University of Kassel
Dept. of Measurement and Control
© Prof. Dr.-Ing. Andreas Kroll

Exercise 2: Linear Model Selection and Regularization

1. Getting Familiar with Stepwise Model Selection in Matlab

In this exercise, you will generate simulated data and use this data to perform stepwise subset selection. Make sure to use `rng(1000)` prior to starting part (a) to ensure consistent results.

(a) Use the `randn()` function to generate a predictor X of length $N = 1000$, as well as a noise vector ϵ of length $N = 1000$ with a standard deviation of 0.8.

(b) Generate a response vector Y of length $N = 1000$ according to the model

$$Y = \beta_0 + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3 + \epsilon \tag{1}$$

where $\beta_0 = 2$, $\beta_1 = 3$, $\beta_2 = -1$, $\beta_3 = 2$.

(c) Split your data set into a training data set containing the first 100 observations and a test data set containing the remaining 900 observations.

(d) Use a for loop to train a set of polynomial models with increasing degree containing the predictors X, $X^2, \ldots, X^9$. For each model, determine the RMSE, AIC, BIC, and R^2 on the training data set. Additionally, for each model, predict the response for the test data set and determine the corresponding RMSE.

(e) Plot the resultant values of the criteria from (d) over the number of predictors. Discuss the resulting graphs. For which model does the test set RMSE take on its minimum value?

(f) Now, use the `stepwiselm()` function to perform a stepwise selection in order to get the best regressor subset out of X, $X^2, \ldots, X^9$. What is the best model according to different model performance criteria (AIC, BIC, R^2)?

2. Getting Familiar with Regularization in Matlab

In this exercise, you will use the data from exercise 1 to perform model selection using lasso.

(a) Use the same set up as in exercise 1, (a) to (c), to generate a training data set.

(b) Use the `logspace()` function to generate a sequence of 100 logarithmically spaced values for λ in between 10^2 and 10^{-5} specifying the degree of regularization.

(c) Generate a regression matrix like in exercise 1 (d) containing $X, X^2, \ldots, X^9$. Use the `lasso()` function to estimate models for the given sequence of λ. Perform 10-fold cross validation (CV) to select the optimal value of λ. (Use `[Beta,FitInfo] = lasso()` to get further information about the fits).

(d) Create plots of the 10-fold cross-validation error and the model coefficients, respectively, as a function of λ using the `lassoPlot()` function. Use a logarithmical scale of the x-axis in the plot of the coefficients. Discuss the resultant graphs and pick the model with the minimum MSE regarding the CV.

(e) Perform a ridge regression by using the `ridge()` function for a sequence of 100 logarithmically spaced values for λ in between 10^4 and 10^{-5} (set the scaled value in the `ridge()` function to zero). Plot the resultant coefficients as a function of lambda using `semilogx()`. Furthermore, invert the x-axis by the use of `set(gca, 'XDir','reverse')`. Compare the results with the results obtained with lasso.

(f) Finally, estimate a full model using least squares. Plot the resultant coefficients together with the coefficients obtained with lasso and ridge for the value of λ obtained in (d). Additionally, determine the RMSE for each model on the test data and discuss the results.

3. Model Selection for mpg Data Set

In this exercise, you will apply model selection to the mpg data set.

(a) Load the `carsmall` data set and choose the displacement, weight, horsepower, and acceleration as potential input variables to predict mpg. Store all variables in a matrix and use `data(any(isnan(data),2),:) = []` to get rid of NaN values. Plot the correlation between predictors and output and determine the corresponding correlation coefficients using `corrcoef()`. What can be concluded from the results?

(b) Try the selection approaches explored in the former exercises, such as stepwise regression and lasso, to find a parsimonious linear model. Present and discuss the results for the approaches that you consider. Additionally, estimate the full model

and compare the resulting models regarding their RMSE, AIC, and BIC values as well as the RMSE obtained using a 10-fold cross-validation.

(c) Now we are interested in the dependency of mpg just on horsepower. Use the selection approaches to find a suitable polynomial model for prediction.