Compulsory exercise 2

TMA4268 Statistical Learning V2019

Michail Spitieris, Andreas Strand and Mette Langaas

Deadline: April 12 at 16 on Bb

Contents

Introduction **1**
Supervision . . . 1
Practical issues . . . 1
R packages . . . 2

Problem 1: Regression [6 points] **2**
a) Know your data! . . . 3
b) Modelling `logprice` and `carat` . . . 3
c) Best subset selection . . . 3
d) Variable selection with the lasso . . . 4
e) Regression tree . . . 4
f) Random forest . . . 4
g) Comparing results . . . 4

Problem 2: Unsupervised learning [3 points] **4**
a) Understanding PCA . . . 5
b) Hierarchical clustering . . . 6

Problem 3: Flying solo with diabetes data [6 points] **7**

References **8**

(Last changes: 13.03 first version)

Introduction

Maximal score is 15 points. You need a score of 6/15 for the exercise to be approved. Your score will make up 15% points of your final grade.

Supervision

All supervision is in Smia.

- Thursdays 16.15-18, in addition
- Monday April 1 at 8.15-10
- Thursday April 4 at 14.15-16

Practical issues

- Maximal group size is 3 - join a group (self enroll) before handing in on Bb.
- Remember to write your names and group number on top of your submission.

- The exercise should be handed in as one R Markdown file and a pdf-compiled version of the R Markdown file (if you are not able to produce a pdf-file directly please make an html-file, open it in an internet 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.
- In *all* 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: https://www.math.ntnu.no/emner/TMA4268/2019v/CompEx2mal.Rmd. NB: remember that we can't write full latex in R Markdown, and using `$` to start and stop math mode, or `$$` to start and end equation mode should be sufficient for what you need.
- Please save us time and NOT submit word or zip - or only Rmd - that only result in us contacting you to ask to upload correct file formats.

R packages

You need to install the following packages in R to run the code in this file.

```
install.packages("knitr") #probably already installed
install.packages("rmarkdown") #probably already installed
install.packages("bestglm")# for subset selection with categorical variables
install.packages("glmnet")# for lasso
install.packages("tree") #tree
install.packages("randomForest") #for random forest
install.packages("ElemStatLearn") #dataset in Problem 2
BiocManager::install(c("pheatmap")) #heatmap in Problem 2
```

The following are packages that might be useful.

```
install.packages("ggplot2")
install.packages("GGally") # for ggpairs
install.packages("caret") #for confusion matrices
install.packages("pROC") #for ROC curves
install.packages("e1071") # for support vector machines
install.packages("nnet") # for feed forward neural networks
```

Problem 1: Regression [6 points]

In this problem we will work with understanding and applying different regression methods, and will use the `diamonds` data set from the `ggplot2` package. The `diamonds` data set consists of

- `price` (in US dollars) - which will be our response,

and quality information (9 covariates) for around 54000 diamonds. According to Wickham (2016) (Section 3.10) there are four C's of diamond quality:

- `carat` (weight),
- `cut` (quality of the cut: Fair/Good/VeryGood/Premium/Ideal),
- `colour` (from worst J to best D) and
- `clarity` (from worst to best: I1, SI1, SI2, VS1, VVS1, VVS2, IF).

In addition there are five physical measurements:

- `depth` (total depth percentage, calculated from x, y and z),
- `table` (width of top of diamond relative to widest point),
- `xx` (length in mm),

- `yy` (width in mm) and
- `zz` (depth in mm)

explained with a drawing in Figure 3.1 of the Wickham (2016) book (ebook available for free for NTNU students). There are no missing data. In addition we have made two additional transformed variables:

- `logprice`: logarithm (base 10) of the price, and
- `logcarat`: logarithm (base 10) of the carat.

To be able to do the analyses without too much hassle (running time for analyses and plotting) we constructed a smaller version of the data set, and divided it into a training set **`dtrain`** and a test set **`dtest`**, to be loaded from the course web-page as follows.

```
all=dget("https://www.math.ntnu.no/emner/TMA4268/2019v/data/diamond.dd")
dtrain=all$dtrain
dtest=all$dtest
```

Our aim will be to predict the price of the diamond, based on some or all of the covariates. We would also like to understand which of the covariates are important in predicting the price and how the covariates are related to the price.

a) Know your data!

Q1: Would you choose `price` or `logprice` as response variable? Justify your choice. Next, plot your choice of response pairwise with `carat`, `logcarat`, `color`, `clarity` and `cut`. Comment.

b) Modelling `logprice` and `carat`

Use the local regression model $\texttt{logprice} = \beta_0 + \beta_1 \texttt{carat} + \beta_2 \texttt{carat}^2$ weighted by the tricube kernel K_{i0}.

Q2: What is the predicted price of a diamond weighting 1 carat. Use the closest 20% of the observations.
Q3: What choice of β_1, β_2 and K_{i0} would result in KNN-regression?

c) Best subset selection

Q4: Describe how you can perform model selection in regression with AIC as criterion.
Q5: What are the main differences between using AIC for model selection and using cross-validation (with mean squared test error MSE)?
Q6: See the code below for performing model selection with `bestglm()` based on AIC. What kind of contrast is used to represent `cut`, `color` and `clarity`? Write down the final best model and explain what you can interpret from the model.
Q7: Calculate and report the MSE of the test set using the best model (on the scale of `logprice`).

```
library(bestglm)
ds=as.data.frame(within(dtrain,{
  y=logprice    # setting reponse
  logprice=NULL # not include as covariate
  price=NULL    # not include as covariate
  carat=NULL    # not include as covariate
  }))
fit=bestglm(Xy=ds, IC="AIC")$BestModel
summary(fit)
```

d) Variable selection with the lasso

Q8: Build a model matrix for the covariates `~logcarat+cut+clarity+color+depth+table+xx+yy+zz-1`. What is the dimension of this matrix?
Q9: Fit a lasso regression to the diamond data with `logprice` as the response and the model matrix given in Q8. How did you find the value to be used for the regularization parameter?
Q10: Calculate and report the MSE of the test set (on the scale of `logprice`).

e) Regression tree

Q11: A regression tree to model is built using a *greedy* approach. What does that mean? Explain the strategy used for constructing a regression tree.
Q12: Is a regression tree a suitable method to handle both numerical and categorical covariates? Elaborate.
Q13: Fit a (full) regression tree to the diamond data with `logprice` as the response (and the same covariates as for c and d), and plot the result. Comment briefly on you findings.
Q14: Calculate and report the MSE of the test set (on the scale of `logprice`).

f) Random forest

Q15: Explain the motivation behind bagging, and how bagging differs from random forest? What is the role of bootstrapping?
Q16: What are the parameter(s) to be set in random forest, and what are the rules for setting these?
Q17: Boosting is a popular method. What is the main difference between random forest and boosting?
Q18: Fit a random forest to the diamond data with `logprice` as the response (and the same covariates as before). Comment on your choice of parameter (as decribed in Q16).
Q19: Make a variable importance plot and comment on the plot. Calculate and report the MSE of the test set (on the scale of `logprice`).

g) Comparing results

Q20: Finally, compare the results from c (subset selection), d (lasso), e (tree) and f (random forest): Which method has given the best performance on the test set and which method has given you the best insight into the relationship between the price and the covariates?

Problem 2: Unsupervised learning [3 points]

We will use a data set called `nci` from the Friedman, Hastie, and Tibshirani (2001) book. The data set consists of measurement on the activity of $p = 6830$ genes for $n = 64$ samples from cancer tumors. Our aim here is to use unsupervised methods to study the relationship among the cancer samples. The data is on a transformed scale, where a negative number means that the gene has a lower regulation than some reference sample and a positive number gives a higer regularization. The following names are given for the cancer tumors (R print-out). It is known that K562 cells are leukemia cells and MCF7 are breast cancer cells. Observe the one sample labelled as `UNKNOWN`. More information: http://genome-www.stanford.edu/nci60/

Let $\mathbf{X}$ be a $n \times p$ matrix of these observations, and $\mathbf{Z}$ a standardized version, where column (gene) means are 0 and column standard deviations are 1. Further, let $\mathbf{R}$ be the $p \times p$ covariance matrix of the gene-standardized observations, which can be estimated as

$$\hat{\mathbf{R}} = \frac{1}{n-1} \sum_{i=1}^{n} (\mathbf{Z}_i - \bar{\mathbf{Z}})(\mathbf{Z}_i - \bar{\mathbf{Z}})^T$$

where $\bar{\mathbf{Z}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Z}_i$.

Studying the relationship between the different tumor samples is hard due the the high dimension of the gene data, and we know that the gene expressions of different genes may be correlated. One way to proceed is to calculate *principal components*, which we find using the eigenvectors of the covariance matrix $\hat{\mathbf{R}}$.

a) Understanding PCA

Q21: What is the definition of a principal component score, and how is the score related to the eigenvectors of the matrix $\hat{\mathbf{R}}$.

Study the R-code below. The eigenvectors of $\hat{\mathbf{R}}$ are found in `pca$rotation` and the principal component scores in `pca$x`.

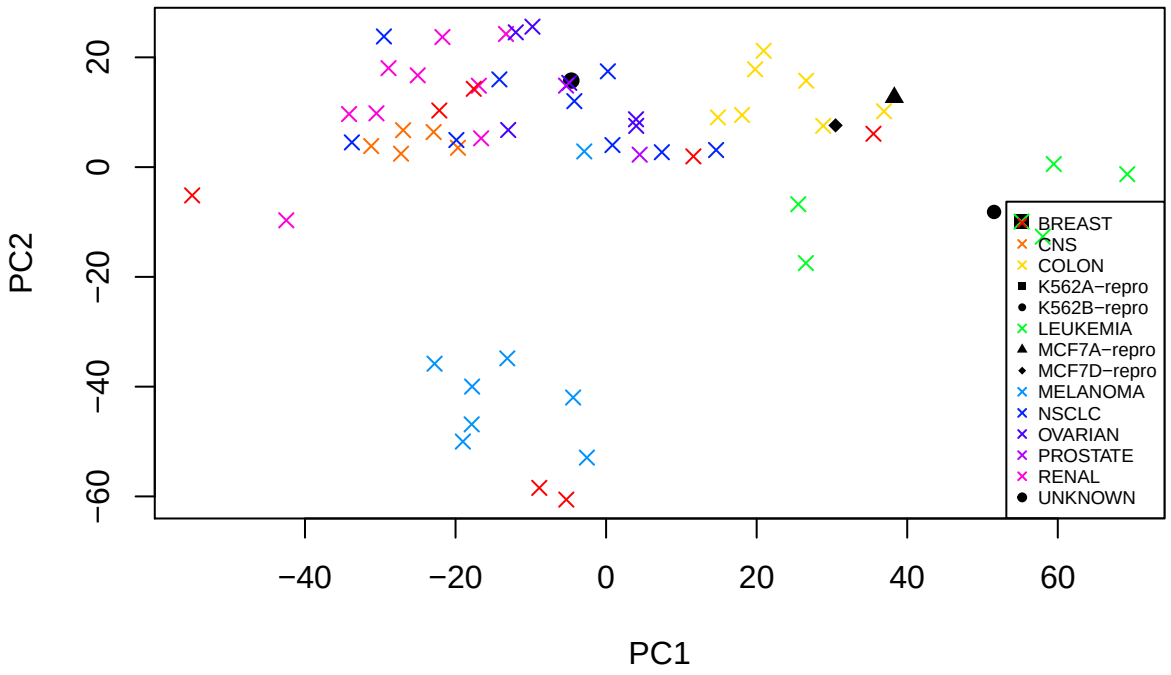
Q22: Explain what is given in the plot with title "First eigenvector". Why are there only $n = 64$ eigenvectors and $n = 64$ principal component scores?

Q23: How many principal components are needed to explain 80% of the total variance in $\mathbf{Z}$? Why is `sum(pca$sdev^2)=p`?

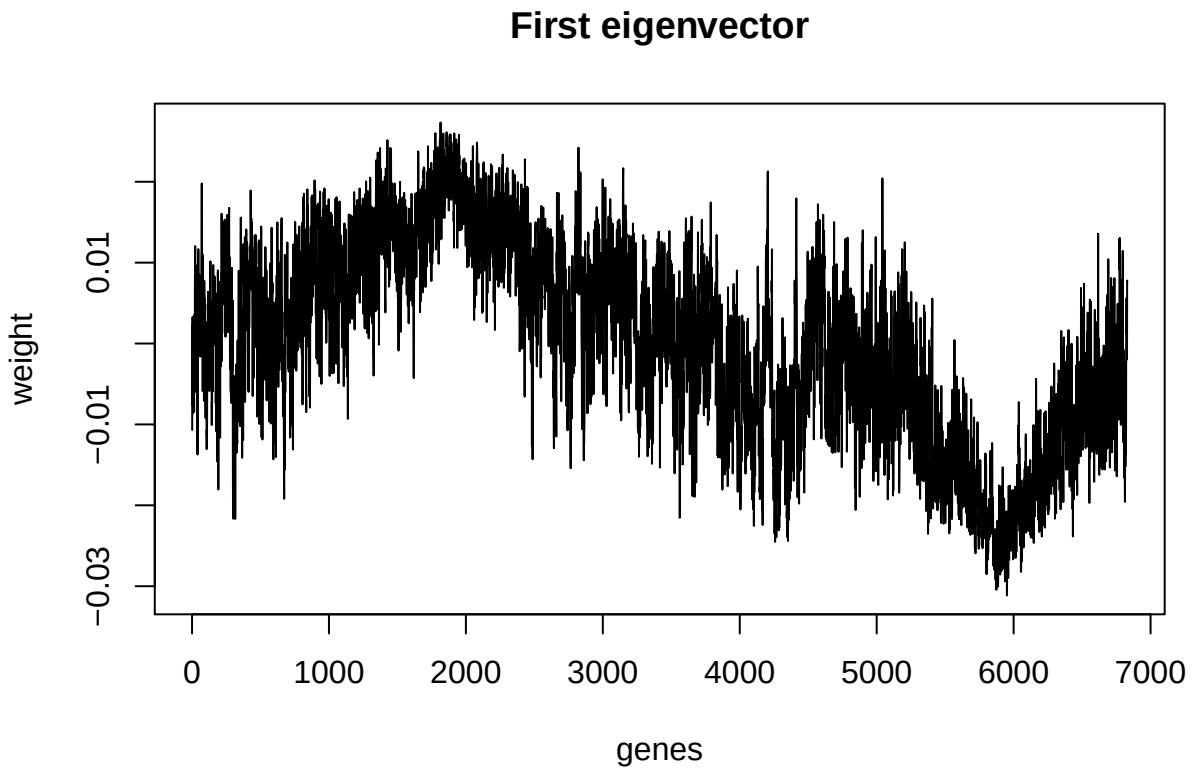
Hint: Total variance is given as the trace of $\hat{\mathbf{R}}$, the trace of a square matrix is the sum of the diagonal elements.

Q24: Study the PC1 vs PC2 plot, and comment on the groupings observed. What can you say about the placement of the `K262`, `MCF7` and `UNKNOWN` samples? Produce the same plot for two other pairs of PCs and comment on your observations.

```
library(ElemStatLearn)
X=t(nci) #n times p matrix
ngroups=length(table(rownames(X)))
cols=rainbow(ngroups)
cols[c(4,5,7,8,14)] = "black"
pch.vec = rep(4,14)
pch.vec[c(4,5,7,8,14)] = 15:19

colsvsnames=cbind(cols,sort(unique(rownames(X))))
colsamples=cols[match(rownames(X),colsvsnames[,2])]
pchvsnames=cbind(pch.vec,sort(unique(rownames(X))))
pchsamples=pch.vec[match(rownames(X),pchvsnames[,2])]

Z=scale(X)

pca=prcomp(Z)
plot(pca$x[,1],pca$x[,2],xlab="PC1",ylab="PC2",pch=pchsamples,col=colsamples)
legend("bottomright",legend = colsvsnames[,2],cex=0.55,col=cols,pch = pch.vec)
```




```
plot(1:dim(X)[2],pca$rotation[,1],type="l",xlab="genes",ylab="weight",main="First eigenvector")
```



b) Hierarchical clustering

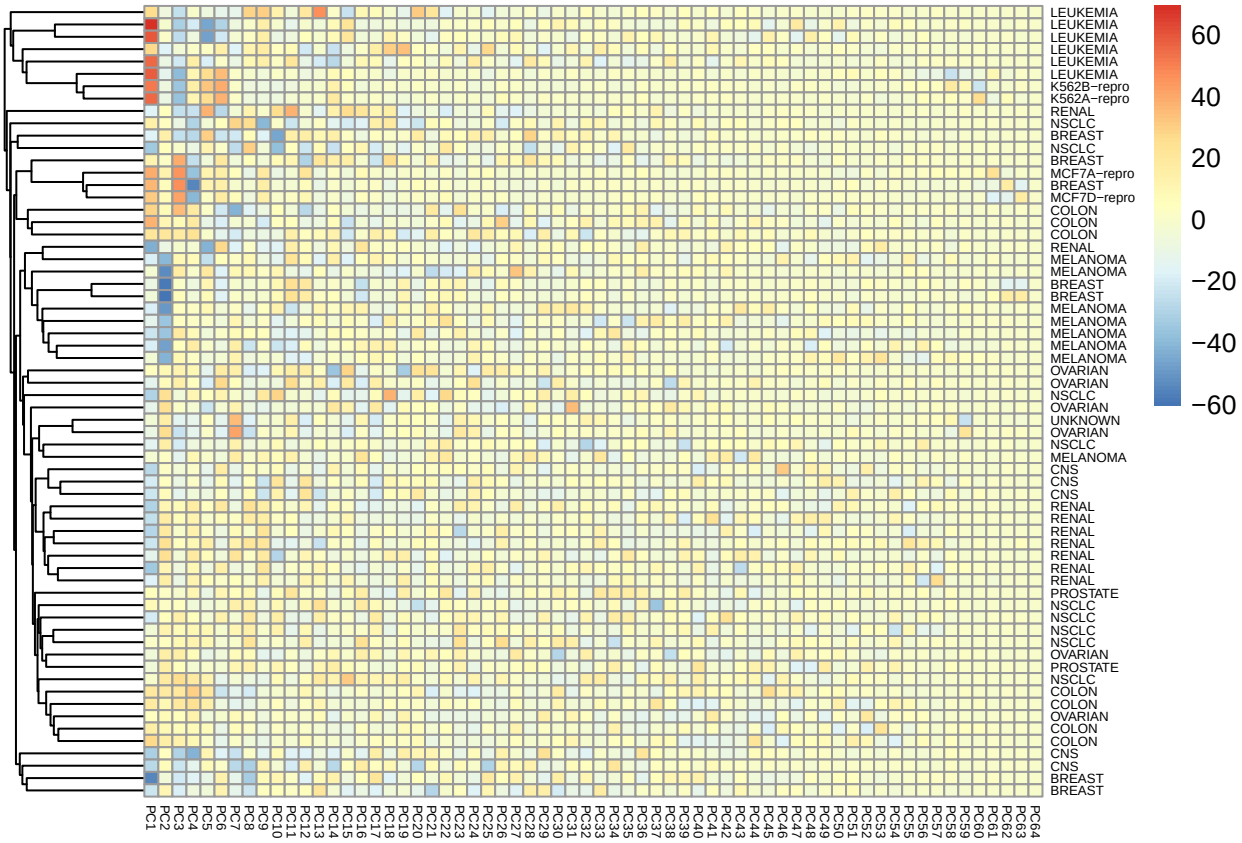
Hierarchical clustering can be used to group the cancer samples, and then we may see which samples that are most similar to the five samples named `K262`, `MCF7` and `UNKNOWN`.

Q25:: Explain what it means to use Euclidean distance and average linkage for hierarchical clustering.

Q26:: Perform hierarchical clustering with Euclidean distance and average linkage on the scaled gene expression in `Z`. Observe where our samples labelled as K562, MCF7 and `UNKNOWN` are placed in the dendrogram. Which conclusions can you draw from this?

Q27:: Study the R-code and plot below. Here we have performed hierarchical clustering based on thefirst 64 principal component instead of the gene expression data in `Z`. What is the difference between using all the gene expression data and using the first 64 principal components in the clustering? We have plotted the dendrogram together with a heatmap of the data. Explain what is shown in the heatmap. What is given on the horizontal axis, vertical axis, value in the pixel grid?

```
library(pheatmap)
npcs=64
pheatmap(pca$x[,1:npcs],scale="none",cluster_col=FALSE,cluster_row=TRUE,clustering_distance_rows = "eucl
```



Problem 3: Flying solo with diabetes data [6 points]

A learning objective in this course is: *The student knows, based on an existing data set, how to choose a suitable statistical model, apply sound statistical methods, and perform the analyses using statistical software. The student knows how to present the results from the statistical analyses, and which conclusions can be drawn from the analyses.*

This problem addresses that objective.

We will use the classical data set of *diabetes* from a population of women of Pima Indian heritage in the US, available in the R `MASS` package. The following information is available for each woman:

- diabetes: `0`= not present, `1`= present
- npreg: number of pregnancies

- glu: plasma glucose concentration in an oral glucose tolerance test
- bp: diastolic blood pressure (mmHg)
- skin: triceps skin fold thickness (mm)
- bmi: body mass index (weight in kg/(height in m)^2)
- ped: diabetes pedigree function.
- age: age in years

We will use a training set (called `ctrain`) with 300 observations (200 non-diabetes and 100 diabetes cases) and a test set (called `ctest`) with 232 observations (155 non-diabetes and 77 diabetes cases). Our aim is to make a classification rule for diabetes (or not) based on the available data.

```
flying=dget("https://www.math.ntnu.no/emner/TMA4268/2019v/data/flying.dd")
ctrain=flying$ctrain
ctest=flying$ctest
```

Q28: Start by getting to know the *training data*, by producing summaries and plots. Write a few sentences about what you observe and include your top 3 informative plots and/or outputs.

Q29: Use different methods to analyse the data. In particular use

- one method from Module 4: Classification
- one method from Module 8: Trees (and forests)
- one method from Module 9: Support vector machines and, finally
- one method from Module 11: Neural networks

For each method you

- clearly write out the model and model assumptions for the method
- explain how any tuning parameters are chosen or model selection is performed
- report (any) insight into the interpretation of the fitted model
- evaluate the model using the test data, and report misclassifiation rate (cut-off 0.5 on probability) and plot ROC-curves and give the AUC (for method where class probabilities are given).

Q30: Conclude with choosing a winning method, and explain why you mean that this is the winning method.

References

Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. 2001. *The Elements of Statistical Learning*. Vol. 1. Springer series in statistics New York.

Wickham, Hadley. 2016. *Ggplot2 Elegant Graphics for Data Analysis*. Springer.