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 *unsupervised* (*parametric*, *non-parametric*, *regression*, *classificiation*) approaches. In the latter case we *do not have a response,* (*cannot do inference*, *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 *inference* (*prediction*, *bias reduction*, *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 *parametric* (*non-parameteric*, *supervised*, *unsupervised*) 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 *higher variance* (*curse of dimensionality*, *higher test MSE* , *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 *regularization* (*model selection*, *scaling*, *transformation of variables*, *bias-variance trade-off*), for example shrinkage methods, dimension reduction or dropout in neural networks.

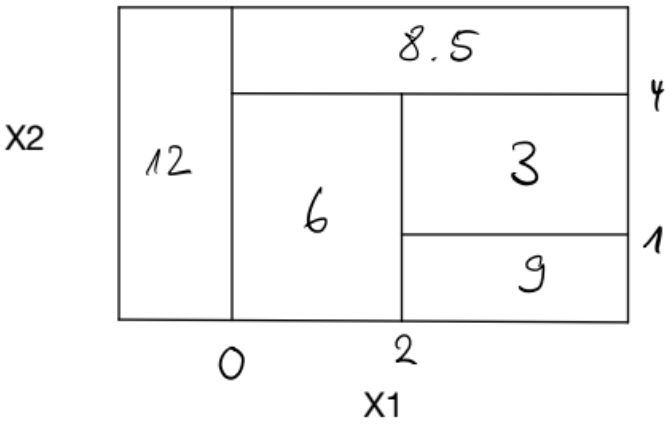
Note: The last sentence was taken out of the grading, because none of the answers are correct:

These approaches increase the robustness of the fitted models, where the main aim is to find the function that minimizes the expected test error, that is, the *irreducible error* (*reducible error*, *bias*, *variance*, *signal-to-noise ratio*).

Problem 2 (7P)

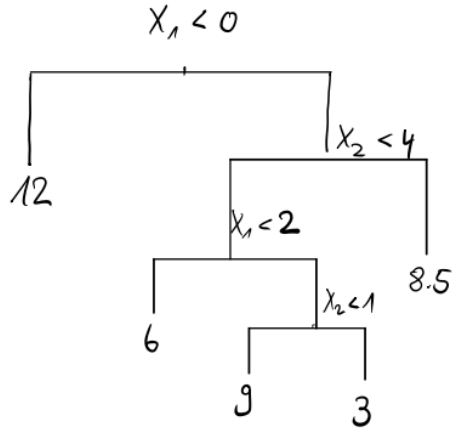
a) (2P)

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.



Solution:

1P for the correct tree structure with cut points, 1P for the correct values on the leafs.



a) (5P)

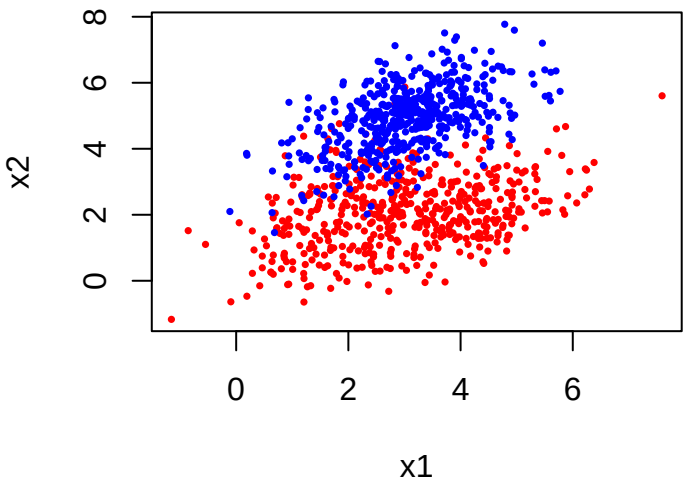
In this problem, we consider a simulated data set with two classes (labelled 0 and 1) and two numerical covariates x_1 and x_2 . Let $x = (x_1, x_2)$ be a column vector with the two covariates. A training set with 500 observations of each class is available, and a scatter plot is given below. We simulate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

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

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

with $\boldsymbol{\mu}_{01} = \begin{pmatrix}2\\2\end{pmatrix}$, $\boldsymbol{\mu}_{02} = \begin{pmatrix}4\\2\end{pmatrix}$, $\boldsymbol{\mu}_{2} = \begin{pmatrix}3\\5\end{pmatrix}$ 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?

Solution:

(i) Say either: The Bayes decision boundary is where the probabilities for the two classes are equal (0.5P for this), or: observations are classified to *the most probable class* (the class with the highest posterior probability; 0.5P for this alternative argument). Using this boundary thus gives *the minimum expected 0/1 loss* (0.5P for the last part, max 1P in total).

(ii) Here we need to find the equation where

$$f_0(x)\pi_0 = f_1(x)\pi_1 \ ,$$

and since $\pi_0 = \pi_1$, we only need to solve

$$f_0(x) = f_1(x) \ .$$

(1P if the students gets here). (Note that log-transformation does not help much here due to the sum in the likelihood for group 0.)

Thus, the equation is given as

$$\begin{aligned} &0.5 \cdot \frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{01})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{01})\right) + 0.5 \cdot \frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{02})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{02})\right) \\ =&\frac{1}{2\pi|\Sigma|} \cdot \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu}_{2})^\top \Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu}_{2})\right) \ , \end{aligned}$$

(1P),

where the students are expected to plug in the actual values for $\boldsymbol{\mu}_{01}$, $\boldsymbol{\mu}_{02}$, Σ etc. (1P for plugging in the values)

(iii) No, the boundary is not linear (and neither quadratic) in the covariates, because by log-transforming we do not get rid of the exponential term due to the sum in group 0.

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

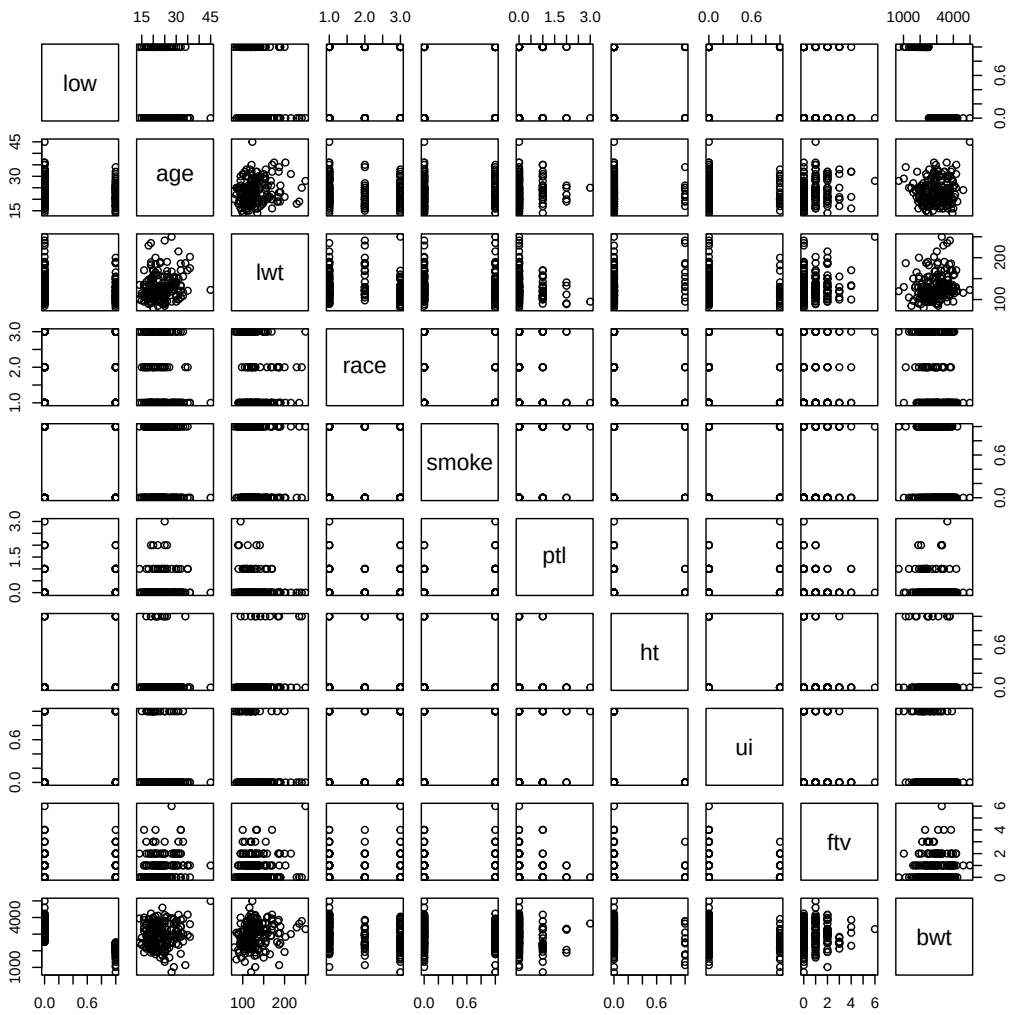
Here we are looking at a regression problem, where we want to understand the factors that affect birth weight of babies. We use the `birthwt` data set from the MASS package, which you can load and investigate using the code below and by typing `?birthwt` into the R console:

```
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 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 ...
##  $ 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.

Solution: (i)

```
r.lm1 <- lm(bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv, d.bw.train)
summary(r.lm1)
```

```
##
## Call:
## lm(formula = bwt ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, data = d.bw.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1676.18  -466.05    29.52   488.72  1772.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2774.220    379.761   7.305 3.13e-11 ***
## age           -4.608     11.681  -0.394 0.693916
## lwt            5.302      2.115   2.507 0.013500 *
## race2       -532.557    181.592  -2.933 0.004014 **
## race3       -326.391    141.538  -2.306 0.022797 *
## smoke       -264.044    133.499  -1.978 0.050197 .
## ptl         -146.597    130.288  -1.125 0.262724
## ht          -687.332    254.689  -2.699 0.007948 **
## ui          -628.922    169.085  -3.720 0.000303 ***
## ftv           -1.865     55.124  -0.034 0.973066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 667.4 on 122 degrees of freedom
## Multiple R-squared:  0.2695, Adjusted R-squared:  0.2157
## F-statistic: 5.002 on 9 and 122 DF,  p-value: 9.836e-06
```

(ii) -533 grams

(iii) F-test using the anova function in R:

```
anova(r.lm1)
```

```
## Analysis of Variance Table
##
## Response: bwt
##            Df   Sum Sq Mean Sq F value    Pr(>F)
## age         1   301426  301426  0.6768 0.4122979
## lwt         1  3267325 3267325  7.3362 0.0077297 **
## race        2  3311398 1655699  3.7176 0.0270876 *
## smoke       1  3332287 3332287  7.4820 0.0071615 **
## ptl         1  1016195 1016195  2.2817 0.1334964
## ht          1  2655528 2655528  5.9625 0.0160476 *
## ui          1  6165125 6165125 13.8426 0.0003018 ***
## ftv         1      510     510  0.0011 0.9730662
## Residuals 122 54335487  445373
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is small (<0.05), thus yes, there is evidence.

(iv) The two values should be reported here: $R^2 = 0.267$ and $R^2_{adj} = 0.216$. The difference in the two R^2 values indicates that there are some unneccesary variables in the model. Note: Students don't need to interpret the R^2 itself, only the differnces.

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?

Solution

(i) The model and its output are as follows (1P for the correct model, no point if the interaction is lacking and -0.5P for other mistakes):

```
r.lm2 <- lm(bwt ~ age + lwt + race + ptl + ht + ui + ftv + smoke * age,
    d.bw.train)
summary(r.lm2)
```

```
##
## Call:
## lm(formula = bwt ~ age + lwt + race + ptl + ht + ui + ftv + smoke *
##     age, data = d.bw.train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1696.40  -518.00    26.93   516.30  1398.18
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2286.122    427.968   5.342 4.39e-07 ***
## age           14.421     14.089   1.024  0.30810
## lwt            5.268      2.078   2.535  0.01251 *
## race2       -451.590    181.752  -2.485  0.01433 *
## race3       -268.483    141.252  -1.901  0.05972 .
## ptl         -131.475    128.156  -1.026  0.30699
## ht          -676.711    250.241  -2.704  0.00783 **
## ui          -689.382    168.124  -4.100 7.50e-05 ***
## ftv            8.813     54.347   0.162  0.87145
## smoke       1044.110    577.136   1.809  0.07291 .
## age:smoke    -55.732     23.945  -2.328  0.02160 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 655.6 on 121 degrees of freedom
## Multiple R-squared:  0.3008, Adjusted R-squared:  0.2431
## F-statistic: 5.207 on 10 and 121 DF,  p-value: 2.483e-06
```

Interpretation: Yes, the effect of smoking changes with age (0.5P). Older women have an even higher risk of low birth weight when they smoke than younger women. The $R^2 = 0.30$ has clearly improved, underlining that the interaction term is relevant (0.5P).

(ii) The overall effect of age is given by $\beta_{age} + \beta_{age:smoking}$, thus we have to multiply this term by 10 to obtain the expected reduction in birth weight: -413.1. Here there is an all-or-nothing grading (1P for correct value, 0P otherwise).

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
```

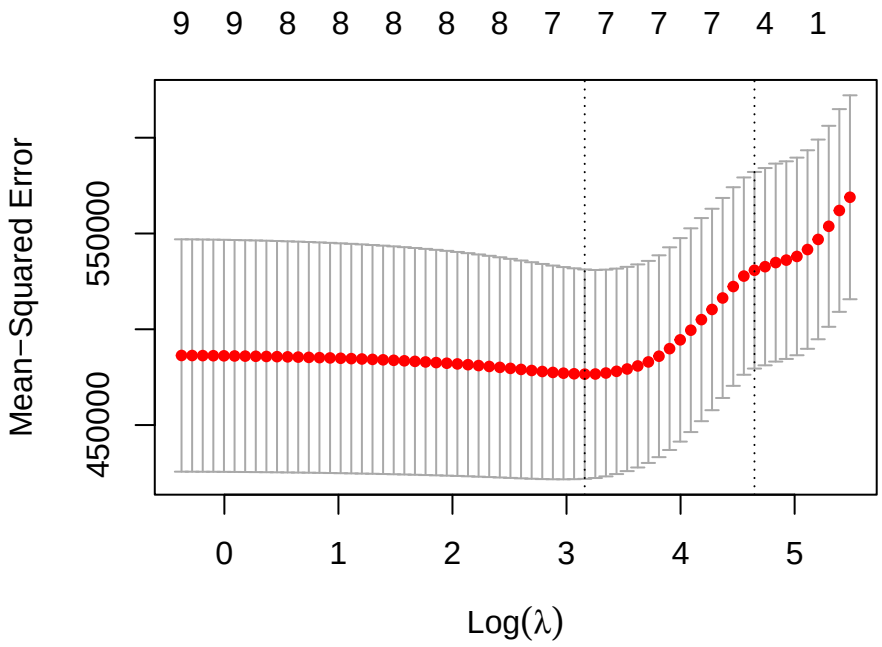
```
library(glmnet)
set.seed(10)
cv.lasso <- cv.glmnet(x.train, y.train, alpha = ...)
plot(cv.lasso)
cv.lasso$...
bw.lasso <- glmnet(..., ..., alpha = ..., lambda = ...)
```

Solution

(i)

```
library(glmnet)
set.seed(10)
cv.lasso <- cv.glmnet(x.train, y.train, alpha = 1)
```

```
plot(cv.lasso)
```




```
cv.lasso$lambda.min
```

```
## [1] 23.55878
```

```
bw.lasso <- glmnet(x.train, y.train, alpha = 1, lambda = cv.lasso$lambda.min)

coef(bw.lasso)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) 2685.110986
## age            .
## lwt            4.337968
## race2       -396.970442
## race3       -225.086479
## smoke       -187.480612
## ptl         -129.089274
## ht          -552.987923
## ui          -573.098758
## ftv            .
```

Alternatively, with `lambda.1se`:

```
cv.lasso$lambda.1se
```

```
## [1] 104.38
```

```
bw.lasso.1se <- glmnet(x.train, y.train, alpha = 1, lambda = cv.lasso$lambda.1se)

coef(bw.lasso.1se)
```

```
## 10 x 1 sparse Matrix of class "dgCMatrix"
##                       s0
## (Intercept) 2778.223411
## age            .
```

```
## lwt            1.446888
## race2        -20.086476
## race3          .
## smoke          .
## ptl          -31.634403
## ht          -121.384488
## ui          -374.208781
## ftv            .
```

λ is chosen using cross-validation (0.5P).

(ii) There are two patterns to notice: The coefficients are shrunken (0.5P) and some are zero (0.5P). For ridge we would also have expected to see shrinkage (0.5P), but none of the coefficients had gone to zero (0.5P). Here I expect to explicitly hint that the coefficients are getting smaller/shurunken (not just that some are close to zero for ridge, for example).

d) (3P)

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

Solution:

```
r.lm1 <- lm(bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv, d.bw.train)
r.pred.lm1 <- predict(r.lm1, newdata = d.bw.test)
mse.a <- mean((r.pred.lm1 - d.bw.test$bwt)^2)

r.lm2 <- lm(bwt ~ age + lwt + race + smoke * age + ptl + ht + ui + ftv,
    d.bw.train)
r.pred.lm2 <- predict(r.lm2, newdata = d.bw.test)
mse.b <- mean((r.pred.lm2 - d.bw.test$bwt)^2)

r.lasso <- predict(bw.lasso, newx = x.test)
mse.c <- mean((r.lasso - d.bw.test$bwt)^2)

# Also for the case where the larger lambda is used in d):
r.lasso.1se <- predict(bw.lasso.1se, newx = x.test)
mse.c.1se <- mean((r.lasso.1se - d.bw.test$bwt)^2)

c(mse.a, mse.b, mse.c, mse.c.1se)
```

```
## [1] 406158.6 417858.8 396418.9 431157.5
```

Lasso performs best if lambda.min was used, otherwise standard regression 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)
```

Solution:

Here there is a lot of freedom. The students will probably use a GAM or a regression tree.

In a regression tree, the students must explain the choice for the number of trees and mtry (if either is lacking, -1P for each). It's not enough to say "I choose the default no of trees from the lecture/function" or "I chose ntree=500", for example – ntrees should be chosen *large enough* (i.e., it is no tuning parameter), and that should be made clear. Note that ntrees is not a tuning parameter, so the students should not optimize it. If students choose a GAM, they also need to explain each term and the choices they made.

Here is one solution: The idea is to use only variables that are left from the Lasso and then apply a cubic spline on the only continuous variable `lwt`:

```
library(gam)
r.gam <- gam(bwt ~ bs(lwt, 4) + race + smoke + ptl + ht + ui, data = d.bw.train)
r.pred.gam <- predict(r.gam, newdata = d.bw.test)
mean((r.pred.gam - d.bw.test$bwt)^2)
```

```
## [1] 392554.2
```

Here is a solution with random forests. Note that the students then need to explain `mtry` (usually p/3 for regression problems) and say that number of trees must be "large enough".

```
library(randomForest)
rf.model <- randomForest(x.train, y.train, mtry = 3, ntree = 500)
yhat.rf <- predict(rf.model, newdata = x.test)
(mean((yhat.rf - y.test)^2))
```

```
## [1] 350918.7
```

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 ($< 2.5kg$), 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")
```

Solution:

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

```
##
## Call:
## glm(formula = low ~ age + lwt + race + smoke + ptl + ht + ui +
##     ftv, family = "binomial", data = d.bw.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
```

```
## -1.5624  -0.8019  -0.5389   0.9842   2.2859
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.2626642  1.4627818   0.863   0.3880
## age         -0.0305166  0.0437289  -0.698   0.4853
## lwt         -0.0185549  0.0085725  -2.164   0.0304 *
## race2        1.3220153  0.6171069   2.142   0.0322 *
## race3        0.5436934  0.5362748   1.014   0.3107
## smoke        0.5196261  0.4905968   1.059   0.2895
## ptl          0.8310422  0.4422122   1.879   0.0602 .
## ht           2.1264726  0.9010048   2.360   0.0183 *
## ui           1.1355797  0.5468967   2.076   0.0379 *
## ftv         -0.0006133  0.2035747  -0.003   0.9976
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 169.39  on 131  degrees of freedom
## Residual deviance: 142.45  on 122  degrees of freedom
## AIC: 162.45
##
## Number of Fisher Scoring iterations: 4
```

(ii) One way to implement this is as follows (but you can also use the `predict()` function):

```
x.data <- c(1, 25, 155, 0, 0, 1, 0, 0, 0, 1)

eta <- t(coef(r.glm)) %*% x.data

(prob <- exp(eta)/(1 + exp(eta)))
```

```
##           [,1]
## [1,] 0.1350238
```

The probability is 0.135.

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
```

Solution:

(i) Here the students must NOT repeat the analysis from (a) with smoke=0, but look at the odds-ratio:

```
exp(coef(r.glm)["smoke"])
```

```
##    smoke
## 1.681399
```

Interpretation: The odds for low birth weight increases by a *factor* of 1.68 for smoking women compared to non-smoking women.

Note: Several students misunderstood the question and did not do any calculation, but only interpreted the result qualitatively. I have 1P for this if the explanation was correct anyway.

(ii)

```
t.pred.glm <- round(predict(r.glm, newdata = d.bw.test, type = "response"),
    0)

(confMat <- confusionMatrix(data = as.factor(t.pred.glm), reference = as.factor(d.bw.test$low))$table)
```

```
##           Reference
## Prediction  0  1
##          0 36  8
##          1  7  6
```

```
1 - (sum(diag(confMat))/sum(confMat[1:2, 1:2]))
```

```
## [1] 0.2631579
```

(iii) Sensitivity: Corresponds to $P(\hat{y} = 1|y = 1)$, where y is the true (reference) value, thus 0.429. Specificity: Corresponds to $P(\hat{y} = 0|y = 0)$, thus 0.837.

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(...)
```

Solution:

(i) 1P to fit the model:

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

rf.bw = randomForest(low ~ age + lwt + race + smoke + ptl + ht + ui +
    ftv, data = d.bw.train, mtry = 2, ntree = 1000, importance = TRUE)
```

The most critical choice is `mtry`. Since we have 8 variables, and $\sqrt{8} = 2.8$, we can choose `mtry` =2 or =3. Both are ok, but need to be justified (0.5P). The number of trees is less critical but needs to be "large enough" (0.5P for the argument).

(ii)

```
t.pred.rf <- as.factor(predict(rf.bw, d.bw.test, type = "class"))
(confMat.rf <- confusionMatrix(data = t.pred.rf, reference = d.bw.test$low)$table)
```

```
##           Reference
## Prediction  0  1
##          0 40 11
##          1  3  3
```

```
1 - (sum(diag(confMat.rf))/sum(confMat.rf[1:2, 1:2]))
```

```
## [1] 0.245614
```

Sensitivity: Corresponds to $P(\hat{y} = 1|y = 1)$ where y is the true (reference) value, thus 0.214.
Specificity: Corresponds to $P(\hat{y} = 0|y = 0)$, thus 0.93.

Observation: In comparison to logistic regression, the misclassification error is lower for the random forest. However, The sensitivity for the forest is extremely low, and in a medical context that might be undesirable because we might overlook a risk patient. Therefore, logistic regression might be preferable.

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

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

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.

Solution: (iv) because the model becomes more flexible/less regularized and can thus fit to the traning data.

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.

Solution: (iv) because the smaller s the more the estimates are biased.

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  0.348307413  1.83881782
## t2* -0.03052885 -0.007057594  0.04626627
## t3* -0.01855679 -0.002464106  0.01102647
## t4*  1.32200285  0.020596973  0.98453760
## t5*  0.54395564  0.022521134  0.61590998
## t6*  0.51987574  0.006733729  0.54902431
## t7*  0.83095237  0.155702816  0.57702957
## t8*  2.12658820  0.876754508  3.98381366
## t9*  1.13555340  0.117687249  0.87502310
```

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.

Solution: (ii)

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.

Solution: (iv)

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

Solution: (i) is correct - remember that there is one bias-node in each layer!

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

Solution: FALSE, FALSE, TRUE, TRUE

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.

Solution: TRUE, TRUE, FALSE, FALSE

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.

Solution:

TRUE, TRUE, FALSE, FALSE

(i) and (ii) are correct. (iii) is wrong, because in 2-fold CV we would fit the model twice (once with each half of the data), while the validation set approach uses only one half of the data to fit the model. (iv) is wrong, because LOOCV is actually very expensive - expect for linear regression, where a formula exists.

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.

Solution

(i) False
(ii) True/False (both correct due to ambiguity)
(iii) False: The penalty term is $\int g''(t)^2 dt$.
(iv) True: because the local regression is based on k-nearest neighbor algorithm.

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.

Solution (i) False (X_2^3 does not cancel out). (ii) False (iii) True (iv) True