M5MS10 Machine Learning
Computer Lab 1

Dr Ben Calderhead

1 Laboratory Exercise

In this computer lab you will gain experience of using R or Matlab to perform supervised learning with linear regression models. You will investigate the effect that increasing model complexity has on the mean squared error (MSE), and consider the use of cross validation as a means of choosing the most appropriate predictive model.

1.1 Training, Testing and Cross-validated Errors

You can download the code for this laboratory from the Imperial Blackboard. There is a function for generating data from a given polynomial, and a function for computing the average error using leave-one-out cross validation (LOOCV).

- Write a function that computes the MSE using all the data for a given order of polynomial. Generate data using the function provided. You should be able to adapt code from the LOOCV function provided. What happens to the MSE as the model becomes more complex? Split the data into a training set and a test set. Learn the parameters using the training set, again for increasing orders of polynomials, and plot the MSE for both the training sets and test sets. What happens to the training and testing errors as model complexity increases?

- Now consider the LOOCV code. Plot the average MSE using LOOCV for increasingly complex models. Does LOOCV do a good job of identifying the correct model? Have a look in the code that generates the data and change it to generate data from 4th and 5th order polynomials. Does LOOCV still do a good job of identifying the correct model?

- Finally, download the long jump dataset we saw in the first lecture. Use these tools to identify the most appropriate polynomial model for this data. Does LOOCV do a good job this time? Discuss your findings.