# FYS-STK4155 Week 36

Edvard B. Rørnes, Isak O. Rukan, Anton A. Brekke

September 18, 2024

## Exercise 1

Show that

$$\mathbb{E}[(\boldsymbol{y}-\tilde{\boldsymbol{y}})^2] = \text{Bias}[\tilde{\boldsymbol{y}}] + \text{Var}[\tilde{\boldsymbol{y}}] + \sigma^2 \tag{1}$$

where $\boldsymbol{y}$ is defined by $\boldsymbol{y} = f(\boldsymbol{x})+\boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0,\sigma^2)$ is a normal distributed error and $f(\boldsymbol{x})$ is the approximated function given our model $\tilde{\boldsymbol{y}}$ obtained by minimizing $(\boldsymbol{y}-\tilde{\boldsymbol{y}})^2$ with $\tilde{\boldsymbol{y}} = \boldsymbol{X}\boldsymbol{\beta}$. Explain what the terms mean and discuss their interpretations. Perform then a bias-variance analysis of a simple one-dimensional function by studying the MSE value as a function of the complexity of your model. Use OLS only. Discuss the bias and variance trade-off as function of your model complexity (the degree of the polynomial) and the number of data points.

**Solution:**

For ease of notation we write $f(\boldsymbol{x}) = f$ and simply ignore vector notation since everything is a scalar in the end. Then we have

$$\begin{aligned}
\mathbb{E}[(\boldsymbol{y}-\tilde{\boldsymbol{y}})^2] &= \mathbb{E}[(f+\boldsymbol{\varepsilon}-\tilde{\boldsymbol{y}})^2] = \mathbb{E}[(f-\tilde{\boldsymbol{y}})^2] + 2\underbrace{\mathbb{E}[(f-\tilde{\boldsymbol{y}})\boldsymbol{\varepsilon}]}_{=0} + \underbrace{\mathbb{E}[\boldsymbol{\varepsilon}^2]}_{=\sigma^2} \\
&= \mathbb{E}[((f-\mathbb{E}[\tilde{\boldsymbol{y}}]) - (\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}]))^2] + \sigma^2 \\
&= \mathbb{E}[(f-\mathbb{E}[\tilde{\boldsymbol{y}}]))^2] + \mathbb{E}[(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}])^2] - 2\,\mathbb{E}[((f-\mathbb{E}[\tilde{\boldsymbol{y}}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}]))] + \sigma^2 \\
&= \text{Bias}[\tilde{\boldsymbol{y}}] + \text{Var}[\tilde{\boldsymbol{y}}] + \sigma^2 - 2\,\mathbb{E}[(f-\mathbb{E}[\tilde{\boldsymbol{y}}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{\boldsymbol{y}}])]
\end{aligned}$$

where $\mathbb{E}[(f-\tilde{\boldsymbol{y}})\boldsymbol{\varepsilon}] = 0$ is justified by $\boldsymbol{\varepsilon}$ being independent and we note that the wrong definition of the Bias is given in the problem text (with that definition $\sigma^2$ gets put into the 'Bias'). All that remains is to show that the last term is 0. Since $\mathbb{E}[f] = f$ and $\mathbb{E}[f\,\mathbb{E}[\tilde{\boldsymbol{y}}]] = f\,\mathbb{E}[\mathbb{E}[\tilde{\boldsymbol{y}}]] = f\,\mathbb{E}[\tilde{\boldsymbol{y}}]$ then

$$\begin{aligned}
\mathbb{E}[(f-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])] &= \mathbb{E}[f\tilde{\boldsymbol{y}} - f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - \tilde{\boldsymbol{y}}\,\mathbb{E}[\tilde{y}] + \mathbb{E}^2[\tilde{\boldsymbol{y}}]] \\
&= f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - f\,\mathbb{E}[\tilde{\boldsymbol{y}}] - \mathbb{E}^2[\tilde{\boldsymbol{y}}] + \mathbb{E}^2[\tilde{\boldsymbol{y}}] = 0
\end{aligned}$$

which proves the claim.

The LHS of (1) is the expected value of the MSE which tells us how well the model's predictions match the true data on average. The equation shows that we can decompose this expected MSE into 3 different components.

- Bias: This quantity measures how much the model's average prediction differs from its true value. A high bias implies that the model is underfitting the data of is simply too simplistic.

- Var: The variance measures how much the model's predictions vary when trained on different datasets. It captures the sensitivity of the model to small changes in the training data. A high variance suggests overfitting, meaning it performs well on the training data but may be capturing noise or false patterns.
- $\sigma^2$: This is the irreducible error or noise in the data itself which cannot be explained by the model.

The idea is to minimize the LHS of (1), so clearly we want to minimize both the bias and the variance at the same time. However these are correlated to one another, so lowering the e.g. the bias will in general increase the variance. So Bias-Variance Tradeoff is essentially trying to optimize the complexity of the model such that we neither overfit nor underfit the model such that it can be generalized to other cases. These quantities can then be used as means to fine tune a model.

Since I have already written it in down I just want to quickly show why (I believe at least) the definition given in the problem text is wrong:

$$\begin{aligned}\mathbb{E}[(\boldsymbol{y}-\tilde{\boldsymbol{y}})^2] &= \mathbb{E}[((\boldsymbol{y}-\mathbb{E}[\tilde{y}])-(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}]))^2] \\ &= \mathbb{E}[(\boldsymbol{y}-\mathbb{E}[\tilde{y}])^2]+\mathbb{E}[(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])^2]-2\,\mathbb{E}[(\boldsymbol{y}-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])] \\ &\underbrace{=}_{\text{wrong}} \mathrm{Bias}[\tilde{\boldsymbol{y}}]+\mathrm{Var}[\tilde{\boldsymbol{y}}]-2\,\mathbb{E}[(\boldsymbol{y}-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])]\end{aligned}$$

Then we have

$$\begin{aligned}\mathbb{E}[(\boldsymbol{y}-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])] &= \mathbb{E}[(f+\boldsymbol{\varepsilon}-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])] \\ &= \mathbb{E}[(f-\mathbb{E}[\tilde{y}])(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])]+\mathbb{E}[\boldsymbol{\varepsilon}(\tilde{\boldsymbol{y}}-\mathbb{E}[\tilde{y}])]\end{aligned}$$

I have already shown explicitly that the first term is 0 and the second term is 0 due to the same reasons as above. So with this definition we would get the wrong result that

$$\mathbb{E}[(\boldsymbol{y}-\tilde{\boldsymbol{y}})^2] = \mathrm{Bias}[\tilde{\boldsymbol{y}}]+\mathrm{Var}[\tilde{\boldsymbol{y}}]$$