaning for your tennis-covariates.

Problem 4 (9P)

a) (2P)

Remember the formula for the K-nearest neighbour regression curve to estimate at a covariate value x_0,

$$\hat{f}(x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} y_i \ ,$$

where for a given integer K, the KNN regression first identifies the K points in the training data that are closest (Euclidean distance) to x_0, represented by $\mathcal{N}_0$. It then estimates the regression curve at x_0 as the average of the response values for the training observations in $\mathcal{N}_0$.

Given the set of possible values for K in the KNN regression problem specified in a), explain how 10-fold cross validation is performed, and specify which error measure you would use. Your answer should include a formula to specify how the validation error is calculated.

b) (2P) - Multiple choice

Which statements about validation set appraoch, k-fold cross-validation (CV) and leave-one-out cross validation (LOOCV) are true and which are false? Say for *each* of them if it is true or false.

(i) 5-fold CV will generally lead to more bias, but less variance than LOOCV in the estimated prediction error.
(ii) The validation set-approach is computationally cheaper than 10-fold CV.
(iii) The validation set-approach is the same as 2-fold CV.
(iv) LOOCV is always the cheapest way to do cross-validation.

List of answers:

c) (1P)

We are now looking at a bootstrap example. Assume you want to fit a model that predicts the probability for coronary heart disease (`chd`) from systolic blood pressure (`sbp`) and sex (0=female, 1=male). Load the data in R as follows

```
id <- "1I6dk1fA4ujBjZPo3Xj8pIfnzIa94WKcy"  # google file ID
d.chd <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id))
```

and perform a logistic regression with `chd` as outcome and `sbp` and `sex` as covariates. What is the probability of chd for a male with a sbp=140 in the given dataset?

d) (4P)

We now use the bootstrap to estimate the uncertainty of the probability derived in b). Proceed as follows:

- Use $B = 1000$ bootstrap samples.
- In each iteration, derive and store the estimated probability for `chd`, given `sbp`=140 and `sex`=male.
- From the set of estimated probabilities, derive the standard error.
- Also derive the 95% confidence interval for the estimates, using the bootstrap samples.
- Interpret what you see.