MTH6101 Introduction to Machine Learning

Laboratory week ten

The intention of this laboratory is to analyse the `diabetes` data set (Efron et al. 2004), using ridge regression. The data set had $n = 442$ diabetes patients measured on $p = 10$ baseline variables. A prediction model was desired for the response variable `y`, a measure of disease progression one year after baseline.

Before you **start** your `RStudio` session, install and load the following libraries: `cvTools`, `lars` and `ridge`.

1. The initial step is to load the data from library `lars` with `data(diabetes)`. Using the instruction `as.matrix`, you will turn the diabetes data into a matrix called `X`, for which you will only use the first 11 columns which will be centered and scaled. As a final step, create a data frame called `DAT` with `X`. Have a look at the column names of the variable you just created.

2. Create validation index variables `Train` and `Test` for an $3:1$ partition of the data. Use the function `cvFolds` from library `cvTools`.

3. Using the command `seq`, create a sequence of 200 values in the range -9 to 6 which are to be exponentiated so that λ ranges from 10^{-9} to 10^{6}. Store this sequence in variable `rangelambda`.

4. Using the training partition and the function `linearRidge` from the library `ridge`, train the **ridge regression** model and store it in a variable termed `LR`. The data for this is `DAT[Train,]`. Use the remaining part of the data to build predictions which are to be stored in variable `PR`. The data for predictions is `DAT[Test,]`.

5. Examine the variables `LR`, `PR` you have created and become familiar with its structure. To compute the validation MSE, allocate fresh observations to matrix `Yobs` with `matrix(nrow=nrow(PR),ncol=ncol(PR),DAT$y[Test])->Yobs` so that you can compute this error every value of lambda with function `apply` and store the error in variable `MSER`.

6. Find which is the value of λ that minimizes MSE. Then plot `MSER` against λ (variable `rangelambda`) and indicate the location of the minimum with a vertical line. For this plot, the horizontal axis has to be with option `log="x"`.

7. Plot the **ridge trace** and give the coefficients suggested by ridge regression.