Machine learning II
Master Data Science
Winter Semester 2019/20

BEUTH HOCHSCHULE FÜR TECHNIK BERLIN
University of Applied Sciences

Prof. Tim Downie

Workshop 2
Non-Linear Regression Models
with solutions

Non-linear modelling: Wage data set

Work through Lab 7.8 in James et al. starting on page 287 until the end of 7.8.2 at the top of p. 294.The last paragraph concerns *loess smoothing* which will look at in detail next week, but it doesn't hurt to work through the few commands here.

The short section on pages 291 and 292 between "Next we consider the task of predicting whether an individual earns more than $250,000 per year" and 'rug plot" is concerned with classification via a logistic regression rather than non-linear regression. This was a subject covered in ML1 but is good revision.

Non-linear modelling: Motorcycle helmet acceleration

Read the help page for the data set `mcycle` (in the MASS library) and 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 following methods. To do this you will need to change some of the variables and parameters, such as the x-axis grid for plotting, the knot points etc.

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? *12.7*

Exercise as homework

This exercise is based on the constrained linear regression model and basis functions section from the lecture, slides 13 and 14.

(a) Use the graphic "B-Splines degree 1" to write a formula for each basis function $b_1(x)$, $b_2(x)$, $b_3(x)$ and $b_4(x)$ in the form

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

(b) The coefficient values 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* $f(x) = 1.65 + 0.73b_2(x) + 4.10b_2(x)$ *(all other terms are zero).*
Using part (a) $f(x) = 1.65 + 0.73(3 - \frac{1}{25}x) + 4.10(\frac{1}{25}x - 2) = -4.36 + 0.1348x$