Statistical and Machine Learning **Exercise 3**
Grün, Bajons **WS 2025/26**

3 Model assessment and selection

Sub-task 1:

In the following we assume that we have a fixed design for the inputs, i.e., $\boldsymbol{X} \in \mathbb{R}^{N \times p}$ is deterministic.

The dependent variable is assumed to result from the following data generating process

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

where $\boldsymbol{y} \in \mathbb{R}^N$, $\beta \in \mathbb{R}^p$ and

$$\mathrm{E}[\boldsymbol{\epsilon}] = \boldsymbol{0}, \qquad \mathrm{Var}[\boldsymbol{\epsilon}] = \sigma^2 \boldsymbol{I}_N,$$

with $\boldsymbol{I}_N$ the identity matrix of dimension N.

For a given training data set $\mathcal{T}$ of size N assume that the OLS estimate for the regression coefficients is given by

$$\hat{\beta}_{\mathcal{T}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}_{\mathcal{T}}.$$

Determine the following:

- Determine the expected training error:

$$\mathrm{E}_{\mathcal{T}} \left[\frac{1}{N} (\boldsymbol{y}_{\mathcal{T}} - \boldsymbol{X}\hat{\beta}_{\mathcal{T}})^\top (\boldsymbol{y}_{\mathcal{T}} - \boldsymbol{X}\hat{\beta}_{\mathcal{T}}) \right].$$

- Determine the expected in-sample error:

$$\mathrm{E}_{\mathcal{T}} \left[\mathrm{E}_{\boldsymbol{y}} \left[\frac{1}{N} (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\mathcal{T}})^\top (\boldsymbol{y} - \boldsymbol{X}\hat{\beta}_{\mathcal{T}}) | \mathcal{T} \right] \right].$$

- Determine the difference between the expected training error and the expected in-sample error.

Sub-task 2:

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 estimate the regression coefficients $\hat{\boldsymbol{\beta}}$ using OLS. Determine the test error for $\hat{\boldsymbol{\beta}}$ using the squared error loss based on the analytical formula as well as estimate the test error for $\hat{\boldsymbol{\beta}}$ based on simulation.

Sub-task 3:

Consider the in-sample prediction error $\mathrm{Err}_{\mathrm{in}}$ and the training error $\overline{\mathrm{err}}$ in the case of squared-error loss:

$$\mathrm{Err}_{\mathrm{in}} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{E}_{Y^0}[(Y_i^0 - \hat{f}(x_i))^2]$$

$$\overline{\mathrm{err}} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2.$$

Establish that the average optimism in the training error is

$$\frac{2}{N}\sum_{i=1}^{N} \mathrm{Cov}(\hat{y}_i, y_i).$$

Sub-task 4:

Assume for N observations $\boldsymbol{y} = (y_1, \ldots, y_N)$ the following model: They are drawn i.i.d. from a Bernoulli distribution with parameter θ. Further assume that the prior distribution for θ is a Beta distribution with parameters $\alpha > 0$ and $\beta > 0$:

$$p(\theta|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1},$$

with $\Gamma(\cdot)$ the gamma function.

- Determine the marginal likelihood

$$p(\boldsymbol{y}|\mathcal{M}) = \int_0^1 \exp(\ell(\boldsymbol{y}|\theta))p(\theta|\alpha, \beta)d\theta.$$

  where $\ell(\boldsymbol{y}|\theta)$ is the log-likelihood of the data assuming that the observations are i.i.d. data from a Bernoulli distribution with parameter θ.

- Evaluate -2 times the exact value of the logarithmized marginal likelihood using the uniform prior (i.e., $\alpha = \beta = 1$) for samples with varying size $N \in \{10, 100, 1000, 10000\}$ where the success probability in each sample is equal to 0.1, i.e., the number of successes in the samples is $\{1, 10, 100, 1000\}$.

Sub-task 5:

Assume for N observations $\boldsymbol{y} = (y_1, \ldots, y_N)$ the following model: They are drawn i.i.d. from a Bernoulli distribution with parameter θ.

- Determine an approximation of the marginal likelihood using a uniform prior based on the Laplace approximation given by

$$p(\boldsymbol{y}|\mathcal{M}) \approx \exp(\ell(\boldsymbol{y}|\hat{\theta}))\sqrt{\frac{2\pi}{\mathcal{J}(\hat{\theta})}},$$

  where $\ell(\boldsymbol{y}|\theta)$ is the log-likelihood of the data assuming that the observations are i.i.d. data from a Bernoulli distribution with parameter θ, $\hat{\theta}$ is the maximum likelihood estimate and $\mathcal{J}(\hat{\theta})$ is the observed information matrix, i.e., the negative second derivative of the log-likelihood function evaluated at $\hat{\theta}$.

- Determine -2 times the logarithm of the approximation and compare the result to the Bayesian information criterion for this model.

- Evaluate the BIC for samples with varying size $N \in \{10, 100, 1000, 10000\}$ where the success probability in each sample is equal to 0.1, i.e., the number of successes in the samples is $\{1, 10, 100, 1000\}$.

Sub-task 6:

We will perform leave-one-out cross-validation (LOOCV) and k-fold cross-validation (kCV) on a simulated data set.

(a) Generate a simulated data set as follows:

```
> set.seed(1)
> x <- rnorm(100)
> y <- x - 2*x^2 + rnorm(100)
```

(b) Create a scatterplot of X against Y. Comment on what you find.

(c) Set a random seed, and then compute the LOOCV and kCV errors based on the mean squared error loss that result from fitting the following four models using least squares:

i. $Y = \beta_0 + \beta_1 X + \epsilon$

ii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$

iii. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon$

iv. $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \epsilon$

(d) Repeat (b) using another random seed, and report your results. Are your results the same as what you got in (b)? Why?

(e) Which of the models in (b) had the smallest LOOCV and kCV error? Is this what you expected? Explain your answer.

Sub-task 7:

We perform a simulation study to assess how good the performance of the lasso is for variable selection:

- The following data generating process is used:
  - Draw a 100-dimensional vector from a standard multivariate normal distribution.
  - Determine the dependent variable with $\epsilon \sim N(0, 0.1)$ (i.e., $\sigma^2 = 0.1$) by:

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

- Draw 100 data sets of size 1000. Split each data set into a training data set containing the first 100 observations and a test data set containing the remaining 900 observations. For each of the 100 repetitions use `glmnet` from the **glmnet** package to fit the lasso model for different values of λ to the training data set and select the λ value where predictive performance is best on the test data set.
- Determine the proportion of correctly included coefficients from all relevant ones (true positive rate) and the proportion of wrongly included coefficients from all irrelevant ones (false positive rate) for each of the 100 data sets and visualize the distribution of the two rates.

Sub-task 8:

Use the `Wage` data set from package **ISLR2** to fit penalized linear regression models.

(a) Omit the variables `logwage` and `region`.

For all categorical variables except for `education` ensure that the baseline / reference level corresponds to the mode category, e.g., `"Married"` for variable `maritl`.

Rescale `year` to have 0 correspond to year 2000.

(b) Split the data into a train and test data where the observations from the last year in the sample are used for testing and all other observations for training.

(c) Create the model matrix for the train and test data using the available covariates with

- difference contrasts for the ordinal variable `education`
- orthogonal polynoms of order 4 for `age`.

To obtain difference contrasts use for example `contr.sdif` from package **MASS**.

(d) Fit a ridge and a lasso regression model with 10-fold cross-validation and visualize the results. Ensure that the coefficient for `year` is not penalized.

Compare the estimated coefficients for ridge and lasso when λ is selected with the $1 - \text{SE}$ rule. Assess which coefficients are selected by lasso and interpret the results.

(e) Determine the mean-square error for the ridge and the lasso model on the test data using the λ obtained for the minimum and the $1 - \text{SE}$ rule and compare the results.

(*Hint*: Functions `coef()` and `predict()` can be used with objects returned by `cv.glmnet` and have an argument `s` which can be specified as `"lambda.min"` and `"lambda.1se"` to select the λ value where the loss is minimized or within one standard error.)