Machine Learning 2014: Project 1 - Regression Report

lukasbi@student.ethz.ch
ajenal@student.ethz.ch
harhans@student.ethz.ch

October 31, 2014

Experimental Protocol

1 Tools

For this exercise we used mainly Matlab. All the statistic computation was performed with Matlab. However, some analysis (scatter plots) was done with VisuLab. The main libraries we used will be found in the Bioinformatics toolbox.

2 Algorithm

We tried different algorithms, but we stuck with ridge regression in its closed form. In our opinion it performed better than the ordinary least squares as well as the lasso algorithm. So basically we solved the equation:

$$\beta = (X' * X + \lambda * I)^{-1} * X' * y \tag{1}$$

Tuning lambda had a relevant impact on the final result and was a challenging task.

3 Features

We did a lot of fine tuning with feature selection and generation for optimizing our result. To generate some basic features we used the provided function: `Matlab.x2fx` with the `'quadratic'` parameter. But we extended our features space with much more non-linear dependencies, where also single features interfere with each other.

To improve the result we did not use all features from the features space but selected the best ones. For finding the best features we used a straight forward approach, where we iterated over the features space and computed the ridge regression for each feature. The ones with minimal error were kept and the others discarded. Important to note is that as soon a feature was selected it was never removed from the selected feature space.

4 Parameters

Vital for feature selection and parametrization is the normalization of the feature vectors. The domain of every dimension of the features have to be dimensionless and normalized to ensure a balanced weighting. The parameters used here were the mean distribution of the training data t.m. mean feature f_{avg} and standard deviation σ_{avg} to f_{avg}. We normalize the training set by subtracting f_{avg} from all points and dividing by σ_{avg}. This is done with the following code:

```
MEAN = mean(training);
STD = std(training);
averagedata = training-repmat(MEAN, size(training,1),1);
normdata = bsxfun(@rdivide, averagedata, STD);
```

5 Lessons Learned

As mentioned above we used Visulab for scatter plots. But this program was not so handy and with Matlab it was easier to visualize essential things like the correlation of different features. For the features selection we also used the `lasso` function provided by Matlab. However, it performed worse than our "brute force" feature selection method. Most probably this was the case because choosing the right `lambda` heavily affects the solution. For solving the minimization problem we also tried the residual sum squared and conjugate gradient approach, but those two methods didn't improve our solution, so we stayed with the ridge regression.