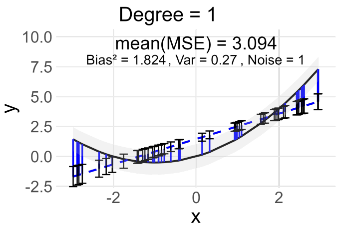
Introduction to Machine Learning

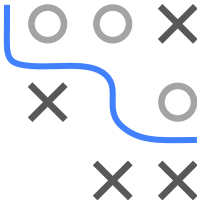
Advanced Risk Minimization
Bias-Variance Decomposition (Deep-Dive)



Learning goals

- Decompose generalization error of a learner specifically for L2 loss into
  - Bias of the learner
  - Variance
  - Inherent noise in the data

BIAS-VARIANCE DECOMPOSITION



Generalization error of learner $\mathcal{I}$: Expected error of model $\hat{f}_{\mathcal{D}_n}$, on training sets of size n, evaluated on a fresh, random test sample

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

Expectation is taken over all training sets **and** independent test sample

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

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

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

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)

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

by using the linearity of the expectation

BIAS-VARIANCE DECOMPOSITION

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

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

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

Plug in the definition of y

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

Reorder terms and use the binomial formula

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

BIAS-VARIANCE DECOMPOSITION

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

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

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