Machine Learning

Spring Semester 2023
Prof. Dr. Peter Zaspel

Assignment Sheet 8.

Submit on **Tuesday, April 11, 2023, 10:00**.

Excercise 1. (Bias-variance trade-off repetition)

In this exercise, we repeat the general idea of the bias-variance trade-off for growing model complexity. To this end, you are asked to provide one plot that gives curves for the irreducible error, the bias, the variance and the total (expected) generalization error for growing model complexity. *(You might have seen such a plot in the lecture material.)* In addition, comment in your own words the "behavior" of the four different curves. Again, in your own words, explain, in connection to these curves, the idea of the bias-variance trade-off.

(4 Points)

Excercise 2. (Bias-variance decomposition for the linear model)

Recall from the lecture that the general bias-variance decomposition is given by

$$EGE(f, \mathbf{x}_0) = \sigma_\varepsilon^2 + \left[\mathrm{E}_T\left(f_T(\mathbf{x}_0)\right) - f_{exact}(\mathbf{x}_0)\right]^2 + \mathrm{E}_T\left(\left(f_T(\mathbf{x}_0) - \mathrm{E}_T\left(f_T(\mathbf{x}_0)\right)\right)^2\right) .$$

While in the lecture notes, we discuss this decomposition more concretely for kNN regression, we do not discuss it for linear regression by least squares. Since deriving the resulting decomposition for linear regression by least squares is a bit involved, we will not do this here. Instead, we want to compute the bias and the variance term for concretely given data and the linear model trained via the least squares estimator.

To achieve this, we first assume to have $f_{exact}(x) = x^2$. Moreover, we consider the four samples $\mathcal{T}_1 = \{x_i^{(1)}, y_i^{(1)}\}_{i=1}^3$, $\mathcal{T}_2 = \{x_i^{(2)}, y_i^{(2)}\}_{i=1}^3$, $\mathcal{T}_3 = \{x_i^{(3)}, y_i^{(3)}\}_{i=1}^3$, $\mathcal{T}_4 = \{x_i^{(4)}, y_i^{(4)}\}_{i=1}^3$ from T with $N = 3$ training samples. These are given as follows

t	i	$x_i^{(t)}$	$\varepsilon_i^{(t)}$	$y_i^{(t)} = f_{exact}(x_i^{(t)}) + \varepsilon_i^{(t)}$
1	1	1.0	−0.1	
1	2	−0.5	0.0	
1	3	3.0	0.2	
2	1	−2.0	0.3	
2	2	−1.5	−0.2	
2	3	0.5	0.1	
3	1	2.0	0.3	
3	2	1.0	0.1	
3	3	−3.0	0.2	
4	1	−1.0	0.3	
4	2	−1.5	0.0	
4	3	2.5	−0.2	

Finally we choose $x_0 = 0$.

a) Compute the output samples of the given training sets, i.e. fill in the remaining cells in the above table.

b) Now that you have all four training sets at hand, compute an estimator for the bias term. To this end, you first build for all four different training sets the linear model using linear regression by least squares, hence you obtain models $f_{\mathcal{T}_1}$, $f_{\mathcal{T}_2}$, $f_{\mathcal{T}_3}$, $f_{\mathcal{T}_4}$. You then need to estimate the expectation, recalling that for a random variable Z an estimator for its mean is given by

$$\mathrm{E}\left(Z\right) \approx \overline{Z} = \frac{1}{M}\sum_{i=1}^{M} z_i ,$$

where the $z_1, \ldots z_M$ are M samples drawn from that random variable. *Hint: Use a computer to find the necessary partial results.*

c) Finally estimate the variance term. Here, it is useful to further observe that the variance term is indeed the variance of $f_T(\mathbf{x_0})$ with respect to the random variable T, hence

$$\mathrm{E}_T\left(\left(f_T(\mathbf{x}_0) - \mathrm{E}_T\left(f_T(\mathbf{x}_0)\right)\right)^2\right) = \mathrm{Var}_T\left(f_T(\mathbf{x_0})\right) .$$

Then use that an unbiased estimator for the variance of a random variable Z is given by

$$\mathrm{Var}\,(Z) \approx \frac{1}{M-1}\sum_{i=1}^{M}(z_i - \overline{Z})^2\,.$$

Hint: Use a computer to find the necessary partial results.

(2+3+3 Points)

Programming Exercise 1. (Visualization of variance in kNN predictor)

In this programming exercise, we further study Example 7.2 from the lecture notes with its associated implementation. If we more carefully read the example, we observe that in the process of the evaluation of the different error contributions, we build $N_T = 100$ different kNN regression predictors for $k = 1, \dots 20$. Note however, that these predictors are only build from the training data $\{(x_i, f_{exact}(x_i)\}_{i=1}^{N}$ and not from $\{(x_i, f_{exact}(x_i) + \varepsilon_i)\}_{i=1}^{N}$.

a) Start by extending the existing implementation such that you also build all these predictors for the training data with the additional noise term.

b) Now, fix $k = 3$ and add a plot that contains the evaluation of all these predictors (with included noise term). Thus you obtain a total of $N_T = 100$ different curves. Try to use some appropriate coloring and transparency to make them properly visible.

   The plot that you will then have created is a visualization of the variance of the predictor.

Reference solutions will only be provided in Python+Matplotlib. The submission format for Python is a Jupyter notebook. The submission format for C/C++ is standard source files. Choose an appropriate format for the Gnuplot-related submission.

(4 Points)