Exercises for Applied Biostatistics II - FS 2020

1. Go through the R introduction provided by the "CodeSchool": `http://tryr.codeschool.com/`

It is suggested that you sign up for a CodeSchool account first; otherwise, you cannot pause the course since you have to restart from the beginning each time you access the web page. With an account, you can also download and print out the completed worksheet in the end.

2. [**Graded**] In this exercise, we consider four data sets constructed by the British statistician Frank Anscombe. Each data set consists of a response variable Y and an explanatory variable X.

The data sets can be made available in R using the command

```
> data(anscombe)
```

After this, a data frame `anscombe` is available with the four variable pairs `(x1, y1)`, `(x2, y2)`, `(x3, y3)`, and `(x4, y4)`.

a) Display each of the 4 data sets as a scatter plot, draw in the regression line and comment on the results.
`R hints:` you can use the command `par(mfrow = c(2, 2))` to get the four plots side-by-side; to draw in the regression line you can use `abline(lm(Y ~ X))`.

b) For the four models, compare the estimated values $\widehat{\beta_0}$ (intercept), $\widehat{\beta_1}$ (slope) and $\widehat{\sigma}^2$ (variance of the residuals), as well as the "quality criterion" R^2.

3. [**Graded**] The file `farm.dat` contains the size A (in acres), the number of cows C and the income I (in $) of 20 farms in the US. You find the data set on ILIAS.

a) Compute an ordinary linear regression of I versus C. Does the income depend on the number of cows?

b) Give the confidence intervals for the expected income without any cows, with 20 cows, and with C = 8.85 cows.

c) Compute an ordinary linear regression of I versus A and a multiple linear regression of I versus A and C. Also compute the correlation between A and C. Finally, based on your results, explain the differences between the three regression models.

4. [**Graded**] In this exercise, we again consider the air pollution data set presented in the lecture. In a study on the contribution of air pollution to mortality, General Motors collected data from 60 US cities. The dependent variable is the age adjusted mortality (variable `Mortality`). The data includes variables measuring demographic characteristics of the cities, variables measuring climate characteristics, and variables recording the pollution potential of three different air pollutants.

The data set can be found in the file `airpollution.csv` available in ILIAS.

a) Get an overview of the data and account for possible problems. Which of the variables need to be transformed?

b) Carry out a multiple linear regression containing all variables. Does the model fit well? Check the residuals.

c) Now take all the non-significant variables out of the model and compute the regression again. Compare your results to part b.).

5. **[Graded]** On February 9, 2014, Swiss voters accepted the initiative "Against Mass Immigration". In this exercise, we will try to predict the acceptance in each canton based on goegraphic and demographic data.

The data set `massimmigration.csv` available in ILIAS contains the following variables:

`canton`	abbreviation of the canton
`yes`	acceptance (fraction of "Yes" votes) in % (response variable)
`area`	area im km^2
`inhabitants`	inhabitants of the canton
`foreigners`	fraction of foreigners in %

a) Plot acceptance versus the fraction of foreigners, and fit a linear model to the data. Does the model fit well? Analyse the residuals.

b) Plot a confidence band and prediction intervals into the plot of a), both for a confidence level of 90%. What is the difference between the two?

c) How well does the fraction of foreigners explain the acceptance in the different cantons? Calculate the coefficient of determination R^2 and the F statistic "by hand", i.e. only using the R functions `resid`, `fitted` and `mean`. Check your results with the output of `summary`.

d) Select the best linear model as follows:

1. Add a variable `density` to the data set, defined as the number of inhabitants per area.
2. Start with the full regression model.
3. As long as there is an explanatory variable with a p-value above 5%:
   - Remove the least significant variable.
   - Keep the new model if the larger model is not significantly better based on an F-test.

6. **[Graded]** Biologists studied the relationship between the length of a bullfrog and how far it can jump.

The resulting data set had 2 variables, `length` (body length, in mm) and `jump` (maximum leap distance, in cm). The variables were fitted in a linear model the output of which is shown in the following:

```
Call:
lm(formula = jump ~ length, data = bullfrog)

Residuals:
    Min      1Q  Median      3Q     Max
-34.864  -5.206   5.589  11.799  21.120

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  51.7416    59.5828   0.868    0.408
length        0.3492     0.3965   0.881    0.401

Residual standard error: 18.15 on 9 degrees of freedom
Multiple R-squared:  0.07933,        Adjusted R-squared:  -0.02296
F-statistic: 0.7755 on 1 and 9 DF,  p-value: 0.4014
```

a) Write down the linear regression model the biologists assumed.

b) How many frogs were included in the study?

c) Is the length of the frogs a good predictor for the maximum leap distance?

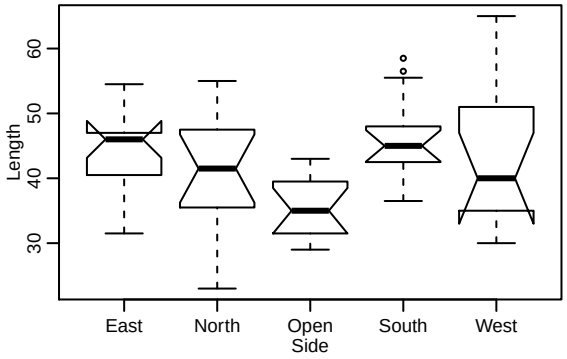
d) Fill out an ANOVA table for the model based on the R output above.

Source of variation	**df (degrees of freedom)**	**SS (sum of squares)**	**MS (mean square)**
regression	...	...	...
errors / resid.	...	...	...
total around global mean	...	...	

7. [**Graded**] A researcher collected daffodils from four sides of a building and from an open area nearby. She wondered whether the average stem length of a daffodil depends on its location. The data set is available as `daffodils.csv` from ILIAS.

a) State the null hypothesis of an ANOVA model for this problem in words and as a formula.

b) A boxplot of the data looks as follows:



Based on the boxplot, does it appear that the null hypothesis is true?

c) Fit an ANOVA model to the data and test the null hypothesis from a) on a significance level of 10%.

d) Does the ANOVA model fit well to the data? Perform a residual analysis.

e) Which locations (sides of the building and open area) are not significantly different on a 5% level? Use Bonferroni adjusted pairwise t-tests.

8. [**Graded**] A researcher studied the flexibility of women after taking different sports courses. The flexibility was measured by the spinal extension, a measure of how far the women could bend her back.

The ANOVA table of the data set looks as follows:

```
Analysis of Variance Table

Response: SpineExtension
          Df  Sum Sq Mean Sq F value   Pr(>F)
Activity   2  7.0357  3.5178  6.0667 0.006882 **
Residuals 26 15.0764  0.5799
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

a) How many groups (sports activities) were in the experiment? How many women participated?

b) What can you say on a 5% level about the null hypothesis that all sport courses lead to the same flexibility?

c) What is the pooled standard deviation, s_{pool}?

9. [**Graded**] In a study of the dietary treatment of anemia in cattle, 144 cows were randomly divided into four treatment groups A, B, C, and D; A was the control group. After a year of treatment, blood samples were drawn and assayed for selenium. The following table shows the mean selenium concentrations in μg/dl. The MS(within) from the ANOVA was 2.071.

Group	Mean	n_i
A	0.8	36
B	5.4	36
C	6.2	36
D	5.0	36

Compute three Bonferroni-adjusted confidence intervals comparing diets B, C, and D to the control (diet A) for a FWER of $\overline{\alpha} = 0.05$. Which comparison-wise significance level α do you have to use to calculate the confidence intervals?

10. [**Graded**] Researchers measured the yield of 5 crop species treated with 4 fertilizers. You find the measurements in the file `fertilizer.dat` on ILIAS.

a) Read in the data set and draw an interaction plot.
R hint: since the factor levels are not encoded as strings, you first have to explicitely convert the variables `CropSpecies` and `Fertilizer` to factors.

b) Fit an ANOVA model without interaction to the data set. Analyse the residuals. Does the model fit well? If not, can you think of a transformation that could make the fit better?

c) Add an interaction term to the model from task b). Is there a significant interaction?

d) Beginning from the model from task c), perform a model selection by remove non-significant terms. In which order do you have to test the terms? With which model do you end up?

11. [**Graded**] A scientist is interested in how genotype of a strawberry plant affects fruit yield. There are three levels of genotype (AA, AB, BB) and ten plots of land, three plants per plot. Each of the three genotypes is present in each plot. The data is in the file `strawb.dat` available on ILIAS.

a) Perform an ANOVA, assuming one-way randomized block design.

b) Repeat the analysis of variance without taking into account land effects.

c) Compare the results in a) and b). Why are the degrees of freedom different? Which result would you use?

12. [**Graded**] The Dutch coastal institute RIKZ measured the species richness at different beaches of the Dutch coast. The data set is available as `RIKZ.txt` in ILIAS and has the following variables:

`Sample`	sample ID
`Richness`	number of species found in a test area
`Exposure`	ordinal variable determining exposure of site, composed of different elements
`NAP`	height of sampling station relative to Normaal Amsterdams Peil $\approx$ mean sea level
`Beach`	index of beach

The aim of this exercise is to predict species richness with the explanatory variables.

a) Read in the data set with R. Look at the structure of the data frame you get. Do you need some manual corrections to get the data in the right format?

b) Fit a linear regression model using all explanatory variables. Look at the summary output of your fit, and analyse the residuals. Which problems do you see? How could you solve them?

c) Try to improve the quality of your model from task b) by transforming some of the variables in the RIKZ data set. Redo the model fit with the transformed variables. Does the model fit better now?

d) Continue with the transformation you chose in c). Perform model selection by iteratively removing explanatory variables. Which model do you end up with?

13. [**Graded**] Below, you find a list of explanatory variables that could appear in biological studies. For each of them, decide wether it makes more sense to model them as fixed or as random, and give a short explanation.

1. The index of a patient (subject) in a medical study. For each of the subjects, different physiological quantities were measured in order to predict the response to a certain drug.
2. The index of a biological replicate in a gene expression study, in which different cells were grown from the same strain of yeast.
3. The yeast strain in a gene expression study, in which cells from 5 different yeast strains were compared.

4. The type of a machine used for sequencing a virus genome in a study which compares sequencing errors produced by different technologies and at different parts of the genome.
5. The litter a rat in a behavioural study comes from.

14. [**Graded**] The data set `oxboys.csv` (available on ILIAS) consists of the heights of 26 boys from Oxford, each measured on 9 different occasions (at different ages). The data set consists of the following variables:

`subject`	ID of the boys
`age`	*centered* age
`height`	height in cm
`occasion`	index of measurement for each boy

a) Plot the height of the boys against the age. Fit a linear model to the data (ignoring the subjects), and draw the regression line into the plot. Is the linear model appropriate?

b) How much of the variance can be explained by the regression? Determine the R^2 value (from an R output).

c) Fit a random intercept model to the data set, using `subject` as the random effect. Again, plot the height vs. the age, and add the 26 regression lines for the different boys to the plot.

d) Now, fit a random intercept and slope model to the data set, and plot the individual regression lines over the data points. Do you think the random slope gives better fits to the data?

e) Validate the quality of the random intercept and slope model form d) with a Tukey-Anscombe and a Q-Q plot of the residuals. Does the model fit well?

f) The `summary()` function does not output an R^2 value for a mixed effects model, as opposed to fixed effects models. Calculate the R^2 value for the model in d) "by hand" (see Exercise 5.c). Compare it to the R^2 value of the fixed effects model from b), and explain the difference.

15. [**Graded**] The `Pastes` data set contains 60 quality measurements (variable `strength`) of a chemical paste delivered in different batches. From 10 randomly selected delivery batches (variable `batch`, values 'A' to 'J') three casks (variable `cask`, values 'a' to 'c') were randomly sampled and analyzed twice. This means that we have 30 samples in total (variable `sample`, values 'A:a' to 'J:c') and two measurements were carried out on each.

The data set can be loaded as follows:

```
> library(lme4)
> data(Pastes)
```

a) Assume you would have forgotten your knowledge on random effects and want to analyse the data set with an one- or two-way ANOVA. How do you have to specify the model, and what are the problems of such an approach?

b) Assuming you have not forgotten anything on random effects. Fit a two-way random effects model to the data set. Which model formula is appropriate? Which remarkable feature does the data have?

c) Does the strength significantly deviate between different casks? And between different batches?

16. [**Graded**] The data set `video.csv` (available on ILIAS) contains measurements of a study in which psychologists measured the ability of probands of different age groups in learning to play a video game. The data set has the following variables:

`id`	initials of the subjects
`age`	age of the probands
`trial`	1: first trial, 2: second trial, ..., 5: fifth trial
`score`	score in the video game (response variable)

You can treat `trial` as a numerical variable.

a) Plot the score against the trial for all probands. Fit a linear model to the data and draw the regression line into the plot.

b) Write down the model equation for the linear model you fitted in a). Which assumptions do you need? Are the assumptions fulfilled in this data set? If no, how could we improve the model? Motivate your answers.

c) Consider the model formula

$$\texttt{score} \sim \texttt{age} * \texttt{trial} + (1 + \texttt{trial}|\texttt{id})$$

How is such a model called? Translate this model formula into a mathematical formula, and write down the assumptions. Fit the model in R.

d) Assess the random effects from the model in c). Remove the non-significant ones.

e) Assess the fixed effects from the model in d). In which order do you have to check their significance? Remove the non-significant ones.

17. [**Graded**] In this exercise, we again consider the cement data set presented in the lecture. The data set is available in the R package `wle` and can be loaded as follows:

```
> data(hald, package = "wle")
```

You now have a matrix called `hald` which you can standardize and transform into a data frame as follows:

```
> cement <- data.frame(scale(hald))
> names(cement) <- c("y", "x1", "x2", "x3", "x4")
```

`y` is the (standardized) response variable (heat evolved in a cement mix), `x1` to `x4` are explanatory variables indicating the composition of the cement mix.

a) Start with an empty linear model and perform forward selection using the function `stepAIC` from the `MASS` package. Which linear model do you get in the end?

b) Repeat the analysis of a) on two reduced data sets:

1. the `cement` data set without row 3
2. the `cement` data set without row 10.

Do you get the same model as in a)?

c) For the model of a) and the two models of b), calculate the difference $\hat{\beta}_2 - \hat{\beta}_4$, where $\hat{\beta}_i$ stands for the fitted regression coefficient associated to variable X_i. What do you observe? How do you explain that?

d) Now, fit a linear model using ridge regression to the `cement` data set. Use a lambda range of $\lambda \in [0, 1]$. Which is the best λ value according to the GCV value?

e) Indicate the coefficients of the ridge solution from d) for the determined optimal λ. As in task b), repeat the ridge fit for the data set without row 3 and 10, respectively, both with the optimal λ from task d). What do you observe?

18. [**Graded**] Prostate-specific antigen (PSA) is an enzyme whose level is used as a (controversial) indicator of prostate cancer; however, a high level may also indicate other prostate diseases. The `Prostate` data set examines the correlation between PSA levels and other clinical measures in men who were about to receive a radical prostatectomy. The data set is included in the R package `lasso2` and can be loaded as follows:

```
> data(Prostate, package = "lasso2")
```

It contains the following variables:

`lcavol`	log(cancer volume)
`lweight`	log(prostate weight)
`age`	age
`lbph`	log(benign prostatic hyperplasia amount)
`svi`	seminal vesicle invasion
`lcp`	log(capsular penetration)
`gleason`	Gleason score
`pgg45`	percentage Gleason scores 4 or 5
`lpsa`	log(prostate specific antigen); response variable

a) Fit a linear model to the data set using LASSO regression. Use the function `cv.glmnet` from the R package `glmnet`. Which is the optimal regularization parameter λ? Which coefficients are non-zero in the corresponding model?

b) Fit a linear model using ridge regression. Use again the function `cv.glmnet`, this time with parameter `alpha = 0`. Which is the optimal regularization parameter λ here?

c) Plot the LASSO and ridge traces of the models from a) and b) side-by-side. What are the differences, and where do they come from?

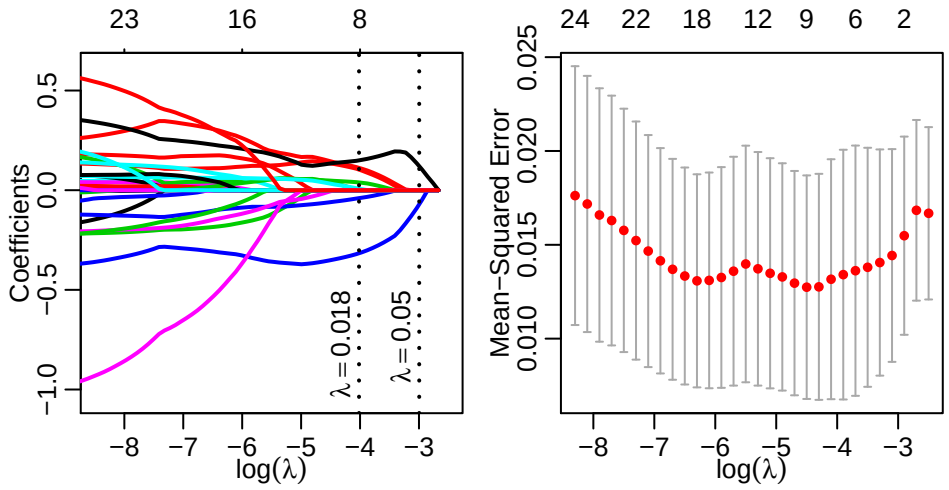
d) The so-called "elastic net" is a regularization approach between ridge and LASSO. With the function `(cv.)glmnet`, it can be used by specifying a parameter $\alpha \in (0, 1)$. Its penalty term is

$$\frac{1-\alpha}{2} \sum_{i=1}^{p} \beta_i^2 + \alpha \sum_{i=1}^{p} |\beta_i|$$

Use a value of $\alpha = \frac{1}{2}$ to fit the `Prostate` data set. Does this approach still do model selection? Compare its estimated mean squared error for the optimal regularization parameter λ to that of the LASSO and ridge fit; is it better or worse?

19. **[Graded]** We again consider the pyrimidine data set presented in the lecture. It contains 74 activitity measurements of the enzyme DHFR in a bacterium in the presence of different pyrimidines characterized by 26 physico-chemical properties. Those properties are quantified by the variables X_1 to X_{26}; the activitity of DHFR is the response variable Y.

The following figure shows the regularization path (left) and the result of a leave-one-out cross validation (right):



a) What is the optimal range for the regularization parameter λ?

☐ $[-5, -3]$ ☐ $[0.0001, 0.001]$ ☐ $[0.005, 0.05]$ ☐ $[0.1, 1]$ ☐ $[3, 5]$

b) When choosing $\lambda = 0.018$ (left dotted line in left plot), we get the following coefficients (rounded):

```
27 x 1 sparse Matrix of class "dgCMatrix"
                    1
(Intercept)  0.600
x1           .
x2           .
x3           .
x4          -0.040
x5           .
x6           .
x7           .
x8           0.103
x9           0.001
x10          .
x11          .
x12          .
x13          .
x14          .
x15          0.039
x16          .
x17          0.008
x18          .
x19          .
x20          0.112
x21          .
x22         -0.316
x23          .
x24          .
x25          0.150
x26          .
```

How many variables are selected in this model? Write down the corresponding model equation.

c) Assume now that we choose $\lambda = 0.05$ (right dotted line in left plot). Which variables remain in the model in this case?

☐ X_4, X_8, X_{22} ☐ X_4, X_{20} ☐ X_{14}, X_{22}, X_{25} ☐ X_{15}, X_{17}, X_{22} ☐ X_{22}, X_{25}

20. **[Graded]** We again look at the prostate data set from exercise 18 (see there for a description). The data set is included in the R package `lasso2` and can be loaded as follows:

```
> data(Prostate, package = "lasso2")
```

With this exercise, you can deepen your knowledge about cross validation as well as improve your R programming skills...

a) Fit a linear model using LASSO (with the `glmnet` package) as in exercise 18.a). Determine the "optimal" regularization parameter in advance. Estimate the generalization performance measured by the mean squared error (MSE) for this value of λ "by hand" (i.e., writing an own R function, not relying on the estimates from `cv.glmnet`). Do you get the same values as `cv.glmnet`?

b) Fit a linear model performing forward selection, using the function `stepAIC` from the `MASS` package. Estimate the MSE also for the model you get from `stepAIC`. How does it perform compared to the LASSO estimate?

c) Finally, fit a linear model *allowing for interaction terms* using LASSO (again using `glmnet`). To do so, you have to specify the design matrix appropriately; use the R function `model.matrix` for that. Again, calculate the MSE for this model. Is it better than the one from a)?
Note: by using a LASSO estimator for a model with interaction terms, it can happen that you get a model with an interaction term but without corresponding main effects. This is a bit strange, but still OK here since we do not try to assign p-values to any of the model terms.

21. **[Graded]** In this exercise, we calculate the Bayes classifier for one explanatory variable (feature) under the assumption that the class-conditional probability densities are normal.

Assume a fish packing plant wants to predict the species of fishes based only on their lightness (see introductory example from the lecture). We want to distinguish between two fish species, sea brass (class 0) and salmon (class 1). Assume the class-conditional probability densities for the length x of the fishes are given by the following normal distributions:

$$X|Y=0 \sim \mathcal{N}(\mu=7, \sigma^2=1) \ ,$$
$$X|Y=1 \sim \mathcal{N}(\mu=3, \sigma^2=1) \ .$$

As a reminder, the density of the normal distribution is given by

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \ .$$

a) Assume that the prior class probabilities are equal for salmon and sea brass: $P[Y=0] = P[Y=1] = \frac{1}{2}$. How does the Bayes classifier look like?

b) Assume now that the fish packing plant processes two times as many sea brass as salmons: $P[Y=0] = \frac{2}{3}, P[Y=1] = \frac{1}{3}$. How does the Bayes classifier look like in this case?

c) Why is the decision threshold in a) and b) not the same? Explain in words or with a figure, but without calculations.

22. **[Graded]** The data set `baby.dat` (available on ILIAS) contains data from a study in which clinicians measured several clinical variables of premature babies; the response variable, `Survival`, notes whether the babies survived (`Survival = 1`) or not (`Survival = 0`).

a) Fit a logistic regression model to the data, using all explanatory variables. Indicate the misclassification rate on the data set.

b) Start with the full model you got in a) and eliminate variables using backward selection. For models fitted with `glm`, you can also use the function `stepAIC` from the `MASS` package. Which model do you end up with? What's its misclassification rate?

c) Estimate the *expected* misclassification rate (for babies that were not in the study) for the model you get in b) using leave-one-out cross validation. In contrast to exercise 18, our error here is not calculated as the square between predicted and measured response variable, but the 0-1-classification error: we count an error of 0 for a correctly classified sample, and and error of 1 for a wrongly classified sample.

d) Comment on the different misclassification rates (or expected misclassification rates) you get in tasks a) to c).

23. [**Graded**] The aim of this exercise is to implement an R function for k-fold cross validation for classification algorithms. k-fold cross validation works as follows:

1. Randomly partition the data set into k subsets of (almost) equal size
2. For each subset $j = 1, \ldots, k$, do:
   a) Remove the j-th subset from the data set
   b) Train a classifier on the data of the remaining $k - 1$ subsets
   c) Predict the class labels y_i of the data points from the removed subset
   d) Calculate the misclassification rate ε_j on the removed subset
3. Report the average of the k misclassification rates: $\frac{1}{k} \sum_{j=1}^{k} \varepsilon_j$

You are completely free to design your implementation, but if you like, you can use the following function skeletons:

```
> kfoldcv <- function(formula, data, k, classifier)
 {
   ## Your code here
 }
```

`formula` and `data` should have the usual meaning; `classifier` is meant to be a function taking `formula`, `data`, `training` and `test` as arguments, fitting a classifier on the training data (corresponding row indices specified by `training`) and giving the predicted class labels for the test data (corresponding row indices specified by `test`) as an output vector. As an example, an implementation for logistic regression is given below:

```
> logRegClassifier <- function(formula, data, training, test)
 {
   # Fit the classifier on the training data and get the index of the predicted class label
   fit <- glm(formula, data[training, ], family = "binomial")
   pred.index <- (predict(fit, newdata = data[test, ], type = "response") >= 0.5) + 1

   # Convert the response index to the correct format: a factor as in the data set
   response <- all.vars(formula)[1] # name of the response variable
   factor(pred.index, levels = 1:2, labels = levels(data[[response]]))
 }
```

Test your approach for k-fold cross validation (i.e. your function `kfoldcv` or similar) by estimating the expected misclassification rate on the babies survival data set (available on ILIAS, see exercise 22) with 5- as well as with 10-fold cross validation. Can you also estimate the expected misclassification rate via *leave-one-out cross validation* with your implementation of `kfoldcv`?

24. [**Graded**] The `BreastCancer` data set contains clinical data of breast cancer tissue of 699 women. Our goal is to predict whether the breast cancer is "benign" or "malignant" (variable `Class`) based on the clinical variables.

The data set is available in the R package `mlbench` and can be loaded as follows:

```
> library(mlbench)
> data(BreastCancer)
```

For all cross validation tasks in this exercise, use your function from exercise 23. If you did not succeed in that exercise, you can instead use leave-one-out cross validation (which may take some time due to the size of the data set), or use the functions from the solutions of exercise 23.

a) Fit an LDA model to the data set using the function `lda` from the R package `MASS`. Do you think the LDA model is appropriate for this data set?

b) Estimate the expected misclassification rate of your model from a) using 10-fold cross validation. You can work with the framework you prepared in exercise 23 and add a new classifier function for LDA, or you can use the function `LDAClassifier` provided in the R solution for exercise 23.

c) Fit a CART model to the data set. Use the R function `rpart` from the homonymous package. Try to let `rpart` overfit the data, then perform manual pruning of the tree using the function `prune`. How many leaves does the resulting decision tree have? How many decisions (= internal nodes) does this correspond to?

d) Estimate the expected misclassification rate of your model from c) using 10-fold cross validation. Which of the models performs better, the one from a) or the one from c)? Is this plausible? You can work with the framework you prepared in exercise 23 and add a new classifier function for CART, or you can use the function `CartClassifier` provided in the R solution for exercise 23.

e) Fit a support vector machine to the data set using the R function `svm` from the package `e1071` with the standard parameters. How many support vectors does the corresponding SVM have? Repeat the fit after standardizing the data beforehand; does this change the number of support vectors?

f) Estimate the expected misclassification rate of your model from e) (fitted on *standardized* data) using 10-fold cross validation. Compare the misclassification rates of the three models from a), c) and e).

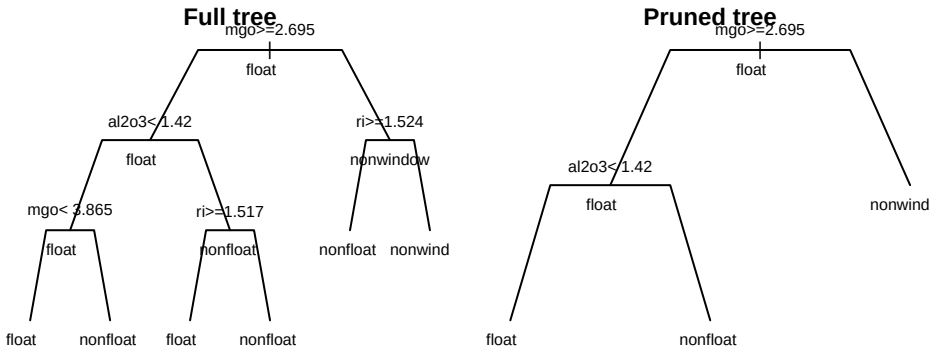
25. [**Graded**] The dataset `glass` lists 9 parameters of different glass types:

`ri`	Refractive index
`na2o`	Sodium oxide (unit measurement: weight percent)
`mgo`	Magnesium oxide
`al2o3`	Aluminium oxide
`sio2`	Silicon oxide
`k2o`	Potassium oxide
`cao`	Calcium oxide
`bao`	Barium oxide
`fe2o3`	Iron oxide

The glass type itself is indicated in the factor `type` which can take one of three values:

`float`	float processed window glass
`nonfloat`	non-float processed window glass
`nonwindow`	non-window glass

a) In the next figure, you see two CART trees fitted to the `glass` data set. The left one is the output from `rpart` when using the default parameters, the right one is a tree we got by manually pruning the tree guided by the plot from `plotcp`.



In a laboratory, you measure the following properties of two glass assays:

	ri	na2o	mgo	al2o3	sio2	k2o	cao	bao	fe2o3
assay 1	1.6	13.0	2.5	1.5	70.0	0.5	8.0	0.2	0.1
assay 2	NA	13.0	3.5	1.5	71.0	NA	NA	NA	0.5

"NA" stands for missing values, i.e. parameters that could not be measured in the laboratory. How would these assays be classified by the full tree and the pruned tree from the plot above? (Possible predictions: `float`, `nonfloat`, `nonwindow`, "no predition possible".)

b) Logistic regression cannot be directly applied to the `glass` dataset with 3 classes. We therefore encode the 3 class problem with 3 binary classification problems:

1. Encode `float` glass with class label $y_i = 1$, `nonfloat` with $y_i = 2$ and `nonwindow` as $y_i = 3$.
2. For $j = 1, 2, 3$, define an indicator variable $Z_{ij} := \begin{cases} 1, & \text{if } Y_i = j, \\ 0, & \text{otherwise} \end{cases}$.
3. Fit a logistic regression model for $Z_{\cdot j}$, $j = 1, 2, 3$, to obtain an estimate $\hat{\pi}_j(x)$ for the probability $\pi_j(x_1, \ldots, x_p) = P[Z_{\cdot j} = 1 \mid X_1 = x_1, \ldots, X_p = x_p]$.
4. Given a sample with features $x_1, \ldots, x_p$, assign it to class $c := \arg\max_j \hat{\pi}_j(x_1, \ldots, x_p)$.

The R output for the three logistic regressions of class j vs. the rest is printed below. Based on these outputs, which glass type would the two assays from a) be assigned to?

```
Class 1 vs. the rest:
=====================
Call:
glm(formula = (glass$type == levels(glass$type)[j]) ~ ., family = binomial,
    data = glass)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.04388  -0.61645  -0.01345   0.69724   1.83599

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  444.2244   369.9216   1.201  0.22981
ri          -524.3150   224.3986  -2.337  0.01946 *
na2o           3.1564     1.8378   1.718  0.08589 .
mgo            6.0824     1.9791   3.073  0.00212 **
al2o3         -0.5402     2.0960  -0.258  0.79661
sio2           3.3943     1.8270   1.858  0.06319 .
k2o            2.9829     2.4032   1.241  0.21452
cao            4.8417     1.9712   2.456  0.01404 *
bao            4.5398     2.6547   1.710  0.08725 .
fe2o3         -2.3060     2.0216  -1.141  0.25400
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 289.15  on 213  degrees of freedom
Residual deviance: 168.52  on 204  degrees of freedom
AIC: 188.52

Number of Fisher Scoring iterations: 8


Class 2 vs. the rest:
=====================
Call:
glm(formula = (glass$type == levels(glass$type)[j]) ~ ., family = binomial,
    data = glass)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8938  -0.8942  -0.5029   1.0852   2.1739

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 413.1911   296.5554   1.393 0.163529
ri           84.7017   166.5029   0.509 0.610956
na2o         -5.8448     1.7844  -3.275 0.001055 **
mgo          -4.9395     1.8504  -2.669 0.007598 **
```

```
al2o3        -3.9722     1.9201  -2.069 0.038569 *
sio2         -5.4353     1.7995  -3.020 0.002524 **
k2o          -5.6850     1.8745  -3.033 0.002422 **
cao          -5.1840     1.8825  -2.754 0.005891 **
bao          -6.4489     1.9072  -3.381 0.000721 ***
fe2o3         0.3689     1.7487   0.211 0.832942
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 278.44  on 213  degrees of freedom
Residual deviance: 236.96  on 204  degrees of freedom
AIC: 256.96

Number of Fisher Scoring iterations: 5


Class 3 vs. the rest:
=====================
Call:
glm(formula = (glass$type == levels(glass$type)[j]) ~ ., family = binomial,
    data = glass)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.63400  -0.12396  -0.03612  -0.00101   2.30170

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3292.972   1102.056  -2.988  0.00281 **
ri           1232.505    444.483   2.773  0.00556 **
na2o           12.635      6.120   2.065  0.03897 *
mgo             8.334      5.802   1.437  0.15086
al2o3          19.869      7.060   2.814  0.00489 **
sio2           15.269      6.688   2.283  0.02242 *
k2o            10.987      6.306   1.742  0.08146 .
cao             9.157      5.867   1.561  0.11859
bao            11.206      6.106   1.835  0.06646 .
fe2o3          -6.654      5.819  -1.143  0.25285
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 235.029  on 213  degrees of freedom
Residual deviance:  44.079  on 204  degrees of freedom
AIC: 64.079

Number of Fisher Scoring iterations: 9
```