Computational Data Analysis - Case 1

Anders Launer Bæk (s160159)

11 Marts 2018

Sparring partner:

- Grétar Atli Grétarsson (s170251)

Selected models

It has been chosen to estimate a linear model with the below mentioned approaches and combine these individual responses in an final ensemble model.

- Ridge regression: $\beta_{ridge} = arg\,min \parallel Y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2$
- Lasso regression: $\beta_{lasso} = arg\,min \parallel Y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1$

Both of the latter mentioned models uses a shrinkage approach in order to force the model coefficients towards zero which reduces the complexity in the model structure.

Preprocess data

Data provided data exists of 101 columns with 1100 rows. 100 of theses 1100 rows does include a valid responds value.

The following pre-processing have been done to the data set:

- There are 442 of 1100 (40%) rows with missing data. It has been chosen to impute the missing value for the each column by inserting the expected value for the given column.
- The column X100 is a categorical variable with 3 levels (A, B, C). It has been chosen to transform X100 in to 3 dummy variables: X100A, X100B, X100C.
- The distribution of each column have been plotted in a histogram in order to investigate the demand for transformations. There is no skewness in the distributions and hereby no need for transformations of the column according to the visual inspections. The levels of the categorical variable (X100) is reasonable balanced in overall, in train set and in the test set (these without responses), see table 1.

Table 1: Percentage balance of the 3 levels in the X100 variable.

X100	All (%)	Train (%)	Test (%)
A	34	29	33
B	34	29	33
C	31	40	32

Prepare data for cross validation train and validate

Is has been chosen to use 5-fold cross validation with a validation set on 30% of the total 100 observations. ** TODO WHY 5-fold **

The latter mentioned two models are trained and validated on exactly the same indices. The MSE have been calculated for each loop in their hyperparameter search.

- The Ridge regression does only depend on λ_{ii} and is calculated on close form as follows: $\beta_{\lambda_{ii}} =$

$(X_{train}^T X_{train} + \lambda_{ii} I) X_{train}^T)^{-1}(Y_{train} - \mu_{Y_{train}})$.

- The approach for the Lasso regression is inspired by the: `glmnet(X_train, Y_train - Y_train_mean, lambda, alpha = 0, standardize = F, intercept=F)` function. The coefficients for λ_{ii} can be extracted by the `coef()` function.

Please notice that the mean of the responds are subtracted. This is due to the characteristic of the shrinking methods. By not center the responds variable around its mean the model estimation will introduce a bias to the model.

The train design matrix will standardized by the native `scale()` in `R` and the validate design matrix will be scaled according to the train design matrix in each fold.

Figure 1 illustrates the MSE as a function of λ for the two models.

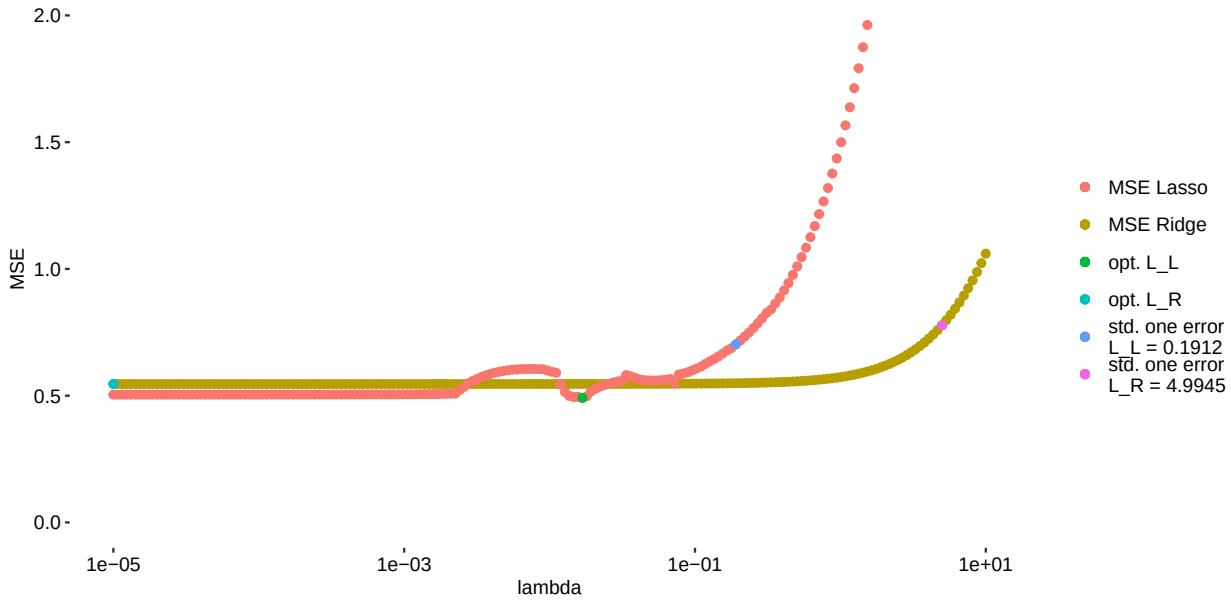

Figure 1: MSE as a function of lambda for Ridge and Lasso regression.

Model Assesment based upon 30% test set

In order to make sure that the best possible estimate of the prediction error, the estimated models are evaluated on the 30% test set. The models uses the optimal parameters found in the cross validation. The parameters are as follows:

- $\lambda_{Ridge} = 4.9945051$
- $\lambda_{Lasso} = 0.1911644$

Table 2 reports the performance metric of the 30% validate set (model selection).

Table 2: Performance of the two models and the combined ensemble model.

Algo	R2	MSE	RMSE
Ridge Regression	0.9972855	0.0207666	0.0521005
Lasso Regression	0.9998977	0.0007824	0.0101126
Ensemble Model	0.9990615	0.0071799	0.0306351

As it appears from the table above, the bottom row is the performance metrics for the ensemble model. The weight ratios between the two models are given below:

- $w_{Ridge} = \frac{(1-0.0521005)}{(1-0.0521005)+(1-0.0101126)} = 0.489166$
- $w_{Lasso} = \frac{(1-0.0101126)}{(1-0.0521005)+(1-0.0101126)} = 0.510834$

Expected rRMSE

The expected `rRMSE` is:

- `rRMSE`=0.0306351

Re-train the model on compleate train data

The models will be re-trained on the complete train data after the unbiased `rRMSE` estimate has been stated. New parameter estimates is obtained and new weight ratios used in the ensemble is calculated.

The new weight ratios between the two models are slightly changed compare to weights found by using the validate set (model selection). The weights for the ensemble model are given below:

- $w_{Ridge} = \frac{(1-0.1687004)}{(1-0.1687004)+(1-0.1745416)} = 0.5017628$
- $w_{Lasso} = \frac{(1-0.1745416)}{(1-0.1687004)+(1-0.1745416)} = 0.4982372$

Reflections

What could be improved and why?

- ...

Predict missing responds

COMMING UP!

- Obtained rRMSE on 1000 observations: `rRMSE=`