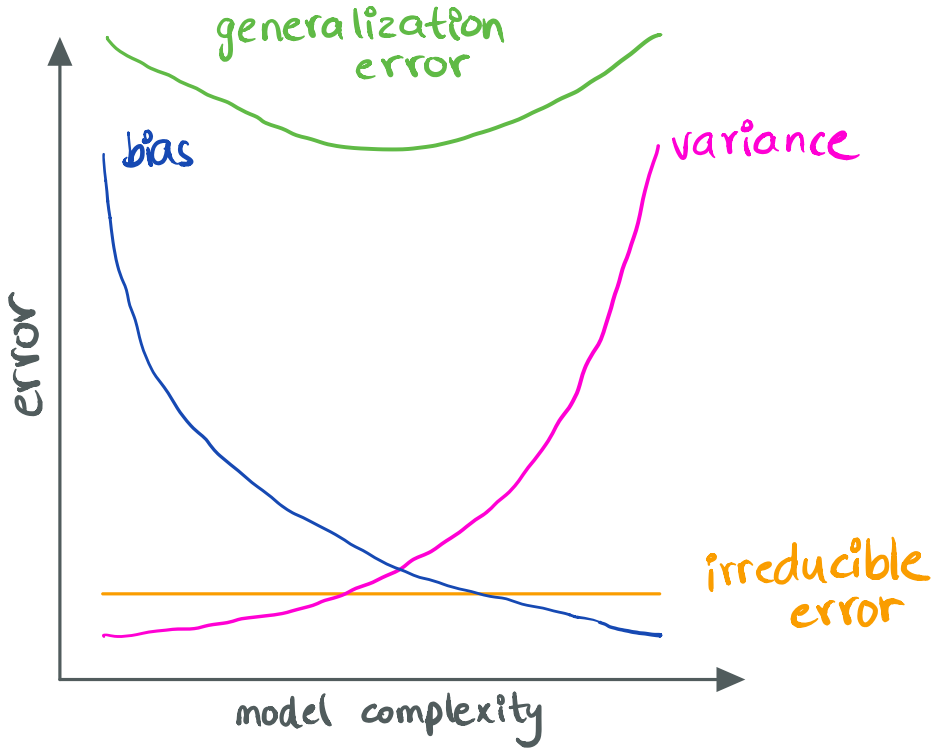
Assignment Sheet 8

Nayan Man Singh Pradhan

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)



The above graph is a bias-variance trade-off graph for growing model complexity. The graph aims to show how different magnitudes of model complexity affects the irreducible error, bias, variance, and generalization error. In order to show this, the graph has model complexity on it's x-axis and error on it's y-axis.

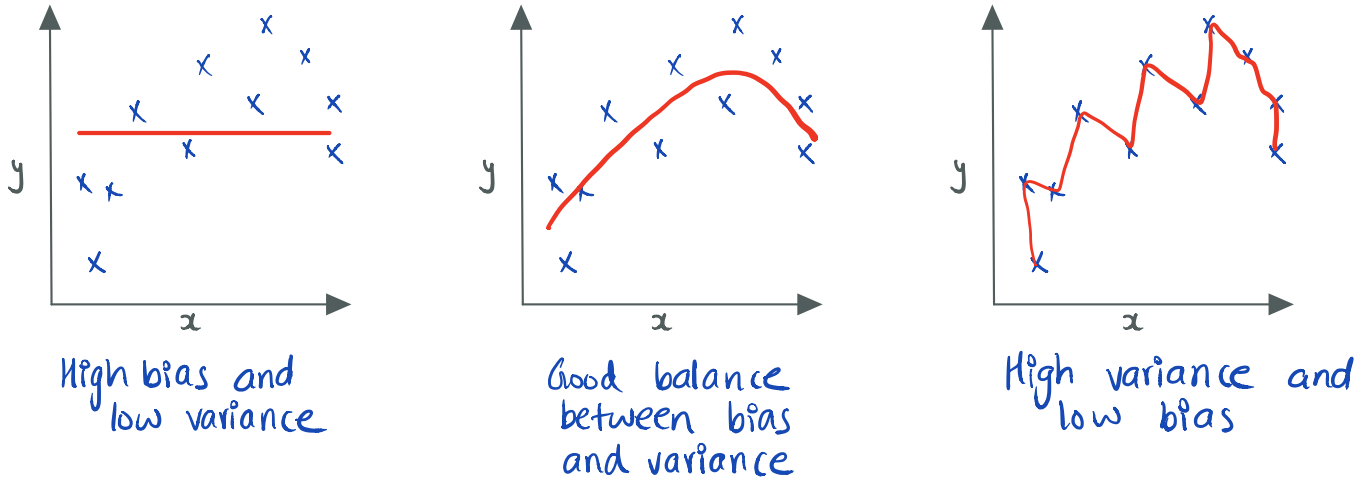
The irreducible error is the error caused by noise in the data. The irreducible error is a linear line parallel to the x-axis. This is because the irreducible error does not change for any value of model complexity. This noise term in given by (\sigma)^2.

Bias is the error caused due to under-fitting of data and Variance is the error caused due to over-fitting of data. When the model complexity is low, there exists high bias and low variance. This is because when the model complexity is low, the predictor function is comparatively simple (low polynomial and less parameters). Having a simple model is more likely to simplify the assumptions made by the model, hence causing under-fitting of data. When the model complexity is high, there exists low bias and high variance. This is because when the model complexity is high, the predictor function is comparatively complex (high polynomial and more parameters). Having a complex model is more likely to strongly fit the data points in the model, hence causing over-fitting of data.

The generalization error is simply the sum of all the errors at the given model complexity. Therefore, the generalization error will always be greater than or equal to the individual errors (bias, variance, and irreducible error). We notice that the generalization error is high for low model complexity (due to the high bias) and also high for high model complexity (due to high variance). The generalization error is comparatively less when the model complexity is in the middle because the sum of the errors for bias, variance, and irreducible error is the minimum when the model complexity is around the middle.

Therefore, the bias-variance trade-off corresponds to the observation that we will never be able to minimize both the bias and variance together. When a model has high bias, it has low variance, and when a model has low bias, it has high variance. Therefore, instead of picking a model with either low variance or low bias, it is better to make a trade-off between both error contributions and chose a model with low generalization error!

A simple example of a model with high bias and low variance, good balance between bias and variance, and high variance and low bias is illustrated below:



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 $f_{exact}(x) = x^3$. Moreover, we consider the four samples $\mathcal{T}_1 = \{x_i^{(1)}, y_i^{(1)}\}_{i=1}^3$, $\mathcal{T}_2 = \{x_i^{(2)}, y_i^{(2)}\}_{i=1}^3$, $\mathcal{T}_1 = \{x_i^{(3)}, y_i^{(3)}\}_{i=1}^3$, $\mathcal{T}_1 = \{x_i^{(4)}, y_i^{(4)}\}_{i=1}^3$ 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	0.1	
1	3	1.0	0.0	
2	1	−0.5	0.1	
2	2	0.5	0.2	
2	3	1.0	0.0	
3	1	0.1	0.0	
3	2	0.5	−0.1	
3	3	1.0	0.1	
4	1	−0.1	−0.1	
4	2	−0.5	0.0	
4	3	−1.0	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.

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	-1.1
1	2	−0.5	0.1	-0.025
1	3	1.0	0.0	1
2	1	−0.5	0.1	-0.025
2	2	0.5	0.2	0.325
2	3	1.0	0.0	1
3	1	0.1	0.0	0.001
3	2	0.5	−0.1	0.025
3	3	1.0	0.1	1.1
4	1	−0.1	−0.1	-0.101
4	2	−0.5	0.0	-0.125
4	3	−1.0	0.2	-0.8

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

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

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.*

Building the linear model using linear regression by least squares for $\mathcal{T}_1$.

$$X = \begin{pmatrix} 1 & -1 \\ 1 & -0.5 \\ 1 & 1 \end{pmatrix}, \quad X^T = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -0.5 & 1 \end{pmatrix}, \quad y = \begin{pmatrix} -1.1 \\ -0.025 \\ 1 \end{pmatrix}$$

$$X^T X = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & -0.5 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 3 & -0.5 \\ -0.5 & 2.25 \end{pmatrix}$$

$$X^T y = \begin{pmatrix} 1 & 1 & 1 \\ -1 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} -1.1 \\ -0.025 \\ 1 \end{pmatrix} = \begin{pmatrix} -0.125 \\ 2.1125 \end{pmatrix}$$

$$X^T X \hat{\beta} = X^T y$$

$$\begin{pmatrix} 3 & -0.5 \\ -0.5 & 2.25 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \begin{pmatrix} -0.125 \\ 2.1125 \end{pmatrix}$$

$$\hat{\beta}_1 = 0.120 \,, \quad \hat{\beta}_2 = 0.965$$

$$f_{\mathcal{T}_1} = 0.119 + 0.965\,\beta$$

I computed the functions $f_{\mathcal{T}_2}$, $f_{\mathcal{T}_3}$ and $f_{\mathcal{T}_4}$ using the previous programming assignments.

$$f_{\mathcal{T}_1} = 0.119 + 0.965\,\beta$$

$$f_{\mathcal{T}_2} = 0.221 + 0.635\,\beta$$

$$f_{\mathcal{T}_3} = -0.296 + 1.259\,\beta$$

$$f_{\mathcal{T}_4} = 0.084 + 0.800\,\beta$$

At point $x_0 = 0$,

$f_{T_1} = 0.119$, $f_{T_2} = 0.221$, $f_{T_3} = -0.296$, $f_{T_4} = 0.084$

$$E = \frac{1}{4}(0.119 + 0.221 - 0.296 + 0.084)$$

$$= 0.032$$

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)

$$\mathrm{Var}_T(f_T(x_0)) = E_T\left((f_T(x_0) - E_T(f_T(x_0)))^2\right)$$

$$= E_T((0 - 0.032)^2)$$

$$= 0.001024 //$$

Now,

$$\mathrm{Var}(z) = \frac{1}{M-1}\sum_{i=1}^{M}(z_i - \overline{z})^2$$

$$= \frac{1}{4-1}\sum_{i=1}^{4}(z_i - \overline{z})^2$$

$$= \frac{1}{3}\left[\left(f_{T_1}(0) - 0.032\right)^2 + \left(f_{T_2}(0) - 0.032\right)^2 + \left(f_{T_3}(0) - 0.032\right)^2 + \left(f_{T_4}(0) - 0.032\right)^2\right]$$

$$= \frac{1}{3} \times 0.153$$

$$= 0.0512 //$$