Machine learning II
Master Data Science
Winter Semester 2023/24 **Prof. Tim Downie**

Worksheet 2
Non-Linear Regression I
with solutions

The R packages you need each week will be given in the worksheet. These should already be installed if you are using the lab computers. If you are using your own computer you need to check if they are installed

```
> find.package("package.name.in.quotes")
```

and if not to install them

```
> install.packages("package.name.in.quotes")
```

To load an already installed package use

```
> require(package.name.nit.in.quotes)
```

or `> library("package.name.not.in.quotes")`
This week you will need the packages `glmnet` and `ISLR2`.

There is a "handwritten" exercise at the end of this worksheet.

Non-linear modelling: Wage data set

Work through Lab 7.8 in James et al. starting on page 311[1] through to the end of part 7.8.2 page 318. Read through the following remarks before you start!

- The short section between p315 "Next we consider the task of predicting whether an individual earns more than $250,000 per year" and p316 "This is often called a rug plot" is concerned with classification using logistic regression rather than non-linear regression. This topic was a subject covered in ML 1 but is good revision.

[1]Section and page numbers refer to the second edition. If you haven't already downloaded this book you can do so at `https://www.statlearning.com/` and clicking on *download the second edition*.

- When you get to the part on natural splines (p 317 towards the bottom), add dashed lines to get the precision using the natural spline estimate in the same way you did for the B-splines at the top of that page. Notice that the precision is noticeably better at the edges.
- The last paragraph in 7.8.2 concerns *loess smoothing*, which we will look at in detail next week. It doesn't hurt to work through the few commands this week though.

Non-linear modelling: Motorcycle helmet acceleration

The the motorcycle helmet acceleration data are part of the MASS library which is pre-installed with R, so enter

```
> library(MASS)
```

to access to the data. Read the help page for the data set `mcycle`.

Plot the acceleration against time in a scatter plot. This is a hard regression problem because the acceleration variable has several phases to it, with different characteristics in each phase.

Use what you have learnt in the previous section to fit the acceleration data using the methods listed below. To do this you will need to consider which variables, parameters and values need to change. For example, the x-axis grid for plotting the predictor functions and choosing 3 or four sensible knot points.

For each method plot the data and the predictor function. Make an informal visual comparison of the different methods.

(a) Polynomial regression with degree 4

(b) Polynomial regression with degree 10

(c) Step function

(d) Constrained piecewise linear regression (use `bs(???,degree=1)`) and 3 knot points

(e) Cubic spline regression using B-splines with 3 knot points. Calculate and plot the confidence interval for the predictor function.

(f) Cubic spline regression using natural splines with 3 knot points. Calculate and plot the confidence interval for the predictor function.

(g) Spline smoothing. Start with `df=4` and slowly increase it's value.

(h) Spline smoothing with cross validation. What is the effective degrees of freedom for the LOOCV optimum? `> fit$df` *12.7*

Exercise as homework

This exercise is based on the constrained linear regression model and basis functions section from the lecture notes, slides 13 to 15.

(a) Use the graphic "B-Splines degree 1" to obtain the formula for each basis function $b_k(x)$ with $k = 1, 2, 3, 4$ in the following form:

$$b_k(x) = \begin{cases} ax + b & \text{for } x_l \leqslant x \leqslant x_m \\ cx + d & \text{for } x_m \leqslant x \leqslant x_u \\ 0 & \text{otherwise.} \end{cases}$$

For each k you should give the numeric values of a, b, c, d, x_l, x_m and x_u.

(b) The coefficient values for the pwl-function are given on slide 12. Use these coefficients Express the predictor function $f(x)$ for $50 \leqslant x \leqslant 75$ as a linear function with numeric coefficients.

(a)

$$b_1(x) = \begin{cases} \frac{1}{25}x & \textit{for } 0 \leqslant x \leqslant 25 \\ 2 - \frac{1}{25}x & \textit{for } 25 \leqslant x \leqslant 50 \\ 0 & \textit{otherwise.} \end{cases}$$

$$b_2(x) = \begin{cases} \frac{1}{25}x - 1 & \textit{for } 25 \leqslant x \leqslant 50 \\ 3 - \frac{1}{25}x & \textit{for } 50 \leqslant x \leqslant 75 \\ 0 & \textit{otherwise.} \end{cases}$$

$$b_3(x) = \begin{cases} \frac{1}{25}x - 2 & \textit{for } 50 \leqslant x \leqslant 75 \\ 4 - \frac{1}{25}x & \textit{for } 75 \leqslant x \leqslant 100 \\ 0 & \textit{otherwise.} \end{cases}$$

$$b_4(x) = \begin{cases} \frac{1}{25}x - 3 & \textit{for } 75 \leqslant x \leqslant 100 \\ 0 & \textit{otherwise.} \end{cases}$$

(b) When $50 \leqslant x \leqslant 75$*, then* $f(x) = 1.65 + 0.73b_2(x) + 4.10b_3(x)$ *(*$b_1(x)$ *and* $b_4(x)$ *are zero in this interval).*
Using part (a) $f(x) = 1.65 + 0.73(3 - \frac{1}{25}x) + 4.10(\frac{1}{25}x - 2) = -4.36 + 0.1348x$