BEUTH HOCHSCHULE
FÜR TECHNIK
BERLIN
University of Applied Sciences

Machine Learning 1

Data Science

Summer Semester 2018 **Prof. Tim Downie**

Workshop 7
Shrinkage Methods

Start R in the usual way.
You will the package `glmnet`, which you will probably need to install. You will also use the package `ISLR`, but if you are up-to-date with the workshops, this will already be installed.

Ridge regression and the lasso using the baseball `Hitters` data

The `Hitters` data set in the package `ISLR` concerns 322 professional baseball players (MLB). Baseball players are categorised into two main groups "pitchers" and "hitters" and this dataset 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 appropriate.

Variables	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
Years	Number of years in ML Baseball (in 1986)
CAtBat, CHits, CHmRun, CRuns, CRBI, CWalks	Career hitting performance statistics
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

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

[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 statistical analyses baseball interests you then a good starting point is to read the book *Moneyball: The Art of Winning an Unfair Game* by Michael Lewis.

The Hitters data contain NA values, which we will remove.

► Work through the few commands on page 244 of Lab 1 Section 6.5 to remove the NAs.

► 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.

► Work through Section 6.6 Lab 2 pages 251 to page 255. A few extra commands have been suggested below; make a note of where they slot in before you start on this section.

▷ At the end of page 252, after obtaining all the coefficients for certain values of lambda, plot how the AtBat coefficient changes with lambda.

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

and try this with a few other variables. Obtain a similar plot with all the variables using

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

▷ At the top of page 254, the R command has been updated, so the command as given in James et al.

```
> ridge.pred=predict (ridge.mod ,s=0, newx=x[test ,], exact=T)
```

now gives an error. When the argument exact=TRUE is specified, the original values of x and y need to be supplied. The appropriate command is:

```
> ridge.pred=predict(ridge.mod,s=0,newx=x[test,],exact=T,x=x[train,],y=y[train])
```

Ridge regression and the lasso using the Auto data set

Repeat analysis in the last section for the Auto data, using the “full model” from Workshop 6

```
horsepower+I(horsepower^2)+cylinders+displacement+weight+acceleration+year
```

Comment on the cross validation result:
The function cv.glmnet uses it’s own grid and the “best” λ value is actually the minimum value of all the λ values investigated 0.74 (the first dotted line in the plot is at the left hand edge). We should always be concerned when an algorithm returns a minimum or maximum near the edge of the search range, it is probable that the true maximum/minimum lier outside this range. We should adapt the search grid to investigate further.

▷ Repeat the cross-validation using a grid from 0 to 1, to find the best lambda in this fine grid.

```
> grid2 = seq (0.0, 1.0, length=100)
> cv.out = cv.glmnet(x[train,], y[train], lambda=grid2, alpha=0)
```

What does the cross validation result suggest in terms of ridge regression? Carry on to the Lasso regression to see if a different result is obtained.

Exercises as homework

Exercise 1

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

(a) The lasso, relative to least squares, is:

- (i) More flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
- (ii) More flexible and hence will give improved prediction accuracy when its increase in variance is less than its decrease in bias.
- (iii) Less flexible and hence will give improved prediction accuracy when its increase in bias is less than its decrease in variance.
- (iv) Less flexible 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.

Exercise 2

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

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

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, 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 error of the noise term (irreducible error).

Exercise 3

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, 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 irreducible error.