Compulsory Exercise 1

TMA4268 Statistical Learning V2022

Emma Skarstein, Daesoo Lee, Stefanie Muff, Department of Mathematical Sciences, NTNU

Hand out date: February 7, 2022

The submission deadline is: Monday February 21, 23:59h using Blackboard

Introduction

Maximal score is 40 points. Your score will make up 10% points of your final grade.

Supervision

We will use the times where we would have lectures and exercises for supervision (4×2 hours). We will announce the exact implementation with physical and online supervision as soon as we know.

Supervision hours:

- Monday, February 14, 08:15-10:00 and 14.15-16.00
- Wednesday, February 16, 14.15-16.00
- Thursday February 17, 08.15-10.00

Remember that there is also the Mattelab forum, and we strongly encourage you to use it for your questions outside the supervision hours – this ensures that all other students benefit from the answers (try to avoid emailing the course staff).

Practical issues (Please read carefully)

- Group size is 2 or 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 (https://mattelab2022v.math.ntnu.no/c/tma4268/9) to find a group. Once you formed a group, log into Blackboard and add yourself to the same group there.
- If you did not find a group even when using Mattelab, you can email Stefanie (stefanie.muff@ntnu.no) and I 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 file!
- 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.
- You may want to work through the R Mardown bonus part in the R course (https://digit.ntnu.no/courses/course-v1:NTNU+IMF001+2020/about)

- 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/tma4268/2022v/subpage6).
- Please **not more than 12 pages** in your pdf-file! (This is a request, not a requirement.)
- 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. It is of course also possible to use more or different packages.

```
install.packages("knitr")  #probably already installed
install.packages("rmarkdown")  #probably already installed
install.packages("ggplot2")  #plotting with ggplot
install.packages("palmerpenguins")
install.packages("ggfortify")  # For model checking
install.packages("MASS")
install.packages("class")
install.packages("pROC")
install.packages("plotROC")
install.packages("boot")
```

Multiple/single choice problems

There will be a few *multiple and 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 several 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 (8P)

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) (2P)

Derive the decomposition of the expected test MSE, $E[y_0 - \hat{f}(x_0)]^2$, into three terms (bias, variance, and irreducible error).

b) (1P)

Explain with words how we can interpret the three terms.

c) (2P) - Multiple choice

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

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

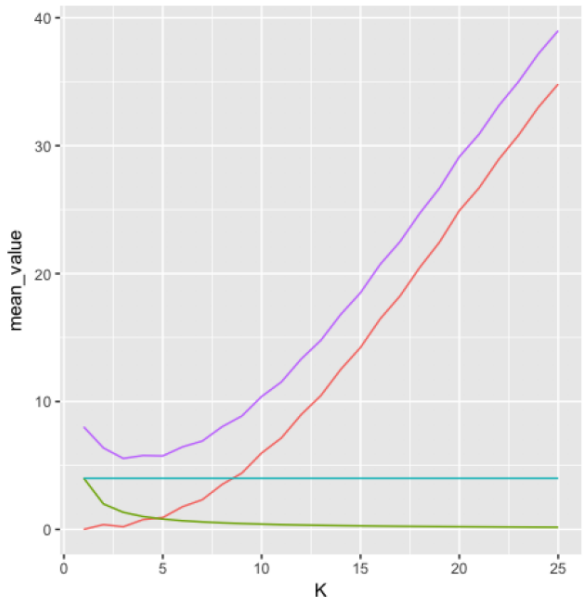

Figure 1: Squared bias, variance, irreducible error and total error for increasing values of K in KNN

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 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.

(iii) In KNN classification, it is important to use the test set to select the value K, and not the training set, to avoid overfitting.
(iv) In a linear regression setting, adding more covariates will reduce the variance of the predictor function.

e) (1P) Single choice

$\mathbf{X} = [x_1, x_2, x_3]^T$ is a 3-dimensional random vector with covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 50 & 33 & 18 \\ 33 & 38 & -10 \\ 18 & -10 & 72 \end{bmatrix}$$

The correlation between element x_1 and x_2 of the vector $\mathbf{X}$ is:

(i) 0.017
(ii) -0.19
(iii) 0.76
(iv) 0.66
(v) 0.10
(vi) 0.3
(vii) It is not possible to calculate the correlation, because this is not a proper covariance matrix.

Problem 2 (9P)

In the following example, Basil has been given a dataset by his boss. The dataset consists of observations of Antarctic penguins who live on the Palmer Archipelago. Basil's boss wonders if he can set up a model to predict the body mass of a given penguin based on some recorded characteristics of the penguin, which are specified in advance based on expert knowledge. However, Basil is a cat, and despite being very clever, he has only a very rudimentary knowledge of statistical techniques and data analysis. In the following code and report, Basil has made a couple of very problematic mistakes.

```
############### =^._.^= ~~~BASIL'S CODE~~~ =^._.^= ###############
############### install.packages('palmerpenguins') # Run if you haven't installed this before.
library(palmerpenguins)  # Contains the data set 'penguins'.
data(penguins)

# Remove island, and year variable, as we won't use those.
Penguins <- subset(penguins, select = -c(island, year))

# Fit the model as specified in advance based on expert knowledge:
penguin.model <- lm(body_mass_g ~ flipper_length_mm + sex + bill_depth_mm * species,
    data = Penguins)

# Look at the model coefficients
summary(penguin.model)$coefficients
```

```
##                                 Estimate Std. Error   t value     Pr(>|t|)
## (Intercept)                -1336.58287 646.922248 -2.066064 3.961450e-02
## flipper_length_mm             17.37877   2.910449  5.971165 6.172012e-09
## sexmale                      432.90151  44.633685  9.698987 1.059323e-19
## bill_depth_mm                 82.98484  22.324227  3.717255 2.370966e-04
```

```
## speciesChinstrap                1460.14721 680.389708  2.146045 3.260954e-02
## speciesGentoo                    644.88114 542.573989  1.188559 2.354811e-01
## bill_depth_mm:speciesChinstrap   -83.53310  37.009147 -2.257093 2.466587e-02
## bill_depth_mm:speciesGentoo       36.17178  34.481962  1.049006 2.949549e-01
```

```r
# Fit final model without sex
final.model <- lm(body_mass_g ~ flipper_length_mm + bill_depth_mm * species, data = Penguins)

summary(final.model)
```

```
##
## Call:
## lm(formula = body_mass_g ~ flipper_length_mm + bill_depth_mm *
##     species, data = Penguins)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -895.42 -226.28  -24.56  207.65 1074.66
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    -4213.176    647.731  -6.505 2.84e-10 ***
## flipper_length_mm                 24.621      3.173   7.760 1.04e-13 ***
## bill_depth_mm                    176.443     22.580   7.814 7.22e-14 ***
## speciesChinstrap                1008.380    771.358   1.307   0.1920
## speciesGentoo                    129.453    608.383   0.213   0.8316
## bill_depth_mm:speciesChinstrap   -61.538     41.978  -1.466   0.1436
## bill_depth_mm:speciesGentoo       78.026     38.545   2.024   0.0437 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 327.3 on 335 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.8364, Adjusted R-squared:  0.8335
## F-statistic: 285.5 on 6 and 335 DF,  p-value: < 2.2e-16
```

REPORT: PREDICTION OF PENGUIN BODY MASS, by Basil Thecat :3

We begin with a linear regression model with body mass as the response, and flipper length, bill depth, species, and sex as covariates, as well as an interaction effect between bill depth and species. In this first model, the sex covariate has the smallest p-value and is thus excluded in the final model to avoid overfitting. The final model can be described depending on the species of the penguin:

$$\hat{y}_{adelie} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + \hat{\beta}_{bill\_depth} x_{bill\_depth}$$

$$\hat{y}_{chinstrap} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + (\hat{\beta}_{bill\_depth} + \hat{\beta}_{bill\_depth:chinstrap}) x_{bill\_depth} + \hat{\beta}_{chinstrap}$$

$$\hat{y}_{gentoo} = \hat{\beta}_0 + \hat{\beta}_{flipper\_length} x_{flipper\_length} + (\hat{\beta}_{bill\_depth} + \hat{\beta}_{bill\_depth:gentoo}) x_{bill\_depth} + \hat{\beta}_{gentoo}$$

(where $\hat{y}_{adelie}$ is the predicted body mass for Adelie penguins, $\hat{\beta}_0$ is the estimated intercept, $x_{flipper\_length}$ is the flipper length covariate, $\hat{\beta}_{flipper\_length}$ is the estimated flipper length coefficient, etc.) Since both of the species coefficients have large p-values, we do not reject the null hypothesis that the species coefficient overall is actually zero. For the interaction effect between species and bill depth, the Gentoo interaction is significant ($p < 0.05$), so the interaction term overall is significant. Based on the coefficient for the dummy variable for the chinstrap penguins being the largest ($\hat{\beta}_{chinstrap} \approx 1008$), we can tell that the chinstrap penguins have the largest body mass.

a) (3P)

Identify three of the mistakes Basil made (there are more than three, but only report three). List them as bullet points along with brief explanations of why these are inappropriate modeling choices.

b) (1P)

In order to make an improved model, you will need to understand the data. Create at least one informative plot that helps you explain at least one of Basil's mistakes and that will justify your modeling choices in the next step. (Plotting the data may even help you *discover* Basil's mistakes in the first place.)

c) (5P)

Redo Basil's analysis including code and report, this time doing it right (4P). Evaluate the fit of the model with at least one graphical tool (1P).

Problem 3 (13P)

We will now consider the Palmer penguin dataset again, but this time looking at classifying the species of the penguins for a given body mass and flipper length. Since there are three penguin species, for simplicity we will define the goal to be to classify a penguin as belonging to the species Adelie or *not* Adelie, giving us a two-class classification problem instead of three.

The following code modifies the data set for this simplified setting, converts the variables to numeric (because the `knn` function can't handle the `int` class, and will give an error), and removes any missing observations. Please remember to use the same seed when you split the data into training and test set.

```
library(tidyverse)
library(GGally)
# Create a new boolean variable indicating whether or not the penguin is an
# Adelie penguin
Penguins$adelie <- ifelse(Penguins$species == "Adelie", 1, 0)

# Select only relevant variables and remove all rows with missing values in body
# mass, flipper length, sex or species.
Penguins_reduced <- Penguins %>% dplyr::select(body_mass_g, flipper_length_mm, adelie) %>%
    mutate(body_mass_g = as.numeric(body_mass_g), flipper_length_mm = as.numeric(flipper_length_mm)) %>%
    drop_na()

set.seed(4268)

# 70% of the sample size for training set
training_set_size <- floor(0.7 * nrow(Penguins_reduced))

train_ind <- sample(seq_len(nrow(Penguins_reduced)), size = training_set_size)

train <- Penguins_reduced[train_ind, ]
test <- Penguins_reduced[-train_ind, ]
```

a) (5P)

(i) (1P) Fit a **logistic regression** model using the training set, and perform the classification on the test set, using a 0.5 cutoff.

(ii) (1P) Fit a **QDA** model using the training set, and perform the classification on the test set, using a 0.5 cutoff.

(iii) (1P) Finally, do the same as in (i) and (ii) using **KNN** with $k = 25$ (use the `knn` function from the `class` package).

R-hints: In the `knn()` function set `prob=T` to ensure you get the class probabilities that you then need in d):

```
knnMod = knn(train = ..., test = ..., cl = ..., k = 25, prob = T)
```

(iv) (2P) Calculate the sensitivity and specificity for the three predictions performed on the test set in (i) - (iii).

b) (5P)

(i) Present a plot of the ROC curves and calculate the area under the curve (AUC) for each of the classifiers in a) (1P for each model).
(ii) Briefly discuss the ROC curves and the AUC. Which model performs best and worst (1P)?
(iii) If the task is to create an interpretable model, which model would you choose (1P)?

R-hints:

- To obtain $P(y = 1)$ from the `knn()` output you have to be aware that the respective probabilities

```
attributes(knnMod)$prob
```

are the success probability for the actual class where the categorization was made. So if you want to get a vector for $P(y = 1)$, you have to use $1 - P(y = 0)$ for the cases where the categorization was 0:

```
probKNN = ifelse(knnMod == 0, 1 - attributes(knnMod)$prob, attributes(knnMod)$prob)
```

- You might find the functions `roc()` and `ggroc()` from the package `pROC` useful, but there are many ways to plot ROC curves.

c) (1P) Single choice

We are again looking at the logistic regression model that you fitted to the training data in a).

According to this model, how would the odds that an observed animal is from the *Adelie* species change if the body mass increases by 1000 g? (the flipper length stays the same)

i) We add 0.712.
ii) We multiply it with 0.002.
iii) We multiply by 2.038.
iv) We multiply by 0.712.
v) We add 2.038.
vi) We multiply by 1000.

d) (2P)

Plot the full data (including both training nd test set) with the two covariates as the x- and y-axis, and use color and some other attribute of your choice (e.g. shape or highlight) to visualize the true species (adelie/not

adelie) as well as the predicted species from the best model in b) (note that the model should only be fitted with the training data as in b), but you are showing the data and predictions for both training and test data in the same plot).

Problem 4 (10P)

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

b) (2P)

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`), sex (0=female, 1=male) and smoking status (0=no, 1=yes). Load the data in R as follows

```
id <- "1chRpybM5cJn4Eow3-_xwDKPKyddL9M2N"  # 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`, `sex` and `smoking` as covariates. What is the probability of chd for a non-smoking male with a sbp=150 in the given dataset?

c) (4P)

We now use the bootstrap to estimate the uncertainty of the probability derived in b). Use $B = 1000$ bootstrap samples and proceed as follows:

- In each iteration, derive and store the estimated probability for `chd`, given `sbp`=150, `sex`=male and `smoking`=0 (1P for implementing the bootstrap).
- From the set of estimated probabilities, derive the standard error (1P).
- Derive the 95% quantile interval for the bootstrap samples (that is, the interval with limits at 2.5% and 97.5%) (1P).
- Interpret what you see. What is the expected probability and what are plausible values? (1P)

d) Multiple choice - 2P

We continue with the same dataset to study some properties of the bootstrap method. Below we estimated the standard errors of the regression coefficients in the logistic regression model with `sex`, `sbp` and `smoking` as predictors using 1000 bootstrap iterations (column `std.error`). These standard errors can be compared to those that we obtain by fitting a single logistic regression model using the `glm()` function (in Problem 4b). Look at the R output below and compare the standard errors that we obtain from the bootstrap with those we get from the `glm()` function (note that the `t1*` to `t4*` variables are sorted in the same way as for the `glm()` output).

```r
set.seed(4268)
library(boot)
boot.fn <- function(data, index) {
    return(coefficients(glm(chd ~ sbp + sex + smoking, family = "binomial", data = data,
        subset = index)))
}
boot(d.chd, boot.fn, 1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d.chd, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##        original        bias    std. error
## t1* -6.65883685  0.0142486669  2.54155686
## t2*  0.03877165 -0.0001413491  0.01921332
## t3* -1.34351384 -0.0462122125  0.33868464
## t4*  0.41031080 -0.0220322661  0.33473663
```

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

(i) The bootstrap relies on random sampling the same data without replacement.
(ii) The estimated standard errors from the `glm()` function are smaller than those estimated from the bootstrap, which indicates a problem with the bootstrap.
(iii) In general, differences between the estimated standard errors from the bootstrap and those from `glm()` may indicate a problem with the assumptions taken in logistic regression.
(iv) The p-values from the `glm()` output are probably slightly too small.