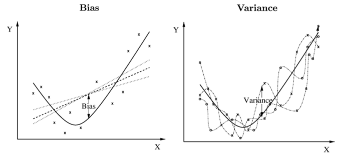
Introduction to Machine Learning

Bias-Variance Decomposition



Learning goals

- Understand how to decompose the generalization error of an inducer into
  - Bias of the inducer
  - Variance of the inducer
  - Noise in the data

BIAS-VARIANCE DECOMPOSITION

Let us take a closer look at the generalization error of a learning algorithm $\mathcal{I}_{L,O}$. This is the expected error an induced model, on trainings sets of size n, when this is applied to a fresh, random test observation.

$$GE_n\left(\mathcal{I}_{L,O}\right) = \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}_{xy}, (\mathbf{x}, y) \sim \mathbb{P}_{xy}}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right) = \mathbb{E}_{\mathcal{D}_n, xy}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)$$

We therefore need to take the expectation over all training sets of size n, as well as the independent test observation.

We assume that the data is generated by

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon\,,$$

with normally distributed error $\epsilon \sim \mathcal{N}(0, \sigma^2)$ independent of $\mathbf{x}$.

BIAS-VARIANCE DECOMPOSITION

By plugging in the $L2$ loss $L(y, f(\mathbf{x})) = (y - f(\mathbf{x}))^2$ we get

$$\begin{aligned} GE_n\left(\mathcal{I}_{L,O}\right) &= \mathbb{E}_{\mathcal{D}_n,xy}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right) = \mathbb{E}_{\mathcal{D}_n,xy}\left(\left(y - \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)^2\right) \\ &= \mathbb{E}_{xy}\underbrace{\left[\mathbb{E}_{\mathcal{D}_n}\left(\left(y - \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)^2 \mid \mathbf{x}, y\right)\right]}_{(*)} \end{aligned}$$

Let us consider the error $(*)$ conditioned on one fixed test observation $(\mathbf{x}, y)$ first. (We omit the $\mid \mathbf{x}, y$ for better readability for now.)

$$\begin{aligned} (*) &= \mathbb{E}_{\mathcal{D}_n}\left(\left(y - \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)^2\right) \\ &= \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(y^2\right)}_{=y^2} + \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})^2\right)}_{(1)} - 2\underbrace{\mathbb{E}_{\mathcal{D}_n}\left(y\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)}_{(2)} \end{aligned}$$

by using the linearity of the expectation.

BIAS-VARIANCE DECOMPOSITION

$$\mathbb{E}_{\mathcal{D}_n}\left(\left(y - \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)^2\right) = y^2 + \underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})^2\right)}_{(1)} - 2\,\underbrace{\mathbb{E}_{\mathcal{D}_n}\left(y\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)}_{(2)} =$$

By using that $\mathbb{E}(z^2) = \mathrm{Var}(z) + \mathbb{E}^2(z)$, we see that

$$= y^2 + \mathrm{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) + \mathbb{E}^2_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) - 2y\mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x}))\right)$$

Plug in the definition of y

$$= f_{\text{true}}(\mathbf{x})^2 + 2\epsilon f_{\text{true}}(\mathbf{x}) + \epsilon^2 + \mathrm{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) + \mathbb{E}^2_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) - 2(f_{\text{true}}(\mathbf{x}) + \epsilon)\mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x}))\right)$$

Reorder terms and use the binomial formula

$$= \epsilon^2 + \mathrm{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) + \left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)^2 + 2\epsilon\left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)$$

BIAS-VARIANCE DECOMPOSITION

$$(*) = \epsilon^2 + \text{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right) + \left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)^2 + 2\epsilon\left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)$$

Let us come back to the generalization error by taking the expectation over all fresh test observations $(\mathbf{x}, y) \sim \mathbb{P}_{xy}$:

$$\begin{aligned} GE_n\left(\mathcal{I}_{L,O}\right) &= \underbrace{\sigma^2}_{\text{Variance of the data}} + \mathbb{E}_{xy}\underbrace{\left[\text{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x}, y\right)\right]}_{\text{Variance of inducer at } (\mathbf{x}, y)} \\ &+ \mathbb{E}_{xy}\underbrace{\left[\left(\left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)^2 \mid \mathbf{x}, y\right)\right]}_{\text{Squared bias of inducer at } (\mathbf{x}, y)} + \underbrace{0}_{\text{As } \epsilon \text{ is zero-mean and independent}} \end{aligned}$$

BIAS-VARIANCE DECOMPOSITION

$GE_n\left(\mathcal{I}_{L,O}\right) =$

$$\underbrace{\sigma^2}_{\text{Variance of the data}} + \mathbb{E}_{xy} \underbrace{\left[\mathrm{Var}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x}, y\right)\right]}_{\text{Variance of inducer at } (\mathbf{x}, y)} + \mathbb{E}_{xy} \underbrace{\left[\left(\left(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}\left(\hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right)^2 \mid \mathbf{x}, y\right)\right]}_{\text{Squared bias of inducer at } (\mathbf{x}, y)}$$

1. The first term expresses the variance of the data. This is **noise** present in the data. Also called Bayes, intrinsic or unavoidable error. No matter what we do, we will never get below this error.
2. The second term expresses how the predictions fluctuate on test-points on average, if we vary the training data. Expresses also the learner's tendency to learn random things irrespective of the real signal (overfitting).
3. The third term says how much we are "off" on average at test locations (underfitting). Models with high capacity have low **bias** and models with low capacity have high **bias**.

BIAS-VARIANCE DECOMPOSITION

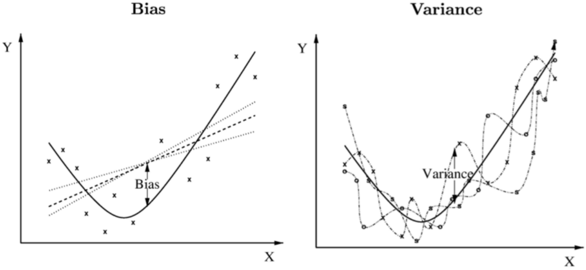


Figure: *Left*: A model with high bias is unable to fit the curved relationship present in the data. *Right*: A model with no bias and high variance can, in principle, learn the true pattern in the data. However, in practice, the learner outputs wildly different hypotheses for different training sets.