Computational Data Analysis - Case 1

Anders Launer Bæk (s160159)

14 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: $\beta_{ridge} = \underset{\beta}{arg\,min} \parallel Y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_2^2$
- Lasso regression: $\beta_{lasso} = \underset{\beta}{arg\,min} \parallel Y - X\beta \parallel_2^2 + \lambda \parallel \beta \parallel_1$
- Elastic net: $\beta_{elastic\,net} = \underset{\beta}{arg\,min} \parallel Y - X\beta \parallel_2^2 + (1-\alpha) \parallel \beta \parallel_1 + \alpha \parallel \beta \parallel_2^2$ where $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1} = 0.5$ in order to have a hybrid bwtween the Rigde and Lasso regression.

The three models use a shrinkage approach in order to force the model coefficients towards zero which reduces the complexity in the model structure.

Preprocess data

The data provided exists of 101 columns with 1100 rows out of which 100 rows includes a valid respond 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 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 has 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) are acceptable balanced 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 100 train observations.

By incresing the number of folds to 10 or to a more drametically approaches such as the leave-one-out cross

validation it will usaually result in a lower bias but a greather variance in the estimate of the parameters.

The three 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 only depends 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 calculating 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 of the regression for a given λ_{ii} can be extracted by the `coef()` function.
- The approach for calculating the Elastic net regression is inspired by the: `glmnet(X_train, Y_train - Y_train_mean, lambda, alpha = 0.5, standardize = F, intercept=F)` function. It has been chosen to force $\alpha = 0.5$ which is the mid point between the Ridge regression and the Lasso regression.

Please notice that the mean of the responds are subtracted in both models. 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 be standardized by the `scale()` function 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 three models.

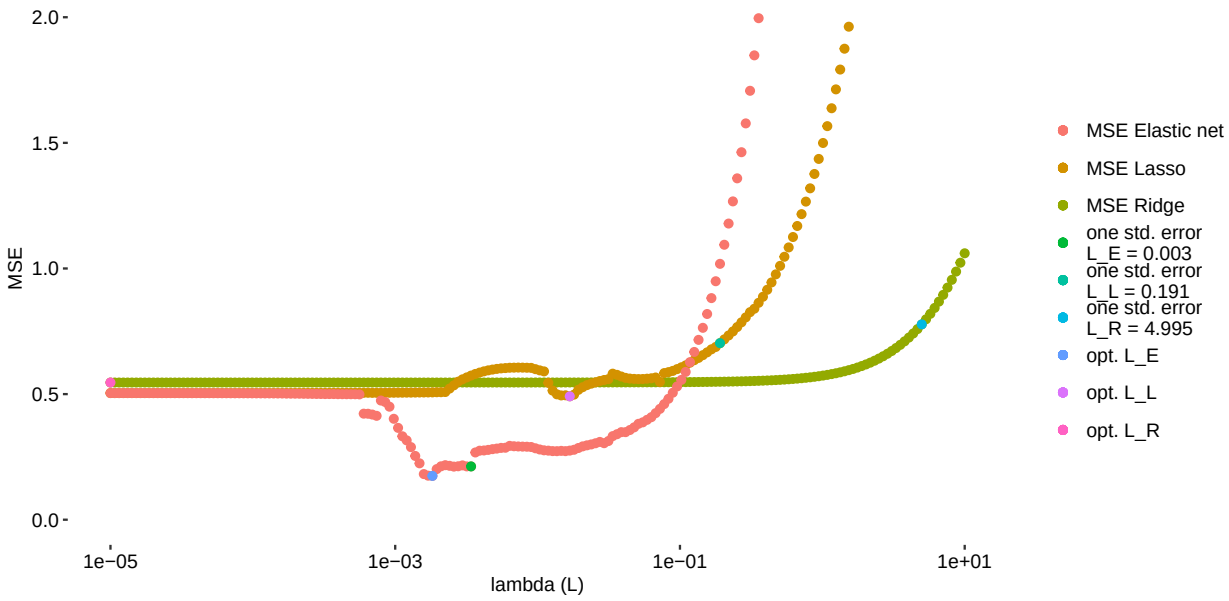

Figure 1: MSE as a function of lambda for the Ridge regression and the Lasso regression.

Model Assessment based upon 30% test set

The estimated models are evaluated on the 30% test set in order to ensure the best possible estimate of the `RMSE` of the unseen data. The models use the optimal parameters plus the one standard error rule which is found by the 5-fold cross validation. The parameters are as follows:

- $\lambda_{Ridge} = 4.9945051$
- $\lambda_{Lasso} = 0.1911644$
- $\lambda_{Elastic\,net} = 0.0034093$

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

Table 2: Performance of the three 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
Elastic net	0.9999970	0.0000231	0.0017388
Ensemble Model	0.9995700	0.0032899	0.0207373

The bottom row of table 2 is the performance metrics for the ensemble model. The weight ratios between the three models are given below:

- $w_{Ridge} = \frac{(1-0.0521005)}{(1-0.0521005)+(1-0.0101126)+(1-0.0017388)} = 0.3228488$
- $w_{Lasso} = \frac{(1-0.0101126)}{(1-0.0521005)+(1-0.0101126)+(1-0.0017388)} = 0.3371496$
- $w_{Elastic} = \frac{(1-0.0017388)}{(1-0.0521005)+(1-0.0101126)+(1-0.0017388)} = 0.3400017$

Expected RMSE

The expected `RMSE` is:

- `RMSE`=0.0207373

Re-train the model on complete train data

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

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

- $w_{Ridge} = \frac{(1-0.1687004)}{(1-0.1687004)+(1-0.1745416)+(1-0.096746)} = 0.3247249$
- $w_{Lasso} = \frac{(1-0.1745416)}{(1-0.1687004)+(1-0.1745416)+(1-0.096746)} = 0.3224432$
- $w_{Elastic} = \frac{(1-0.096746)}{(1-0.1687004)+(1-0.1745416)+(1-0.096746)} = 0.3528319$

Achieved RMSE of the missing respondses

COMING UP!

- Achieved RMSE of the 1000 observations: `RMSE=`