Machine Learning 2013: Project 1 - Regression Report

fregli@student.ethz.ch
ganzm@student.ethz.ch
sandrofe@student.ethz.ch

October 26, 2013

Experimental Protocol

This project was performed using Matlab only. To reproduce test results presented in this report the following steps have to be taken:

- Unzip the sourcefolder containing Matlab code and data sets
- Run learn.m to optain both files testresult.csv and validationresult.csv

1 Tools

As stated in Section Experimental Protocol the only tool which is needed is Matlab - no fancy special commands, no additional libraries are required.

2 Algorithm

The algorithm performs the following steps an can be started by running learn.m

- Read csv file
- Normalize training input and result data
- create feature vectors
- perform cross validation and obtain weight vector w
- generate validation and testing output

2.1 Which part of the code does what

learn.m This is the main script which can be run to obtain the submitted results.

```
x = trainingData(:,1:14);
y = transform(trainingData(:,15));

% prepare features
xfeatures = [y, x, x.^2, x.^3, x.^4, sqrt(x), log2(x)];

% calculate correlation
corr = corr(xfeatures, xfeatures);

% only consider how good the y feature correlates
c = corr(:,1);
```

Figure 1: Code Sample to find correlation

trainData.m Performs Ridge Regression and solves the following problem

$$\min_w \sum_{i=1}^{n} \left(y_i - w^T x_i\right)^2 + \lambda ||w||_2^2$$

using the exact solution:

$$w = \left(X^T X + \lambda I\right)^{-1} X^T y$$

As a result we obtain the weight vector w and an the RMSE.

extractFeatures.m This is the part of the code where features are created and transformed.

3 Features

3.1 Find good features

TODO

4 Parameters

4.1 Parameter λ

By iteratively applying cross validation best parameter λ is determined as seen in Code Sample to find best λ. We chose an arbitrary number of parameters (50 in this sample) and apply cross validation. The parameter which yields the smallest error is then chosen.

5 Lessons Learned

Most of the time spent for the project was used to search for good features. At the start of the project it was mostly unclear how to find such features. Getting a test score which meets our expectation by

```
% perform crossvalidation
lambdaValues = logspace(-6, 2, 50); % hyper parameter
meanErrs = zeros(size(lambdaValues));

for i=1:size(lambdaValues,2)
    [meanErrs(i), W, errorTest] = crossvalidation(Xnorm, Ynorm, lambdaValues(i));
end
```

Figure 2: Code Sample to find best λ

simply try and error features did not yield the expected result.

We later automated our search for good features by applying the methods described in Find good features. This lead to much better result.