Machine Learning 1
Data Science

Summer Semester 2018 **Prof. Tim Downie**

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Workshop 6
Cross-Validation and Bootstrapping

Start R in the usual way.
Check to see if you have the package R `ISLR` using `require`, and install it if necessary.

Model evaluation using cross-validation

The `Auto` data set

The main part of this workshop will be spent working through Lab 5.3 in James et al. This lab uses the dataset called `Auto` in the `ISLR` package, which was introduced in an earlier section in James et al. This is a similar but slightly larger data set to `mtcars`, which you used in one of the K-means clustering workshops. Spend a few minutes investigating the data set. Read the help file

```
> ?Auto
```

and quickly work through the commands on page 49–51 from

`> names(Auto)` onwards.

Side remark: Notice that on page 50 you attached the data frame `Auto` using `> attach(Auto)`. Be careful when using this command! In particular do not attach it a second time (e.g. when you work through the lab on page 191). When you have finished analysing this dataset you *ought to* detach it using

```
> detach(Auto)
```

To see which data frames have been attached and which packages have been loaded type the command

```
> search()
```

Quadratic regression model

In Chapter 3 a quadratic regression model with fuel consumption `mpg` as the outcome variable and `horsepower` as predictor variable was proposed. Before you start working on Lab 5.3 plot the data and overlay linear and quadratic regression functions.

```
plot(mpg~horsepower,data=Auto)
lm.hp =lm(mpg~horsepower ,data=Auto)
abline(lm.full)
lm.quad.hp =lm(mpg~horsepower+I(horsepower^2) ,data=Auto)
summary(lm.quad.hp)

#define the predictor function and call it fq
fq<-function(x)
  lm.quad.hp$coefficients[1]+lm.quad.hp$coefficients[2]*x+
    lm.quad.hp$coefficients[3]*x^2
curve(fq,40,230,add=TRUE)
```

Lab: Cross-validation

Work through Lab 5.3.1 to 5.3.3 on pages 190 to 194.

The Bootstrap

For the Bootstrap section in Lab 5.3, skip the section *Estimating the Accuracy of a Statistic of interest*. Instead run the code below, which is a similar example to the one in the lecture. This does not use the code in the R package `boot`, instead it uses basic R commands, so that you can learn the overall bootstrap approach. After you have run the code below, jump straight to the subsection *Estimating the Accuracy of a Linear Regression Model* on page 195.

```
#Create the original sample of length 75
xsamp<-rexp(75,???) #choose a parameter value somewhere between 0.01 to 0.05
hist(xsamp,breaks=10,main="Histogram of sample",xlab="x")

mean(xsamp) #mean of original sample

#take one resample (length 75 with replacement)
#and calculate the mean of the resample
resamp<-sample(xsamp,75,replace=T)
mean(resamp)
```

```
B<-100
#100 Bootstrap resamples
bsmean<-rep(NA,B)
for(i in 1:B){
  resamp<-sample(xsamp,75,replace=T)
  bsmean[i]<-mean(resamp) #store the resampled mean for this iteration
}

hist(bsmean,breaks=15,main="Histogram of the bootstrapped means",xlab="mean(x)")

var(bsmean) #bootstrap estimate of the variance of our estimator

quantile(bsmean,c(0.025,0.975)) #95% Conf int
#how do you get a 90% CI?

#add the conf int to the histogram
abline(v=quantile(bsmean,c(0.025,0.975)),col=2)
```

Model Selection using cross-validation

We have decided that a quadratic regression is appropriate for modelling `mpg` dependent on `horesepower`, but there are other variables in the dataset. Is it possible that some of these also have an influence on fuel consumption? It is quite likely that some but not all variables have some influence.

The traditional way of choosing which variables to include is using AIC (Akaike information criterion) and BIC Bayesian information criterion). These criteria give a statistic which penalises how many parameters are fitted in the model and the model which gives the lowest statistic corresponds to the best model using that criteria. The BIC method gives a much higher penalty to introducing new parameters. An overview of the method is given on page 78 in James et al.

We will instead use the leave one out cross validation MSE score for model selection, in a similar way to how we chose the best polynomial for horse power in the first section of this workshop.

We begin by defining our minimal model with `horsepower` and `horsepower`^{2}.

```
glm.fit=glm(mpg~horsepower +I(horsepower^2) , data=Auto)
cv.err =cv.glm(Auto ,glm.fit)
cv.err$delta
```

We will now add each of the other variables in turn and for each model calculate the CV error.

```
cv.error=rep (0,6)
glm.fit=glm(mpg~horsepower +I(horsepower^2), data=Auto)
cv.error[1] =cv.glm(Auto ,glm.fit)$delta[1]
glm.fit=glm(mpg~horsepower +I(horsepower ^2)+ year, data=Auto)
cv.error[2] =cv.glm(Auto ,glm.fit)$delta[1]
glm.fit=glm(mpg~horsepower +I(horsepower ^2)+ weight, data=Auto)
cv.error[3] =cv.glm(Auto ,glm.fit)$delta[1]
glm.fit=glm(mpg~horsepower +I(horsepower ^2)+ acceleration, data=Auto)
cv.error[4] =cv.glm(Auto ,glm.fit)$delta[1]
glm.fit=glm(mpg~horsepower +I(horsepower ^2)+ displacement, data=Auto)
cv.error[5] =cv.glm(Auto ,glm.fit)$delta[1]
glm.fit=glm(mpg~horsepower +I(horsepower ^2)+ cylinders, data=Auto)
cv.error[6] =cv.glm(Auto ,glm.fit)$delta[1]


plot(cv.error,type="b")
```

Which new variable gives the lowest LOOCV MSE? We now define the *current model* to be the minimal model plus this variable

```
mpg~horsepower +I(horsepower ^2)+ mynewvariable
```

Run another iteration by adding each other variable in turn to the current model. Does adding a third variable noticeably decreases the LOOCV MSE? Repeat this process by continuing to add variables until the LOOCV MSE increases or flatlines.

Output the summary table for your chosen best model.

For reference the AIC method suggests all the variables should be in the model:

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

and the BIC method suggests

```
mpg ~ horsepower + I(horsepower^2) + year + weight + acceleration
```