Compulsory Exercise 2

TMA4268 Statistical Learning V2020

Martina Hall, Michail Spitieris, Stefanie Muff, Department of Mathematical Sciences, NTNU

Hand out date: March 16, 2020

Last changes: 16.03.2020

The submission deadline is Thursday, 2. April 2020, 23:59h using Blackboard

Introduction

Maximal score is 50 points. You need a score of 20/50 for the exercise to be approved. Your score will make up 10% points of your final grade.

Supervision

Due to the extraordinary situation, there will be no in-person supervision. Please use the discussion board, which you can access from our course site on Blackboard, then choose "Course Tools" -> "Discussion Board" -> "194_TMA4268_1_2020_V_1" and then choose the "General Questions and Answers" Forum.

Practical issues

- **You work on the same group as for compulsory exercise 1.**
- **However, please do reduce personal contact with your group members to a minimum. Use digital channels wherever possible.**
- The submission guidelines are the same as in compulsory 1.
- 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 your browser and save as pdf - no, not landscape - but portrait please). We will read the pdf-file and use the Rmd file in case we need to check details in your submission.
- In the R-chunks please use both `echo=TRUE` and `eval=TRUE` to make it simpler for us to read and grade.
- Please do not include all the text from this file (that you are reading now) - we want your R code, plots and written solutions - use the same template as for compulsory 1 (https://wiki.math.ntnu.no/tma4268/2020v/subpage6).
- Please not more than 12 pages in your pdf-file! (This is a request, not a requirement.)
- Please save us time and do NOT submit word or zip, and do not submit only the Rmd. This only results in extra work for us!

R packages

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

```
install.packages("knitr")  #probably already installed
install.packages("rmarkdown")  #probably already installed
install.packages("ggplot2")  #plotting with ggplot
install.packages("ISLR")
install.packages("MASS")
install.packages("GGally")
install.packages("glmnet")
install.packages("e1071")
install.packages("tree")
install.packages("leaps")
install.packages("randomForest")
install.packages("gbm")
install.packages("ggfortify")
```

Multiple choice problems

Some of the problems are *multiple choice questions*. This is how these will be graded:

Maximum is 2 points. There are four choices, and each of them can be TRUE or FALSE. If you make one mistake (either wrongly mark an option as TRUE/FALSE) you get 1P, if you have two or more mistakes, you get 0P. Your answer should be given as a list of answers, like TRUE, TRUE, FALSE, FALSE, for example.

Problem 1 (10p)

a) Ridge Regression (2p)

Show that the ridge regression estimator is $\hat{\beta}_{Ridge} = (X^T X + \lambda I)^{-1} X^T y$.

b) (2p)

Find the expected value and the variance-covariance matrix of $\hat{\beta}_{Ridge}$ (1P each).

c) (2P) - Multiple choice

Which of the following statements are true, which false?

(i) For a problem that $p > n$ we can use forward selection but not backward.

(ii) Ridge Regression exploits the trade-off between variance and bias in the MSE, but Lasso doesn't.

(iii) Ridge and Lasso perform variable selection.

(iv) As the tuning parameter λ in ridge regression increases, the variance decreases and the bias increases.

d) (2p)

For the following regression tasks in problems 1–3 we will use the `College` data from the `ISLR` package, which consists of 18 variables and 777 observations (for more details see `?College`). First we split into training and testing sets (50% of the data in each) using the following code:

```
library(ISLR)
set.seed(1)
train.ind = sample(1:nrow(College), 0.5 * nrow(College))
college.train = College[train.ind, ]
college.test = College[-train.ind, ]
```

```
str(College)
```

Using the out-of-state tuition (variable `Outstate`) as response, apply forward selection in order to identify a satisfactory model that uses only a subset of all the variables in the dataset in order to predict the response. Choose a model according to one of the criteria that you know and briefly say why. Write down the model and report the MSE on the test set.

e) (2p)

Now do model selection using the same dataset as in (d) using the Lasso method. How did you select the tuning parameter λ? Report the set of variables that was selected and the MSE on the test set.

Problem 2 (9p)

a) (2p) - Multiple choice

Which of the following statements are true, which false?

(i) A regression spline of order 3 with 4 knots has 8 basis functions.
(ii) A regression spline with polynomials of degree $M-1$ has continous derivatives up to order $M-2$, but not at the knots.
(iii) A natural cubic spline is linear beyond the boundary knots.
(iv) A smoothing spline is (a shrunken version of) a natural cubic spline with knots at the values of all data points x_i for $i = 1, \ldots, n$.

b) (2p)

Write down the basis functions for a cubic spline with knots at the quartiles q_1, q_2, q_3 of variable X.

c) (2p)

We continue with using the `College` dataset that we used in problem 1. Investigate the relationships between `Outstate` and the following 6 predictors (using the training dataset `college.train`):

```
## [1] "Private"     "Room.Board"  "Terminal"    "perc.alumni" "Expend"
## [6] "Grad.Rate"
```

Create some informative plots and say which of the variables seem to have a linear relationship with the response, and which might benefit from a non-linear transformation (like e.g. a spline).

d) (3P)

(i) Fit polynomial regression models for `Outstate` with `Terminal` as the only covariate for a range of polynomial degrees ($d = 1, \ldots, 10$) and plot the results. Use the training data (`college.train`) for this task.
(ii) Still for the training data, choose a suitable smoothing spline model to predict `Outstate` as a function of `Expend` (again as the only covariate) and plot the fitted function into the scatterplot of `Outstate` against `Expend`. How did you choose the degrees of freedom?
(iii) Report the corresponding training MSE for (i) and (ii). Did you expect that?

Problem 3 (9p)

a) (2P) - Multiple choice

Which of the following statements are true, which false?

(i) Regression trees cannot handle categorical predictors.
(ii) Regression and classification trees are easy to interpret.
(iii) The random forest approaches improves bagging, because it reduces the variance of the predictor function by decorrelating the trees.
(iv) The number of trees B in bagging and random forests is a tuning parameter.

b) (4P)

Select one method from Module 8 (tree-based methods) in order to build a good model to predict `Outstate` in the `College` dataset that we used in problems 1 and 2. Explain your choice (pros/cons?) and how you chose the tuning parameter(s). Train the model using the training data and report the MSE for the test data.

c) (2p)

Compare the results (tests MSEs) among all the methods you used in Problems 1-3. Which method perform best in terms of prediction error? Which method would you choose if the aim is to develop an interpretable model?

Problem 4 (12P)

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 `d.train`) with 300 observations (200 non-diabetes and 100 diabetes cases) and a test set (called `d.test`) with 232 observations (155 non-diabetes and 77 diabetes cases). Our aim is to

make a classification rule for the presence of diabetes (yes/no) based on the available data. You can load the data as follows:

```
id <- "1Fv6xwKLSZHldRAC1MrcK2mzdOYnbgv0E"  # google file ID
d.diabetes <- dget(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id))
d.train = d.diabetes$ctrain
d.test = d.diabetes$ctest
```

a) (2P) - Multiple choice

Start by getting to know the *training data*, by producing summaries and plots. Which of the following statements are true, which false?

(i) Females with high glucose levels and higher bmi seem to have a higher risk for diabetes.
(ii) Some women had up to 17 pregnancies.
(iii) BMI and triceps skin fold thickness seem to be positively correlated.
(iv) The distribution of the number of pregnancies per woman seems to be a bit skewed and a transformation of this variable could therefore be appropriate.

b) (4P)

Fit a support vector classifier (linear boundary) and a support vector machine (radial boundary) to find good functions that predict the diabetes status of a patient. Use cross-validation to find a good cost parameter (for the linear boundary) and a good combination of cost *and* γ parameters (for the radial boundary). Report the confusion tables and misclassification error rates for the test set in both cases. Which classifier do you prefer and why? (Do not use any variable transformations or standardizations to facilitate correction).

R-hints: The response variable must be converted into a factor variable before you continue.

```
d.train$diabetes <- as.factor(d.train$diabetes)
d.test$diabetes <- as.factor(d.test$diabetes)
```

To run cross-validation over a grid of two tuning parameters, you can use the `tune()` function where `ranges` defines the grid points as follows:

```
tune(..., formula, kernel = ..., ranges = list(cost = c(...), gamma = c(...)))
```

c) (2P)

Compare the performance of the two classifiers from b) to *one other classification method* that you have learned about in the course. Explain your choice and report the confusion table and misclassification error rate on the test set for your chosen method and interpret what you see. What are advantages/disadvantages of your chosen method with respect to SVMs?

d) (2P) - Multiple choice

Which of the following statements are true, which false?

(i) Under standard conditions, the maximal margin hyperplane approach is equivalent to a linear discriminant analysis.
(ii) Under standard conditions, the support vector classifier is equivalent to quadratic discriminant analysis.
(iii) Logistic regression, LDA and support vector machines tend to perform similar when decision boundaries are linear, unless classes are linearly separable.

(iv) An advantage of logistic regression over SVMs is that it is easier to do feature selection and to interpret the results.

e) (2P) Link to logistic regression and hinge loss.

Look at slides 71-73 of Module 9. Show that the loss function

$$\log(1 + \exp(-y_i f(\boldsymbol{x}_i)))$$

is the deviance for the $y = -1, 1$ encoding in a logistic regression model.
Hint: $f(\boldsymbol{x}_i)$ correponds to the linear predictor in the logistic regression approach.

Problem 5 (10P)

The following dataset consists of 40 tissue samples with measurements of 1,000 genes. The first 20 tissues come from healthy patients and the remaining 20 come from a diseased patient group. The following code loads the dataset into your session with row names decribing if the tissue comes from a diseased or healthy person.

```
id <- "1VfVCQvWt121UN39NXZ4aR9Dmsbj-p9OU"  # google file ID
GeneData <- read.csv(sprintf("https://docs.google.com/uc?id=%s&export=download",
    id), header = F)
colnames(GeneData)[1:20] = paste(rep("H", 20), c(1:20), sep = "")
colnames(GeneData)[21:40] = paste(rep("D", 20), c(1:20), sep = "")
row.names(GeneData) = paste(rep("G", 1000), c(1:1000), sep = "")
```

a) (2P)

Perform hierarchical clustering with complete, single and average linkage using **both** Euclidean distance and correlation-based distance on the dataset. Plot the dendograms. Hint: You can use `par(mfrow=c(1,3))` to plot all three dendograms on one line or `par(mfrow=c(2,3))` to plot all six together.

b) (2P)

Use these dendograms to cluster the tissues into two groups. Compare the groups with respect to the patient group the tissue comes from. Which linkage and distance measure performs best when we know the true state of the tissue?

c) (1P)

With Principal Component Analysis, the first principal component loading vector solves the following optimization problem,

$$\max_{\phi_{11},\ldots,\phi_{p1}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \quad \text{subject to} \quad \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

Explain what ϕ, p, n and x are in this optimization probelm and write down the formula for the first principal component scores.

d) (2P)

(i) (1P) Use PCA to plot the samples in two dimensions. Color the samples based on the tissues group of patients.
(ii) (1P) How much variance is explained by the first 5 PCs?

e) (1P)

Use your results from PCA to find which genes that vary the most accross the two groups.

f) (2P)

Use K-means to seperate the tissue samples into two groups. Plot the values in a two-dimensional space with PCA. What is the error rate of K-means?