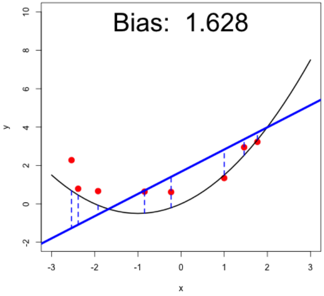
Introduction to Machine Learning

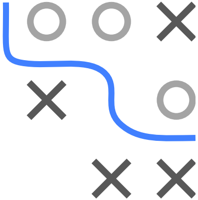
Advanced Risk Minimization
Bias-Variance Decomposition (Deep-Dive)



Learning goals

- Understand how to decompose the generalization error of a learner into
  - Bias of the learner
  - Variance of the learner
  - Inherent noise in the data

BIAS-VARIANCE DECOMPOSITION



Let us take a closer look at the generalization error of a learning algorithm $\mathcal{I}_L$. This is the expected error of an induced model $\hat{f}_{\mathcal{D}_n}$, on training sets of size n, when applied to a fresh, random test observation.

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

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 zero-mean homoskedastic error $\epsilon \sim (0, \sigma^2)$ independent of $\mathbf{x}$.

BIAS-VARIANCE DECOMPOSITION / 2

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

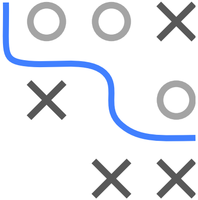
$$
\begin{aligned}
GE_n\left(\mathcal{I}_L\right) &= \mathbb{E}_{\mathcal{D}_n, xy}\left(L\left(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)\right) = \mathbb{E}_{\mathcal{D}_n, xy}\left(\left(y - \hat{f}_{\mathcal{D}_n}(\mathbf{x})\right)^2\right) \\
&\overset{\text{LIE}}{=} \mathbb{E}_{xy}\Bigg[\underbrace{\mathbb{E}_{\mathcal{D}_n}\left(\left(y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

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.)

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

by using the linearity of the expectation.

BIAS-VARIANCE DECOMPOSITION / 3



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

Using that $\mathbb{E}(z^2) = \mathsf{Var}(z) + \mathbb{E}^2(z)$, we see that

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

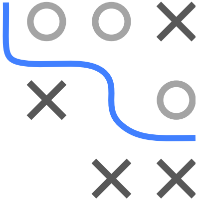
Plug in the definition of y

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

Reorder terms and use the binomial formula

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

BIAS-VARIANCE DECOMPOSITION / 4



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

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

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