Statistical Learning **Homework assignment 4**
Grün **SS 2024**

4 Deep learning with neural networks

Exercise 1:

Generate data from the following additive error model $Y = f(X_1, X_2) + \epsilon$:

- Sum of sigmoids:

$$Y = \sigma(a_1^\top X_1) + \sigma(a_2^\top X_2) + \epsilon,$$

  with

$$a_1 = (3, 3), \qquad a_2 = (3, -3).$$

  - Each X_j, $j = 1, 2$, is a standard Gaussian variate with $p = 2$.
  - ϵ is an independent Gaussian error with variance chosen such that the signal-to-noise ratio as measured by the respective variances equals four.

- Generate a training set of size 100 and a test sample of size 10,000.
- Fit neural networks with weight decay of 0.0005 and varying the number of hidden units from 0 to 10 and record the average test error

$$\mathrm{E}_{\mathrm{Test}}(Y - \hat{f}(X_1, X_2))^2$$

  for each of 10 random starting weights. I.e. for each of the settings with a different number of hidden units fit a neural network 10 times with different initial values.
- Visualize the results and interpret them.

Exercise 2:

The data sets `zip.train` and `zip.test` from package **ElemStatLearn** contain the information on the gray color values of the pixels on a 16×16 pixel image of hand-written digits.

- Visualize for each digit one randomly selected observation (see `?zip.train`).
- Fit a multinomial logistic regression model to the training data and evaluate it on the training and the test data. Determine the overall misclassification rate on the training and the test data and the digit-specific misclassification rates on the test data. Which digits are the most difficult and the easiest to classify?
- Add a positive weight decay of 0.05 when fitting the multinomial logistic regression model to the training data and evaluate the model on the training and the test data. Determine the overall misclassification rate on the training and the test data. Explain why it makes sense to also include weight decay when fitting this multinomial logistic regression model.

Exercise 3:

The data sets `zip.train` and `zip.test` from package **ElemStatLearn** contain the information on the gray color values of the pixels on a 16×16 pixel image of hand-written digits.

- Use only a subset from `zip.train` of size 320 observations with an equal number of observations for each digit to fit a multinomial logistic regression model and a neural network.
- Use all remaining training observations and the test data set to evaluate the fitted models.
- For this small training data overfitting is an issue. Visualize the performance on the test data in dependence of the training epochs when fitting the models.

Exercise 4:

In the following we will estimate a predictive model for the `Default` data from the **ISLR2** pacakge.

- Fit a neural network using a single hidden layer with 10 units and dropout regularization.
- Compare the classification performance of this model with that of linear logistic regression.

Hint: See James et al. (2021, Chapter 10).

Exercise 5:

Consider the `IMDb` dataset from the **keras** package to perform document classification. Restrict the vocabulary to the most frequently-used words and tokens.

- Fit a fully-connected neural network with two hidden layers, each with 16 units and ReLU activation to the data with dictionary size 1000.
- Consider the effects of varying the dictionary size. Try the values 500, 1000, 3000, 5000, and 10,000, and compare the results.

Hint: See James et al. (2021, Chapter 10).