Exercise Sheet 4

Due: 22.11.2023, 10:00

Download the files **f41.csv, f42train.csv, f42test.csv, f43.csv** from ISIS. The last column of each file contains output values, all other columns are input features. Download the file E04_template.ipynb from ISIS.

Exercise 4.1

The goal of this exercise is to investigate the generalization error in dependence of the number of training examples.

Consider the following function

$$f: [-1,1]^2 \to \mathbb{R}, \quad \boldsymbol{x} = (x_1, x_2) \mapsto x_1 \sin(\pi x_2)$$

We define the following single experiment $E(m)$ for number of training examples m:

- Generate a training set with m examples (*1)
- Use multiple linear regression to fit a linear model to the training data
- Compute the training error of the fitted model
- Estimate the generalization error of the fitted model using the test data f41.csv

(*1) Generate training data by drawing independent samples $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_m}$ from the uniform distribution on $[-1,1]^2$ and computing the corresponding output values $y_i = f(\boldsymbol{x_i})$.

For each $m = 2, 3, 4, \dots, 80$ conduct the single experiment $E(m)$ 100 times and plot the average training and test errors in dependence of m. Discuss the results.

Exercise 4.2

In this exercise, we investigate effects of using a training set for cross-validation instead of using it for training only. Conduct the following experiment:

1. For each $k = 1,2,3, \dots, 10$ perform polynomial regression of order k on the f42train-data. Compute the test-MSE for each of the 10 models on the f42test-data. The test data is densely sampled without noise.
2. For each $k = 1,2,3, \dots, 10$ conduct 10-fold cross validation on the f42train-data using polynomial regression of order k.

Plot both the test-MSE and the cross validation error in dependence of the order k in a single plot. State and discuss your observations.

Exercise 4.3

The goal of this exercise it to estimate the generalization error of the following learning method on the f43-data: The learning method uses cross validation to select a suitable regularization parameter and applies polynomial regression of order $k = 10$ with L_2 -regularization. Estimate the generalization error using nested cross validation with 5 folds for both, inner and outer cross validation. Present and discuss your results.