Compulsory Exercise 1

TMA4268 Statistical Learning V2025

Sara Martino, Stefanie Muff, Kenneth Aase, Department of Mathematical Sciences, NTNU

Hand out date: February 7, 2025

The submission deadline is Monday, 24. February 2025, 23:59h, using Blackboard.

Introduction

Maximal score is 50 points. You need a score of 30/50 for the exercise to be approved.

Supervision

Supervision will happen during the usual lectures and exercise session:

- Thursday, February 13, 08:15-10:00 (GL-GE G1)
- Friday February 14, 08:15-10:00 (GL-GE G1)
- Wednesday February 19, 16:15-18:00 (GL-GE G1)

Practical issues

- Maximum group size is 3 - join a group (self enroll) before handing in on Blackboard. We prefer that you do not work alone.
- Please organize yourself via the Mattelab discussion forum to find a group. Once you have formed a group, log into Blackboard and add yourself to the same group there.
- If you did not find a group, you can email Stefanie (stefanie.muff@ntnu.no) and we will try to match you with others that are alone (please use this really only if you have already tried to find a group).
- 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.
- Please check your compiled pdf to make sure it is readable. Resize figures if necessary.
- 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.
- Please **not more than 12 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 need to install the following packages in R to run the code in this file.

```
install.packages("knitr")
install.packages("rmarkdown")
install.packages("ggplot2")
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 estimate by $\hat{f}$. We want to use $\hat{f}$ to make a prediction for an independent new observation (not included 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.

c) (1P)

Explain with words how we can interpret the three terms.

d) (2P) - Multiple choice

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

(i) The bias-variance tradeoff is more relevant in inference than in prediction.
(ii) As the sample size n increases, the expected test MSE will approach zero.
(iii) Given two methods for estimating f, we get the best predictions from the one with the lowest squared bias.
(iv) If σ^2 is very large, we need a very flexible method to estimate f reliably.

e) (2P) - Multiple choice

Figure 1 shows an example of squared bias, variance, irreducible error and total error in a validation set for increasing values of K in a K nearest neighbor (KNN) regression prediction model. Which of the following statements are true and which are false? Say for **each** of them if it is true or false.

(i) As K decreases, the flexibility of the model increases.
(ii) The squared bias always contributes the most to the total error.
(iii) We have enough information to decide on a value of K to use in the KNN model.
(iv) For the plotted range of values of K, overfitting would be preferable to underfitting in this scenario, if we had to choose.

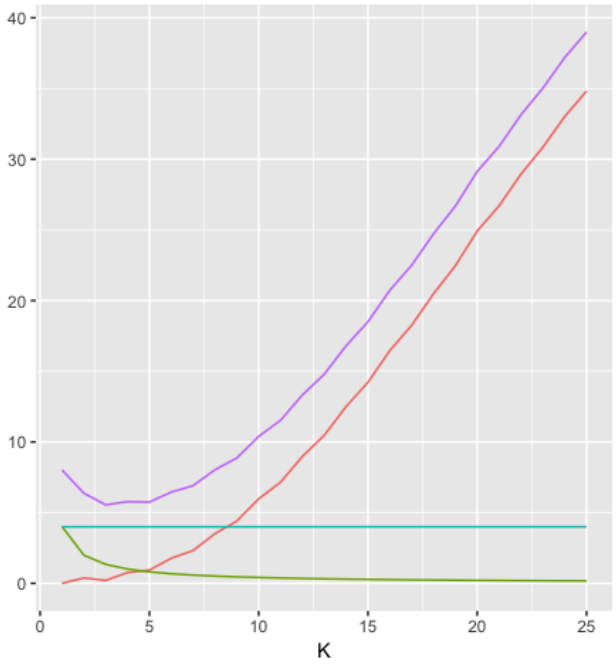

Figure 1: Squared bias (red), variance (green), irreducible error (light blue) and total error (lila) for increasing values of K in KNN.

f) (1P) - Single choice

$\mathbf{X}$ is a 2-dimensional random vector with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 9 & 0.3 \\ 0.3 & 4 \end{bmatrix}$$

The correlation between the two elements of $\mathbf{X}$ is: (i) 0.05 (ii) 0.15 (iii) 0.0083 (iv) 0.60 (v) 0.10

g) (1P) - Single choice

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

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

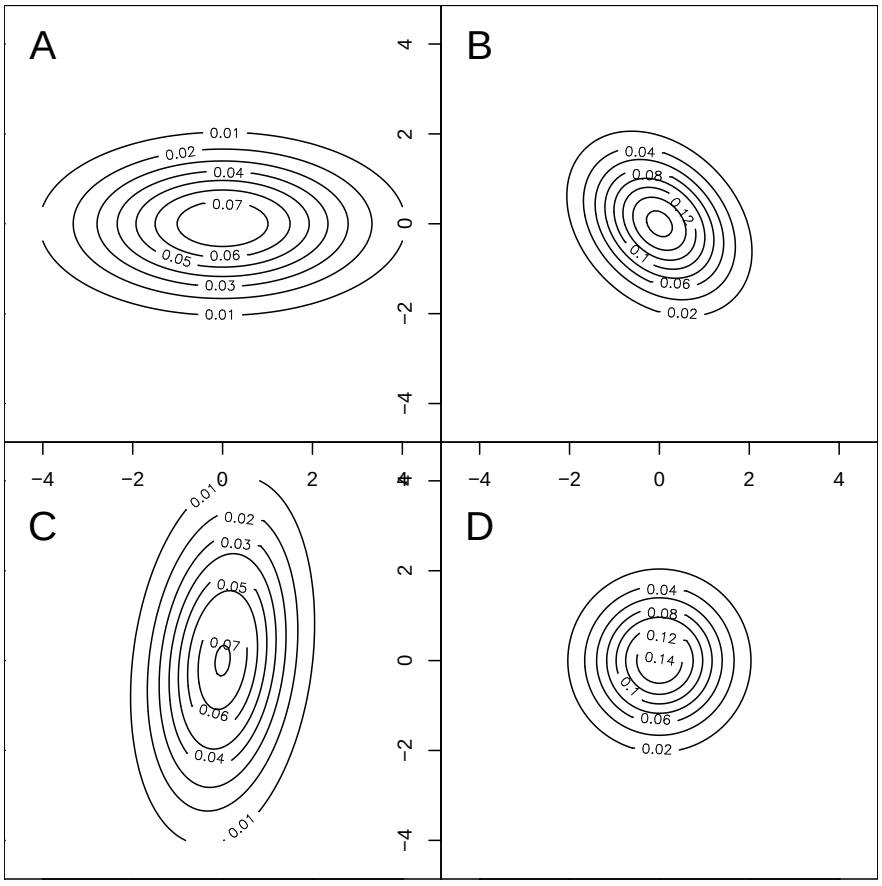

Figure 2: Contour plots

Problem 2 - 15P

We consider a linear regression problem. A group of biologists in Switzerland studied badgers, which mainly eat earthworms. In the badger's excrement one can find a non-digestible part of the earthworm (the muscular stomach). To find out how much energy a badger absorbed by eating earthworms, the biologists wanted to investigate the relationship between the circumference of the muscular stomach and the weight of the earthworm that the badger ate. Therefore, they collected a sample of earthworms, and for each worm the they measured its weight and the circumference of its muscular stomach.

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 genus 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 an exploratory 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 between the variables to see if the linearity assumption makes sense. If this is not the case, one can try to transform the variables.

Make a scatterplot of `GEWICHT` (weight) against `MAGENUMF` (circumference of stomach), 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:

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

c) (3P)

Fit a regression model for an earthworm's weight (`GEWICHT`) given the circumference of its muscular stomach (`MAGENUMF`) and the genus (`Gattung`). **Use the transformed version of the variable(s) from b)**. Use only linear terms that you combine with `+` (no interactions) (1P). After fitting the models, write down the model equations with the estimated parameters for the three genus as three separate equations (1P). Do we find evidence that `Gattung` impacts the weight? (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)

Perform a residual analysis using the `autoplot()` function from the `ggfortify` package. Use the model without interaction term.

- 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, for example, of the case where H_0 is that a slope parameter in a regression model is $\beta = 0$), which of the following statements are true and which are false? Say for **each** of them if it is true or false.

(i) The $1 - p$ is the probability that H_0 is true.
(ii) If the p-value is higher than 0.05, then H_1 is not true.
(iii) p is the probability to observe a data summary under the null hypothesis (H_0) that is at least as extreme as the one observed.
(iv) The p-value tells you the probability that the results happened by random chance.

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 (coded as win=1 or lose=0), based on the number of aces won by each player and the number of unforced errors committed by each player. 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 = TRUE)
```

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 committed by player 1 and 2 in match i. Remember: $Y_i = 1$ represents player 1 winning match i, $Y_i = 0$ represents player 1 loosing match i.

a) (1P)

Use the above expression to show that $\text{logit}(p_i) = \log(\frac{p_i}{1-p_i})$ 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?

```
r.tennis <- glm(Result ~ ACE.1 + ACE.2 + UFE.1 + UFE.2,
                data = tennis,
                family = "binomial")
summary(r.tennis)
```

```
##
## Call:
## glm(formula = Result ~ ACE.1 + ACE.2 + UFE.1 + UFE.2, family = "binomial",
##     data = tennis)
##
## 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 training 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 <- nrow(tennis)
set.seed(1234) # to reproduce the same test and train sets each time you run the code
train <- sort(sample(n, size = n / 2, replace = FALSE))
tennisTest <- tennis[-train, ]
tennisTrain <- tennis[train, ]
```

- Use these variables to fit a logistic regression model for `Result` with two covariates `ACEdiff` and `UFEdiff`, on your training set.
- Using a 0.5 cutoff as decision 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, 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 results for the data in 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

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

- Explain what π_k, $\boldsymbol{\mu}_k$, $\boldsymbol{\Sigma}$ and $f_k(x)$ are in a problem with two covariates (no calculations, only explanations).

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

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

  where

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

- (1P) We use the rule to classify an observation with covariates $\boldsymbol{x}$ to class 1 if $\hat{P}(Y = 1 \mid \boldsymbol{x}) > 0.5$. Write down the formula for the class boundary. 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.

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 (lose), 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 logistic regression and LDA. Would you prefer logistic regression, 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.

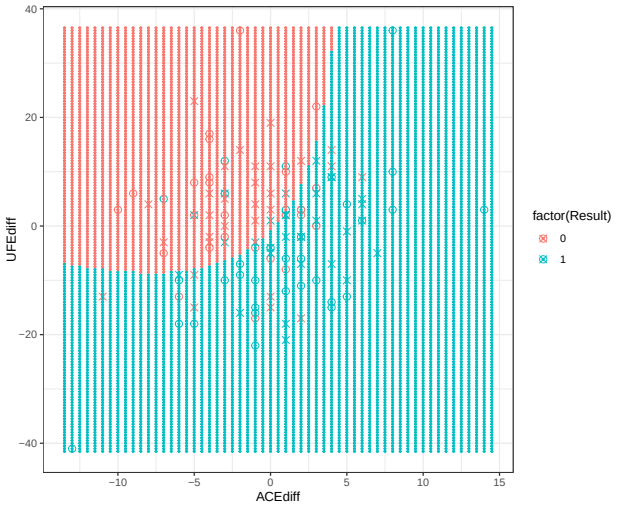

Figure 3: QDA decision boundary

Problem 4 (9P)

a) (2P)

Recall the formula for the K-nearest neighbor regression curve to predict 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 (in Euclidean distance) to x_0, represented by the set $\mathcal{N}_0$. 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 approach, 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 an estimate of the prediction error with less bias, but more variance, than LOOCV.
(ii) 10-fold CV is computationally cheaper than LOOCV.
(iii) The validation set-approach is the same as 2-fold CV.

(iv) LOOCV is a form of bootstrapping.

c) (1P)

We now consider an example of bootstrapping. Assume you want to fit a model that predicts the probability for coronary heart disease (`chd`) from systolic blood pressure (`sbp`) and sex (female coded as 0, male coded as 1). 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 estimated probability of coronary heart disease for a male with a systolic blood pressure of 140?

d) (4P)

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

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