TMA4268 V2023 Exam
TMA4268 Statistical Learning V2023

Stefanie Muff, Department of Mathematical Sciences, NTNU

June 1, 2023

Maximum number of points: 55.5

Warming up

Problem 1 (Fill-in-the-blank text, 4.5P)

Read the whole text and fill in the blanks such that the whole text makes sense (you might only understand which answer is correct after you continued reading):

We have discussed a lot of methods and models in our course, and the methods are broadly divided into supervised and (*parametric*, *non-parametric*, *regression*, *unsupervised*, *classificiation*) approaches. In the latter case we (*cannot do inference*, *do not have a response*, *do not learn from the data*, *tend to overfit*) and therefore are not interested in prediction.

In supervised statistical learning methods there are two main purposes: (*prediction*, *inference*, *bias reduction*, *variance reduction*, *supervised learning*, *unsupervised learning*) and (*prediction*, *bias reduction*, *inference*, *variance reduction*, *unsupervised learning*, *supervised learning*). In both cases we want to learn from data and build a model that relates a set of variables to an outcome, but in the first case we do not interpret the actual model parameters. Some of the methods we learned about were (*non-parameteric*, *supervised*, *unsupervised*, *parametric*) and others were (*non-parametric*, *parameteric*, *supervised*, *unsupervised*), whereas the latter tend to be more more flexible – and thus possibly less biased – than the former ones, but at the cost of (*curse of dimensionality*, *higher test MSE*, *higher variance* , *lower accuracy*). This phenomenon is denoted as (*bias-variance trade-off*, *overfitting*, *underfitting*, *bias*, *variance*, *model selection bias*, *regularization*).

In the course we learned about models with different levels of complexity. For complex models, or in the presence of many variables, we have to be careful that we do not over-fit the data. We learned about several techniques that help prevent over-fitting via (*model selection*, *scaling*, *transformation of variables*, *regularization*, *bias-variance trade-off*), for example shrinkage methods, dimension reduction or dropout in neural networks.

Problem 2 (7P)

a) (2P)

Sketch the tree corresponding to the partition of the predictor space illustrated in the figure. The numbers inside the boxes are the mean of the response y within the regions.
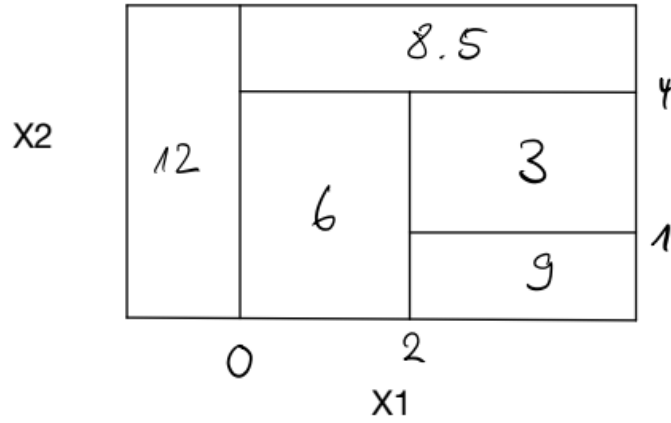
a) (5P)

In this problem, we consider a simulated data set with two classes (labelled 0 and 1) and two numerical covariates x_1 and x_2 . Let $x = (x_1, x_2)$ be a column vector with the two covariates. A training set with 500 observations of each class is available, and a scatter plot is given below. We simulate a data set as follows:

- Prior class probabilities: $\pi_0 = P(Y = 0) = 0.5$ and $\pi_1 = P(Y = 1) = 0.5$.
- Class-specific probabilities

$$P(x|y = 0) = f_0(x) = 0.5 \cdot \frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(x - \mu_{01})^\top \Sigma^{-1}(x - \mu_{01})\right) +$$
$$0.5 \cdot \frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(x - \mu_{02})^\top \Sigma^{-1}(x - \mu_{02})\right)$$

$$P(x|y = 1) = f_1(x) = \frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(x - \mu_{2})^\top \Sigma^{-1}(x - \mu_{2})\right)$$

with $\mu_{01} = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$, $\mu_{02} = \begin{pmatrix} 4 \\ 2 \end{pmatrix}$, $\mu_2 = \begin{pmatrix} 3 \\ 5 \end{pmatrix}$ and $\Sigma = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.
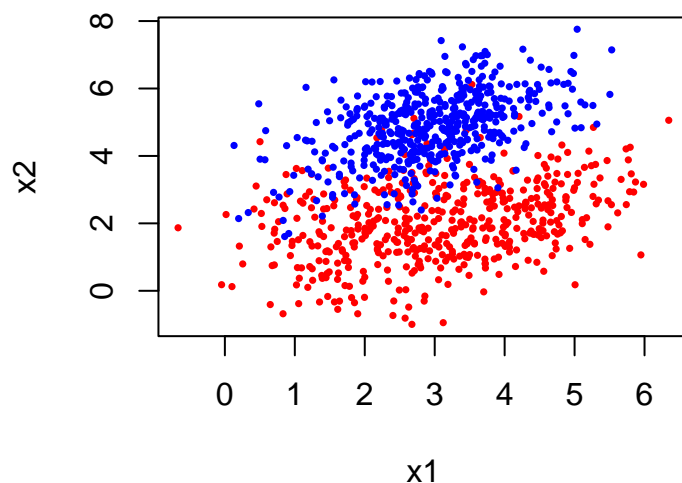
Given that $\pi_0 = \pi_1 = 0.5$, and the knowledge about the class-specific distributions $f_0(x)$ and $f_1(x)$ given above, the aim is to derive the equation for the Bayes decision boundary to find the Bayes classifier.

   (i) (1P) Explain what the Bayes decision boundary actually is.
  (ii) (3P) Write down the equation to be solved with the actual values (you are not supposed/asked to solve the equation), and explain what the unknowns are.
 (iii) (1P) Is the resulting boundary linear in the covariates? Why?

## Problem 3 – Data analysis 1 – regression (17P)

Here we are looking at a regression problem, where we want to understand the factors that affect birth weight of babies. We use the **birthwt** data set from the MASS package, which you can load and investigate using the code below and by typing **?birthwt** into the R console:

```r
library(MASS)
data(birthwt)

d.bw <- birthwt

# Race is a categorical variables, so we have to convert it:
d.bw$race <- as.factor(d.bw$race)

# Look at the data, for example using:
pairs(d.bw)
```
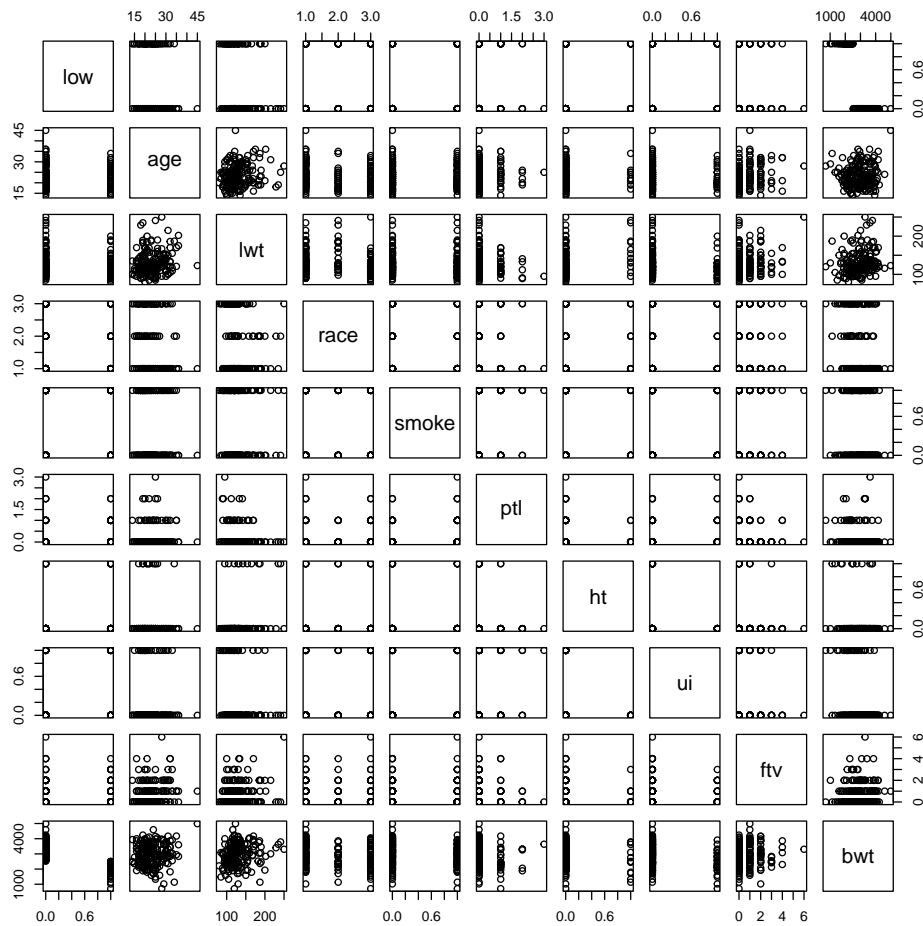
```
str(d.bw)
```

```
## 'data.frame':    189 obs. of  10 variables:
##  $ low  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ age  : int  19 33 20 21 18 21 22 17 29 26 ...
##  $ lwt  : int  182 155 105 108 107 124 118 103 123 113 ...
##  $ race : Factor w/ 3 levels "1","2","3": 2 3 1 1 1 3 1 3 1 1 ...
##  $ smoke: int  0 0 1 1 1 0 0 0 0 1 1 ...
##  $ ptl  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ht   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ ui   : int  1 0 0 1 1 0 0 0 0 0 0 ...
##  $ ftv  : int  0 3 1 2 0 0 1 1 1 0 ...
##  $ bwt  : int  2523 2551 2557 2594 2600 2622 2637 2637 2663 2665 ...
```

In order to assess the robustness of our models, we split the data into a training and a test set as follows:

```
set.seed(1234)
samples <- sample(1:189, 132, replace = F)
d.bw.train <- d.bw[samples, ]
d.bw.test <- d.bw[-samples, ]
```

a) (4P)

(i) (1P) Fit a linear regression model on the training data, where you use birth weight in grams (`bwt`) as the response and all variables **except low** as predictors. Report the regression coefficients, standard errors and p-values (use the standard way to do this in R).

(ii) (1P) What is the expected difference in birth weight for babies of black women compared to white women?

(iii) (1P) Is there evidence for `race` being relevant in the model? Say which test you carry out and report the respective p-value.

(iv) (1P) Compare R^2 and R^2_{adj} of the model and interpret what you find.

b) (3P)

(i) (2P) A medical researcher is interested to know whether the effect of the mother's smoking status during pregnancy changes with the age of the mother. Expand the model from a) such that it accounts for the possibility that the effect of smoking depends on age and interpret the results, still using the training data. In particular, answer the question of the researcher and compare R^2 to the one from a).

(ii) (1P) According to your modeling output, when the mother is smoking, how much does the expected birth weight change for a mother age 35 compared to a mother age 25, given that all other variables are the same?

c) (3P)

(i) (1P) Carry out Lasso regression on the training set excluding the variable `low` (like in a) and b)), and say how you choose λ.

(ii) (2P) Compare the regression coefficients from the Lasso with those from the linear regression in a). What pattern(s) do you notice? Would you see the same if you had carried out ridge regression?

R-hints:

```
x.train <- model.matrix(bwt ~ . - low, data = d.bw.train)[, -1]
y.train <- d.bw.train$bwt
x.test = model.matrix(bwt ~ . - low, data = d.bw.test)[, -1]
y.test = d.bw.test$bwt

library(glmnet)
set.seed(10)
cv.lasso <- cv.glmnet(x.train, y.train, alpha = ...)
plot(cv.lasso)
cv.lasso$...
bw.lasso <- glmnet(..., ..., alpha = ..., lambda = ...)
```

d) (3P)

Report the MSEs for the test set for all the models fit in a), b) and c). Which method performs best?

e) (4P)

Finally choose a more flexible method than those used above to find a model with *lower test error*. Explain your model and the choices you make. Code without explanation will not give full score.

R-hints:

```
library(gam)
library(randomForest)
```

Problem 4 – Data analysis 2 – classification (12P)

We look at the same data set as in Problem 3. Use the same code to load and prepare the data:

```
library(MASS)
data(birthwt)

d.bw <- birthwt

# Race is a categorical variables, so we have to convert it:
d.bw$race <- as.factor(d.bw$race)
# Also convert low for later use:
d.bw$low <- as.factor(d.bw$low)
```

In order to assess the robustness of our models, we split the data into a training and a test set as follows:

```
set.seed(1234)
samples <- sample(1:189, 132, replace = F)
d.bw.train <- d.bw[samples, ]
d.bw.test <- d.bw[-samples, ]
```

The aim is to both *understand* why and *predict* whether a newborn baby has low birthweight ($< 2.5 kg$), using the binary indicator variable `low` for being underweight.

a) (3P)

(i) (1P) Fit a logistic regression model that can be used to predict whether a baby will be underweight, including all predictor variables **except bwt** as predictors. Report the coefficients, standard errors and p-values (use the standard way to do this in R).

(ii) (2P) Report the probability to give birth to an underweight baby for a female with the following characteristics:

- age=25
- lwt (weight in pounds)=155
- race=white
- smoke = 1 (yes)
- ptl=0
- ht=0
- ui=0
- ftw=1

R-hints:

```
glm(..., family = "binomial")
```

b) (4P)

(i) (2P) A medical doctor is interested in the overall effect that smoking has on low birth weight, based on logistic regression. Use the regression output from a) and explain to the doctor how smoking affects the chance for low birth weight. We are interested in a general statement that is not dependent on the value of the other variables.

(ii) (1P) Report the error rate for the test data using the fitted model from a).

(iii) (1P) Calculate sensitivity and specificity for the test data.

R-hints:

```
glm(..., family = "binomial")
predict(...)
confusionMatrix(data = ..., reference = ...)$table
```

c) (5P)

As a comparison to the logistic regression model we are now using a random forest.

  (i) (2P) Use a random forest to fit a model on the training data. Justify the choice of any parameters that you use.
  (ii) (3P) Report the misclassification error (1P) sensitivity and specificity (1P) and compare the values to the logistic regression model above. Which of the methods is preferable and why (1P)?

R-hints:

```
library(randomForest)
set.seed(4268)

randomForest(...)
```

Multiple and single choice questions

Problem 5 (5P, single choice, 1P each)

a)

We want to estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 \quad \text{subject to} \quad \sum_{j=1}^{p}|\beta_j| \le s$$

for a particular value of s. Which statement is correct?

As we increase s from 0, the training RSS will

  (i) first increase and then start decreasing in an inverted U-shape.
  (ii) first decrease and then start increasing in a U-shape.
  (iii) steadily increase.
  (iv) steadily decrease.
  (v) remain constant.
  (vi) show an unpredictable behavior that is dependent on the data set.

b)

Same situation as in a). Which statement is correct?

As we increase s from 0, the bias will

  (i) first increase and then start decreasing in an inverted U-shape.
  (ii) first decrease and then start increasing in a U-shape.
  (iii) steadily increase.
  (iv) steadily decrease.
  (v) remain constant.
  (vi) show an unpredictable behavior that is dependent on the data set.

c)

We again look at the birth weight data set from Problem 4 to study some properties of the bootstrap method. Below we estimated the standard errors of the regression coefficients in the logistic regression model 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 with the same predictor variables. Look at the R output below and compare the standard errors that we obtain from these two approaches (note that the `t1*` to `t9*` variables are sorted in the same way as for the `glm()` output).

```
r.glm <- glm(low ~ age + lwt + race + smoke + ptl + ht + ui, d.bw.train,
    family = "binomial")
summary(r.glm)$coef
```

```
##               Estimate  Std. Error    z value   Pr(>|z|)
## (Intercept)  1.26251607 1.461824435  0.8636578 0.38777590
## age         -0.03052885 0.043538779 -0.7011876 0.48318593
## lwt         -0.01855679 0.008549737 -2.1704510 0.02997269
## race2        1.32200285 0.617089718  2.1423187 0.03216784
## race3        0.54395564 0.529144418  1.0279909 0.30395412
## smoke        0.51987574 0.483534486  1.0751575 0.28230421
## ptl          0.83095237 0.441190599  1.8834317 0.05964189
## ht           2.12658820 0.900192413  2.3623707 0.01815847
## ui           1.13555340 0.546814571  2.0766700 0.03783203
```

```
library(boot)
boot.fn <- function(data, index) {
    return(coefficients(glm(low ~ age + lwt + race + smoke + ptl + ht +
        ui, d.bw.train, family = "binomial", subset = index)))
}
boot(d.bw.train, boot.fn, 1000)
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
##
## Call:
## boot(data = d.bw.train, statistic = boot.fn, R = 1000)
##
##
## Bootstrap Statistics :
##        original         bias     std. error
## t1*   1.26251607   0.323468739  1.76661377
## t2*  -0.03052885  -0.006888524  0.04589292
## t3*  -0.01855679  -0.002316210  0.01059121
## t4*   1.32200285   0.062557947  0.69690421
## t5*   0.54395564   0.050796715  0.59579583
## t6*   0.51987574   0.015153417  0.54582304
## t7*   0.83095237   0.150903471  0.60891086
## t8*   2.12658820   1.059348986  4.12033986
## t9*   1.13555340   0.099963604  0.71110315
```

Which of the following statements is true?

(i) For some variables there are large differences between the estimated standard errors, which indicates a problem with the bootstrap.

(ii) The differences between the estimated standard errors indicate a problem with the assumptions taken about the distribution of the estimated parameters in logistic regression.

(iii) The bootstrap output indicates that the p-values estimated with logistic regression tend to be too large (i.e., too conservative).

(iv) The bootstrap relies on random sampling of the same data without replacement.

(v) The coefficients from logistic regression seem to be biased.

d)

Imagine a particular data set with only five observations. We carry out hierarchical clustering twice, once using single linkage and once using complete linkage. We obtain two dendograms. At a certain point in the single linkage dendogram the clusters {1,2} and {3,4,5} fuse (fusion 1). In the complete linkage dendogram the clusters fuse as well at a certain point (fusion 2).

Which statement is true?

(i) Fusion 1 will occur higher in the tree.

(ii) Fusion 2 will occur higher in the tree.

(iii) The fusions occur at the same height.

(iv) There is not enough information to tell which fusion is higher.

e)

We are dealing with a fully connected feed-forward neural network for classification into 5 categories. The input dimension of the data is 128 and we have two hidden layers with 32 and 64 nodes in layers 1 and 2, respectively. We fit the network using keras in R and decide to use a mini-batch of size 32, dropout of 20% in each layer and softmax activation in the output layer. Question: How many parameters do we need to estimate in total?

(i) 6565

(ii) 6464

(iii) 5252

(iv) 5171

(v) 6496

(vi) 3392

Problem 6 (10P, multiple choice, 2P each)

a) (2P)

Which of the following methods can be used to select a subset of variables for prediction?

(i) All types of regularization methods

(ii) Convolutional neural networks (CNNs)

(iii) Simple regression trees

(iv) Lasso

b) (2P)

Select all statements that are true for automatic model selection that minimizes AIC via forward or backward selection:

(i) Automatic model selection is not justified when the aim of the model is explanation (inference).

(ii) The procedure may introduce model selection bias.

(iii) The p-values of the final model tend to be too large.

(iv) If AIC is replaced by BIC we can circumvent the problems of the respective model selection procedure.

c) (2P)

Which statements about validation set approach, k-fold cross-validation (CV) and leave-one-out cross validation (LOOCV) are true?

- (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.

d) (2P)

Choose all the statements that are true:

- (i) In principal component regression (PCR) we automatically do variable selection when choosing a small number of PCs.
- (ii) In SVMs, when the cost parameter C is small, we tend to have low bias but high variance, and vice versa.
- (iii) The smoothing spline ensures smoothness of its function, g, by having a penalty term $\int g'(t)^2 dt$ in its loss.
- (iv) The K-nearest neighbors regression (local regression) has a high bias when its parameter, K, is high.

e) (2P)

We are looking at the following non-linear decision boundary for a support vector classifier:

$$f(X_1, X_2) = (X_1 + 2)^2 + (X_2 + 2)^2 - 2X_2^3 = 0 \ .$$

We assume that class 1 fulfills $f(X_1, X_2) > 0$ and class 2 fulfils $f(X_1, X_2) < 0$.

Which of the following statements are true?

- (i) This decision boundary is linear in terms of X_1, X_2, X_1^2 and X_2^2.
- (ii) The decision boundary has the shape of a circle.
- (iii) The point $(x_1, x_2) = (1, -1)$ belongs to class 1.
- (iv) The point $(x_1, x_2) = (1, 3)$ belongs to class 2.