**Choice of assignment:** Predicting the biodegradability of chemicals from QSAR data.

**Data Preparation:** I loaded the data, matrix ***D***, into MATLAB where I then split column 11 into its own array ***z***. This array is the target data of whether a chemical is biodegradable or not. The input data, columns 1 through 10, were placed into another matrix ***X***. I did not perform any normalisation on this data as each column of data was within a similar range, approximately between ± 15.0000. I then chose my training sample size ***n***, I decided I wanted ***n*** to equal three quarters the total size of each column of input data. Three quarters of 1055 does not equal an integer, so my MATLAB code rounds the result down to the closest integer, this meant I had a training sample data of column length 791 and an out of sample (OOS) column length of 264. Both ***X*** and ***z*** were then randomly split into training sample ***x*** and ***z***, and OOS ***x\_star*** and ***z\_star*** with the sizes mentioned above.

**Model building parameters:** There were certain things that had to be chosen in order to build a model, initially the first choice was what type of model I was going to create. I decided to use a create a multilayer perceptron (MLP). I chose to use an MLP because we created an MLP for a different classification problem during one of the lab classes, perhaps there were better model choices but since I was familiar with creating an MLP for a smaller but similar problem, I decided to stick with an MLP. I also needed to choose the number of cross folds ***k***, number of hidden units (# of HUs) ***nhid***, number of iterations ***nits*** and whether the output unit activation function was linear or logistic.

I chose ***k*** to equal 5, I chose that value as 5 or 10 seem to be the usual values, even though there is not a formal rule [1]. Rather than picking a singular # of HUs, I made ***nhid*** an array of 7 elements between 5 and 15. I chose this as I was unsure what the best # of HUs was so within my code, I automated the process on choosing which is the best # of HUs.

[1] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Springer New York, 2013.