The bias-variance tradeoff is an important aspect of ML

Consider the additive error model $Y = f(X) + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2)$. Here we are thinking about choosing an algorithm to learn the predictive $\hat{f}$.

All learning algorithms use a mathematical approach that contains an 'error' term which can be further split into two components:

This is the noise — term in the true relationship that cannot fundamentally be reduced by the model — irreducible error

Reducible error → should be minimized further to max. accuracy

EPE

$$Err(x_0) = \mathbb{E}(Y_0 - \hat{f}(x_0))^2 = \sigma_\varepsilon^2 + \mathbb{E}(f(x_0) - \hat{f}(x_0))^2$$

$$Y_0 = f(x_0) + \varepsilon_0 \qquad = \sigma^2 + \text{Bias}^2 + \text{Variance}$$

$$\boxed{\text{Claim: } \mathbb{E}(Y_0 - \hat{f}(x_0)) = \sigma^2 + \text{Bias}(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0))}$$

Decompose the expected predictn Error into 3 part irreducible error, squared Bias, variance

Proof:

$$EPE = \mathbb{E}\{(Y_0 - \hat{f}(x_0))^2\} \qquad Y_0 = f(x_0) + \varepsilon_0$$

$$= \mathbb{E}\{(f(x_0) + \varepsilon_0 - \hat{f}(x_0))^2\}$$

$$= \mathbb{E}\{\varepsilon_0^2 + 2\varepsilon_0(f(x_0) - \hat{f}(x_0)) + (f(x_0) - \hat{f}(x_0))^2\}$$

$\mathbb{E}(\cdot)$ linear

$\mathbb{E}(X+Y+Z) = \mathbb{E}X + \mathbb{E}Y + \mathbb{E}Z$

$$= \underbrace{\mathbb{E}\varepsilon_0^2}_{\text{Var}\,\varepsilon_0 + (\mathbb{E}\varepsilon_0)^2} + 2\mathbb{E}\{\varepsilon_0(f(x_0) - \hat{f}(x_0))\} + \mathbb{E}\{(f(x_0) - \hat{f}(x_0))^2\}$$

$\mathbb{E}(XY) = 0$ if $X \perp\!\!\!\perp Y$ as $\mathbb{E}X = 0$

$$= \sigma^2 + 0 + \mathbb{E}\{(f(x_0) - \hat{f}(x_0))^2\}$$

$$\# \quad \mathbb{E}\{(f(x_0) - \hat{f}(x_0))^2\} = \mathbb{E}\{(f(x_0) - \mathbb{E}\hat{f}(x_0) + \mathbb{E}\hat{f}(x_0) - \hat{f}(x_0))^2\}$$

$$= \mathbb{E}\{(f(x_0) - \mathbb{E}\hat{f}(x_0))^2 + 2(f(x_0) - \mathbb{E}\hat{f}(x_0))(\mathbb{E}\hat{f}(x_0) - \hat{f}(x_0)) + (\mathbb{E}\hat{f}(x_0) - \hat{f}(x_0))^2\}$$

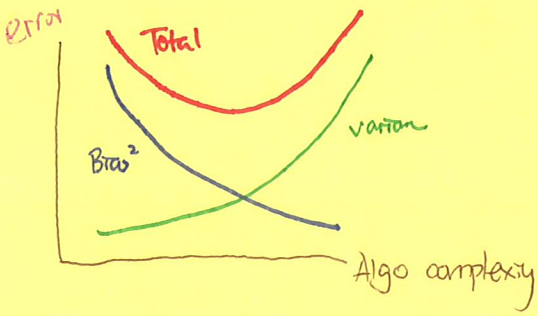
$$= (f(x_0) - \mathbb{E}\hat{f}(x_0))^2 + \underbrace{2\mathbb{E}(\quad)(\quad)}_{0} + \mathbb{E}(\mathbb{E}\hat{f}(x_0) - \hat{f}(x_0))^2$$

$$\mathbb{E}\{(f(x_0) - \hat{f}(x_0))^2\} = (\mathbb{E}\hat{f}(x_0) - f(x_0))^2 + \mathbb{E}\{(\hat{f}(x_0) - \mathbb{E}\hat{f}(x_0))^2\}$$

from train set — error due to overgeneralisation data — error due to fluctuation in the training dataset — from test set

$$= \text{Bias}(\hat{f}(x_0)) + \text{Var}(\hat{f}(x_0))$$

this is often called Bias-variance trade off because improving one will worsen the other



In ML, the ideal algorithm has low bias and can accurately model the true relationship and has low variability, by producing consistent prediction across different datasets.