Technical University of Denmark

Case 1

Course:
02582 Computational Data Analysis

Author(s):
Daniel Thoren (s144222)

March 26, 2019

1 Data cleaning and pre-processing

After reading the CSV file, the data is divided input, X and output, y. Due to the small size of the dataset, **missing values** are replaced by the column's mean or mode, for numerical and categorical columns, respectively. Thereafter the categorical features ($X_{96}, X_{97}, X_{98}, X_{99}, X_{100}$) are quantified. For example X_{96} is replaced by a two-feature matrix $[X_{96A}, X_{96B}]$ (because it only contains `A`s and `B`s) and the matrix is filled using one-hot encoding:

$$\begin{cases} \text{if } X_{96} \text{ is } A, & X_{96A} = 1, \text{else } 0 \\ \text{if } X_{96} \text{ is } B, & X_{96B} = 1, \text{else } 0 \end{cases}$$

Due to the small size of the data, missing input features within an observation, are replaced with the column average. Before any regularization is done, the data is standardized - moving the variables to a zero-mean and making them independent of their scale.

2 Models

Only linear models are considered for solving this problem, including, OLS, LARS, Ridge and ElasticNet. Each method regularizes the weight parameters differently.

Cross-validation is used to estimate the generalization error of these models. In this case 10-fold cross-validation is used, which results in a training set of size 90 and a test set of size 10. Furthermore, cross-validation is used to determine optimal regularization parameters by running 10-fold cross validation across parameter combinations.

2.1 OLS

Ordinary Least Squares overfits the data with zero training error (naturally) and a high test error of 1.004 - as shown in table 1.

Table 1: OLS performance

Model	Training error	Test error
OLS	0.0000	1.004

2.2 LARS

Introducing regularization increases training error and lowers test error. Thus, the LARS model performs much better than OLS. Figure 1 shows the training error and test error as a function of the number of non-zero coefficients.

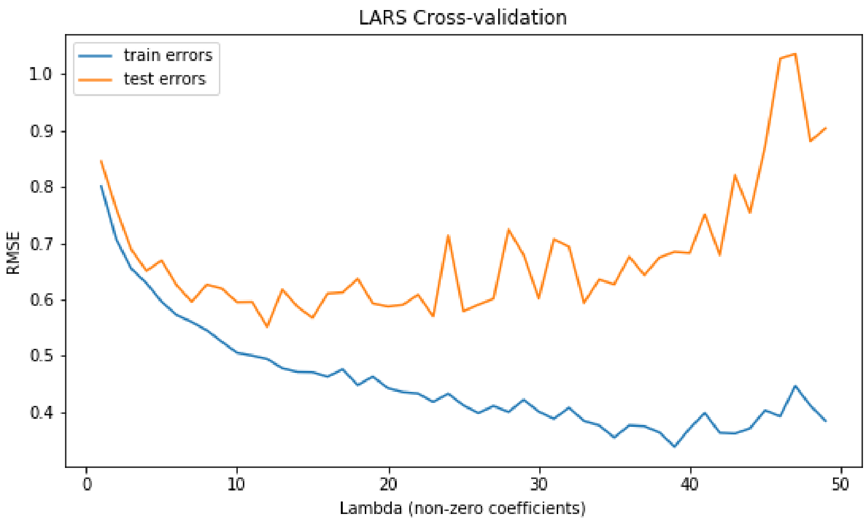

Figure 1: Cross-validation of the LARS model

2.3 Ridge

Figure 2 shows the performance of the Ridge model as a function of the L1-norm. It shows that Ridge performs worse than LARS.

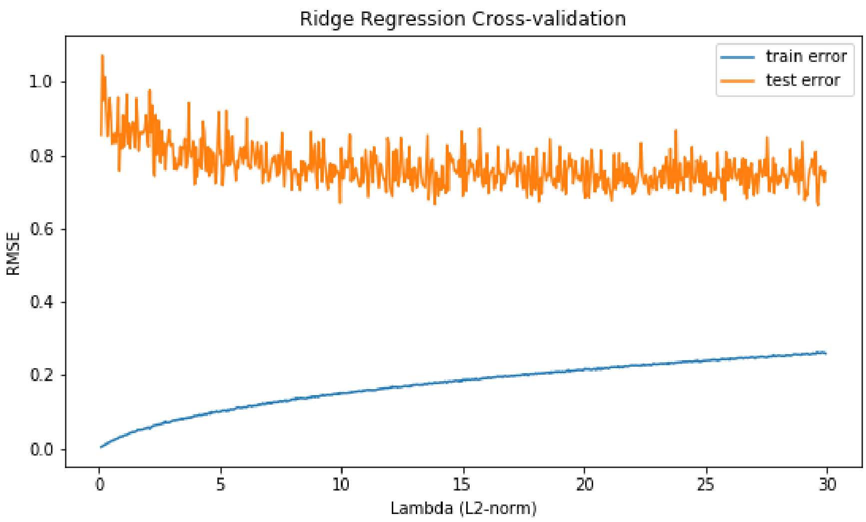

Figure 2: Cross-validation of the Ridge model

2.4 ElasticNet

ElasticNet forms a robust sparse estimate using the combination of L_2-norm shrinkage from Ridge and L_1-norm parameter selection from Lasso. Figure 3 shows the performance of the its by plotting relative mean squared error (RMSE) at differing values of the L2-norm and L1-norm.

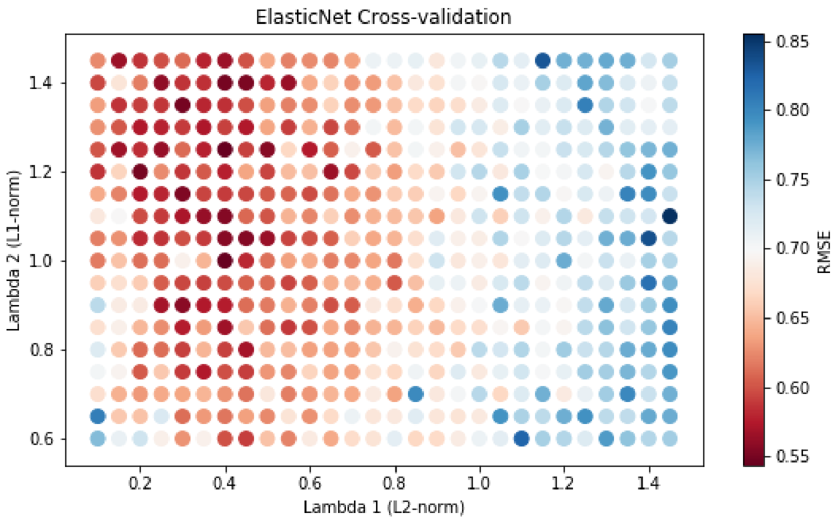

Figure 3: Cross-validation of the ElasticNet model

3 Model Selection

The top five best performing models are all ElasticNet models as shown in table 2.

Table 2: Top 5 models by RMSE (sorted by test error)

Model	L1-norm	L2-norm	Training error	Test error
ElasticNet	0.4	1.0	0.4679	0.5435
ElasticNet	0.4	1.25	0.485	0.5437
ElasticNet	0.2	1.2	0.3794	0.5498
ElasticNet	0.4	1.4	0.4921	0.5502
ElasticNet	0.3	1.35	0.4528	0.5507

The final model is chosen using the one-standard-deviation rule with a test error standard deviation of 0.0649. The final model is shown in table 3.

Table 3: Final model with $+1\sigma$ from the best model

Model	L1-norm	L2-norm	Training error	Test error
ElasticNet	0.55	1.15	0.5264	0.609

The features most heavily weighed by the model - along with their relative weights are shown table 3. The predictions are expected to have an RMSE of around 0.6 based on the model selection process.

Table 4: Weight parameters of the final model

Variable	Weight
X_2	1.412
X_{42}	0.66
X_{56}	3.039
X_{97C}	-0.342
X_{98C}	0.281
X_{99C}	0.039
X_{100D}	0.093