Prof. Dr. Steffen Wagner
Angewandte Statistik
Fachbereich II
Berliner Hochschule für Technik

Exercise 06: Shrinkage Methods

Machine Learning I – SoSe 24

This workshop covers hierarchical clustering and soft clustering. At the end of the worksheet there are a couple of written exercises for you to do at home, which should be good practice for the exam.

1 Preparations

1.1 RStudio Project

1. Open your Machine Learning 1 RStudio Project
2. Create an R Script file to perform this exercise

1.2 Required Packages

For this exercise you require the following additional R packages. Please make sure that you have installed them on your computer before coming to the workshop session for the case that Eduroam is not working.

```
# check if packages can be loaded, i.e. they are already installed
library(glmnet)          # shrinkage regression
library(ISLR2)           # for data sets
```

If you get an error at this stage, you need to install the packages.

2 Ridge and Lasso Regression

2.1 The baseball `Hitters` data

The `Hitters` data set in the package `ISLR2` concerns 322 professional baseball players (MLB). Baseball players are categorised into two main groups "pitchers" and "hitters" and this data-set is restricted to the second group. The aim is to develop a supervised learning model with `salary` as the outcome variable. The resulting model can be used to predict a players salary given a players statistics.[1]

You don't need know much about baseball to analyse these data, but a brief summary of the type of variables is useful.

Variable	Explanation
Salary	Player's annual Salary in 1987. The outcome variable for our supervised learning model
AtBat, Hits, HmRun, Runs, RBI, Walks	Hitting performance statistics in 1986. The larger the number the better.
Years	Number of years in ML Baseball (in 1986)
CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks	Career hitting performance statistics. The larger the number the better.
League, Division, NewLeague	Nominal variables indicating in which league and division the hitter played and if this changed during the 1986 season
PutOuts, Assists, Errors	Fielding performance statistics in 1986. Put-outs and assists are good and number of errors is bad.

Hits and Home runs are/were considered the most important of a hitter's performance statistics.

Now, use Section 6.5.1/6.52 in *James et al*[2] to apply shrinkage regression to the predict hitter's salary.

1. Data preparation: The `Hitters` data contain NA values, which we will remove. Check the beginning of Section 6.5.1 in *James et al* to to find out how to handle this issue.
2. Data exploration: Spend a few minutes getting to know the data. In particular produce a histogram of the outcome variable `Salary` and some scatter plots with `Salary` on the y-axis. As some of the variables have skewed distributions, you might want to plot some axes on a log scale. The `plot()` command takes an argument `log="x"`, `log="y"` or `log="xy"` to do this.
3. modeling: Jump to Section 6.5.2 *Ridge Regression and the Lasso*. A few extra commands are suggested below. Apply them before using the `predict()` function.
   - Plot coefficient estimates for variable `AtBat` as a function of `lambda`:

```
plot(grid, coef(ridge.mod)["AtBat",], log = "x", typl = "l", xlab = "lambda")
```

   - Try this with a few other variables.
   - Obtain a similar plot with all the variables using

[1] In the 80s and 90s several baseball players went to a player's tribunal to argue that they were being underpaid, using such models as evidence. If the subject of data analysis in baseball interests you, a good book to read is *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis.

[2] https://www.statlearning.com/

```
plot(ridge.mod, xvar = "lambda")
```

4. Predictions and Lasso: Carry on working through the remaining part of Section 6.5.2 including the section on the Lasso.

Solution:

R file `Shrinkage_hitters.R` in Moodle
Link: https://lms.bht-berlin.de/mod/resource/view.php?id=1197543

2.2 The `mtcars` data set

Remark: You have come across two similar data sets relating to mechanical details of motor-cars: `datsets::mtcars` and `ISLR2::Auto`.

In many ways the data set `Auto` is better because there are 392 models of car and 9 variables. In `mtcars` there are only 32 models of car and 11 variables. Because ridge regression and the lasso give similar results to linear models when the number of observations is large compared to the number of variables, we will use `datasets::mtcars`. If we use `Auto`the best value for lambda will $\lambda = 0$.

Your task is to repeat the Ridge and Lasso model fitting process introduced with the `Hitters` data set in the previous section.

1. To set up the data use:

```
data("mtcars")
x <- model.matrix(mpg ~ hp + I(hp^2) + cyl + disp + drat + wt + qsec + vs +
                    am + gear + carb,
                  data = mtcars)[, -1]   # Why  [, -1]
y <- mtcars$mpg
grid <- 10^seq(10, -2, length = 100)
```

   We are including `I(hp^2)` (squared horsepower) as we found this to be significant in one of the previous workshops for the `Auto` data.

2. Ignore any warning messages of the form:
   `Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold`

3. Answer the following questions as you work through the exercise.

   a) What is the Test MSE for the full model, and for the null model?
   b) What is the best ridge regression value for λ?
   c) What is the Test MSE for the best lambda model.
   d) What is the best lasso value for λ, and which variables in the best lasso model are non zero?
   e) What is the Test MSE for the best lambda model?

Solution:

R file `Shrinkage_mtcars.R` in Moodle
Link: https://lms.bht-berlin.de/mod/resource/view.php?id=1197542

3 Understanding the concept

3.1 MC questions I

Indicate which of (i) through (iv) is correct. Justify your answer.

a) The lasso, relative to least squares, is:
   i) Fits the data better and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
   ii) Fits the data better and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
   iii) Less over-fitting and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
   iv) Less over-fitting and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.

b) Repeat (a) for ridge regression relative to least squares.

Solution:

Only iii) is true. The least squares model is unbiased but can overfit. By shrinking the coefficients the variance will decrease and the bias increase. This holds for a) and b).

3.2 MC questions II

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\sum_{j=1}^{p}\widehat{\beta}_j x_{ij}\right)^2 \quad \text{subject to} \quad \sum_{j=1}^{p}\widehat{\beta}_j^2 \leq s,$$

for a particular value of s.

For parts a) through e), indicate which of (i) through (v) is correct. Justify your answer.

a) As we increase s from 0 until the least squares estimates are found, the training RSS will:
   i) Increase initially, and then eventually start decreasing in an inverted U shape.
   ii) Decrease initially, and then eventually start increasing in a U shape.
   iii) Steadily increase.
   iv) Steadily decrease.
   v) Remain constant.

b) Repeat (a) for test RSS.

c) Repeat (a) for variance.

d) Repeat (a) for squared bias.

e) Repeat (a) for the variance of the noise term.

Solution:

a) Training RSS will decrease because with largers s the betas can fit the data better. The only correct answer is iv).b) Test RSS will initially decrease as the predictions start to adapt to the observed data, but at some point the model starts overfitting and the Test RSS will start increasing. Answer is ii).c) (iii) the variance increases.d) iv) the squared bias decreases.e) v) remains constant.

3.3 MC questions III

Suppose we estimate the regression coefficients in a linear regression model by minimizing

$$\sum_{i=1}^{n}\left(y_i-\widehat{\beta}_0-\sum_{j=1}^{p}\widehat{\beta}_j x_i\right)^2+\lambda\sum_{j=1}^{p}\widehat{\beta}_j^2$$

for a particular value of λ.

For parts (a) through (e), indicate which of i) through v) is correct. Justify your answer.

a) As we increase λ from 0 until the least squares estimates are found, the training RSS will:
   i) Increase initially, and then eventually start decreasing in an inverted U shape.
   ii) Decrease initially, and then eventually start increasing in a U shape.
   iii) Steadily increase.
   iv) Steadily decrease.
   v) Remain constant.

b) Repeat (a) for test RSS.

c) Repeat (a) for variance.

d) Repeat (a) for squared bias.

e) Repeat (a) for the variance of the noise term.

Solution:

This question is the reverse of the previous question, large λ corresponds to small s. So the answers are: a-iii, b-ii, c -iv, d-iii, e-v.