Statistical Learning **Homework assignment 1**
Grün **SS 2024**

1 Introduction and penalized regression

Exercise 1:

Use the diabetes data set to fit different linear models. The data set is available in the R package **lars** and can be loaded using:

```
> data("diabetes", package = "lars")
```

The dependent variable is contained in `diabetes$y`, the model matrix in `diabetes$x` for the linear regression.

- Set a random seed and split the data set into a training and test data set such that 400 observations are used for training and the remaining ones for testing. Explain why it might be good to randomly select 400 observations from the available data set instead of using the first 400.
- The covariates are only available in standardized form. Explain if this is an issue for the subsequent analysis.
- Analyze the pairwise correlation structure between the covariates as well as the covariates and the dependent variable. Interpret the results and explain how these correlations impact model selection.
- Fit a linear regression model containing all explanatory variables. Inspect the model and evaluate the in-sample fit as well as the performance on the test data based on the mean squared error (MSE).
- Fit a smaller model where only the covariates are contained which according to a t-test are significant at the 5% significance level conditional on all other variables being included. Evaluate the performance in-sample as well as on the test data. Compare this model to the full model using an F-test.
- Use stepwise regression based on the AIC to select a suitable model. Evaluate the performance in-sample as well as on the test data and compare this model to the full model using an F-test.
- Use best subset selection to select a suitable model based on the AIC. Evaluate the performance in-sample as well as on the test data and compare this model to the full model using an F-test.

Summarize the results in a table containing the regression coefficients of the different models as well as the in-sample and the test data performance.

Exercise 2:

We use the wage data set to fit different linear models. The data set is available in the R package **ISLR2** and can be loaded using:

```
> data("Wage", package = "ISLR2")
```

- Fit a linear regression model to predict `Wage`. Omit the variable `logwage` before analysis, specify non-linear effects for the variable `age` and use suitable contrasts for the variable `education`.
- Use best subset selection to determine a suitable model.
- Assess and explain if it makes a difference if you include the polynom of the original variable `age` or orthogonal polynoms constructed using `poly(age, k)`.

Exercise 3:

Assume the following data generating process:

$$y = x + x^2 + \epsilon,$$

with $\epsilon \sim N(0, \sigma_\epsilon^2)$ and $x \sim N(0, \sigma_x^2)$ and x and ϵ independent.

- Determine analytically the test error using the squared error loss given parameter estimates $\hat{\boldsymbol{\beta}}$.
- Assume $\sigma_\epsilon^2 = \sigma_x^2 = 1$. Draw a sample of size $N = 40$ as training data and determine the test error using the squared error loss using the analytical formula as well as simulation when estimating the regression coefficients $\boldsymbol{\beta}$ using OLS.
- Assume $\sigma_\epsilon^2 = \sigma_x^2 = 1$. Estimate the expected test error for samples of size $N = 40$ used as training data for the squared error loss when estimating the regression coefficients $\boldsymbol{\beta}$ using OLS and test data sets of suitable size drawn from the data generating process.

Exercise 4:

Assume the following data generating process:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

with $\epsilon \sim N(0, \sigma_\epsilon^2 \boldsymbol{I})$ and a fixed covariate matrix $\boldsymbol{X}$ with N rows and p columns.

- Determine analytically the in-sample error using the squared error loss given parameter estimates $\hat{\boldsymbol{\beta}}$.
- Determine analytically the expected in-sample error using the squared error loss and OLS estimates for the regression coefficients.

Exercise 5:

Use artificial data to perform LASSO and ridge regression. Set a random seed before the analysis.

- Draw 100 observations from a 100-dimensional standard multivariate normal distribution. This is the matrix of covariates `X` of dimension 100 × 100.
- Draw 100 observations for the dependent variable given by

  $$y = \sum_{i=1}^{10} x_i + \epsilon,$$

  with $\epsilon \sim N(0, 0.1)$.
- Fit LASSO and ridge models with different values of λ using function `glmnet` from package **glmnet**.

  Note: Note that the default is `intercept = TRUE`. Keep this default to fit a model including an intercept to `X` and `y`.
- Create the default plots for the returned objects and interpret them. Create also the plots where the argument `xvar` is set to `"lambda"` and interpret them. Point out the specific differences between the solutions obtained for LASSO and ridge regression.
- Determine the number of non-zero coefficients and the model fit as measured by the `deviance()` (= RSS) in dependence of λ for LASSO and ridge. Visualize these results and comment on them.

  Note: You can use `predict(fit, type = "nonzero")` to obtain the indices of variables with non-zero coefficients for the `fit` object returned by `glmnet()` which contains the λ sequence used in `fit$lambda`.