Machine Learning 2014: Project 1 - Regression Report

anguyen@student.ethz.ch
spark@student.ethz.ch
frenaut@student.ethz.ch

October 18, 2014

Experimental Protocol

As first analysis, each feature was plotted against the response variable in order to spot possible non-linearities and non-significant parameters.
Afterwards, the parameters were estimated as described in the next sections.

1 Tools

Both the analysis and the estimation were carried out in *MATLAB*.
Git was used to keep track of progress and versioning of our algorithm.

2 Algorithm

Ridge regression has been used to estimate the model's parameters. The optimal penalizing parameter λ was chosen by minimizing the prediction error estimated using *Cross Validation* (the number of subsets was set to $\sqrt{n}$, where n is the sample size.) (Figure 1).

3 Features

In a first step, we exponentiated some of the given features (e.g. $x_i \rightarrow x_i^a$ with $a \in [-2 : 0.5 : 5]$ and $a \neq 0$).
Next, we combined some of the transformed features in order to consider possible interactions.
Hence, our final model has the form $y = \beta_0 + ... + \beta_i x_i^a + ... + \beta_z x_j^b x_k^c + ...$

4 Parameters

Cross Validation was used to find how to transform and combine the parameters. In details:

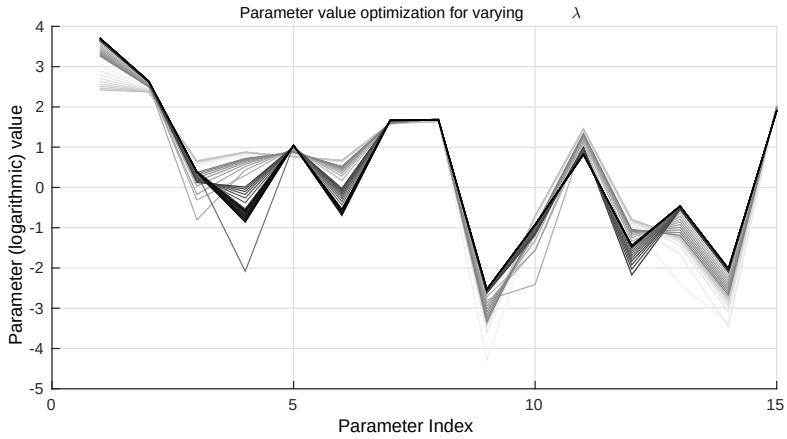

Figure 1: Parameter evolution during prediction error minimization. Darker values are closer to the optimal λ.

- For each feature x_i, the exponent a is chosen so that the estimated prediction error is minimized.
- For each pair of feature x_i and x_j, the combination term $x_i x_j$ is included only if the estimated prediction error decreases by a certain percentage.

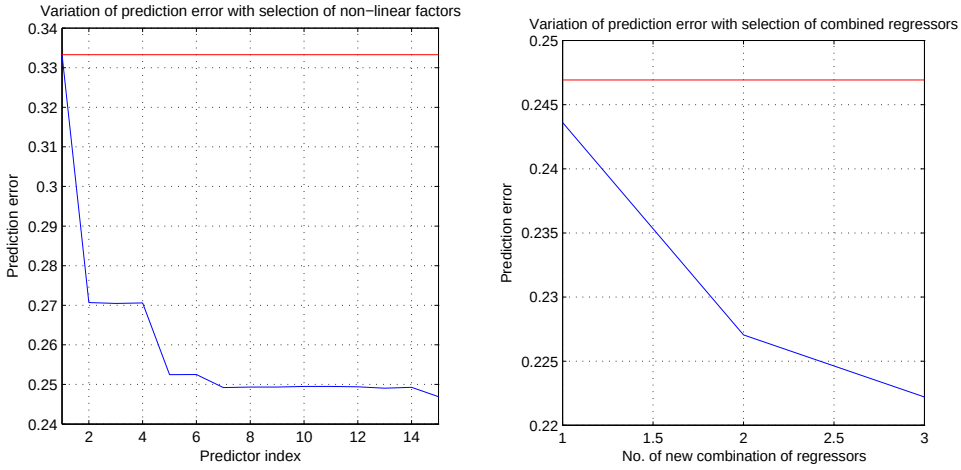

Figure 2: Error reduction through feature transformation and combination.

5 Lessons Learned

Other strategies that we tried include feature discarding[1], response transformation and different feature transformations (e.g. $log(\cdot)$). The resulting worse performance is most likely given by a loss of information (in the case of the discarded feature) or wrong model assumptions.

[1]We computed the parameters β on normalized data and excluded the features corresponding to the lowest $|\beta|$.