Machine Learning 2015: Project 1 - Regression Report

rabha@student.ethz.ch
rsridhar@student.ethz.ch
xandeepv@student.ethz.ch

October 31, 2015

Experimental Protocol

To reproduce the results, the new user just need to specify the correct filepaths in our R code for both Validation and Training Data Files. The output of the validation or test data is written to a file with the output for the model predicted values for the Processor delay in microseconds. The code has been made in such a way that no user inputs are required during the code run. The code expects the data files to have the Column Headers and first 14 columns with the different characterisitcs of the processor. The Traning data file should have the 15^{th} column with the measured Processor delay in microseconds.

1 Tools

Our model is implemented in R Language with the use of MASS library, which is the R implementation of 'Modern Applied Statistics with S' by W.N Venables and B.D. Ripley.

2 Algorithm

Our Algorithm for Regression is implemented with the help of Four user defined functions. The first fucntion $crossval$ is used to implement the N-fold Cross Validaton on Ridge Regression.The second function gcv gives the optimal Lambda value for General Cross Validation (GCV) Error for the Ridge Regression. The third function $greedyFeatures$ implements a greedy forward feature selection method to get new features from the Processor Characteristics by applying different transformations. This function implements the possible transformations that could be applied to the Processor Characteristic data like $x^2, \sqrt{|x|}, \log|x|, x, x^3, x\log|x|, x^2\log|x|, x^2\log|x|^2$ for any characteristic parameter x. This function is called with the user's preference for the number of new features to be extracted. The fourth function $transformFeatures$ is used to transform any data to the new features extracted from the greedy feature selection done with the above function.

3 Features

The fucntion $greedyFeatures$ can take number of new features to be extracted as input and extract that many number of features and we have taken 11 features for our final model. The Table 1 shows the number of New features added (other than the 14 Characteristic Parameters) and the corresponding Kaggle scores.

Table 1: Kaggle score for New Features

Sl No.	Features	Kaggle Score
1	8	662.46542
2	10	659.94534
3	11	656.90319
4	12	662.30074
5	13	668.63267

4 Parameters

Our greedy approach algorithm gave the below new features for the 13 features (ordered according to the selection) and we ended up in the model using first 11 features along with the other 14 Processor characterisitcs.

1. logabs(Width)*depth
2. logabs(LSQ)*Width
3. Id(logabs(Width)*depth)*Width
4. logabs(RFRead)*Id(logabs(Width)*depth)*Width
5. logabs(LSQ)*logabs(LSQ)*Width
6. logabs(LSQ)
7. logabs(RFRead)*logabs(RFRead)*Id(logabs(Width)*depth)*Width
8. logabs(ROB)
9. logabs(IQ)*logabs(LSQ)*logabs(LSQ)*Width
10. sqrtabs(Branches)*RFWrite
11. sq(RFRead)*RFSize
12. cubic(l1dcahce)*RFSize
13. xxlogabsq(RFWrite)*sq(RFRead)*RFSize

where logabs is $\log|x|$, sqrtabs is $\sqrt{|x|}$, xxlogabsq is $x^2\log|x|^2$, sq is x^2 and cubic is x^3

5 Lessons Learned

We started with simple models and then to polynomial fits. The Kaggle scores for these models were very low and we understood the need to automate feature selection. We came up with the greedy algorithm for feature selection. From the features selected we could understand that the log transformation could enhance the results. Also we could see that the Processor Characteristics of width, depth, LSQ, RF Reads and RF writes have major impact on the Processor delays. Improving these parameters greatly decrease the Processor Delay. This is technically true about the effect of processor characteristics on the Processor delay.