Overfitting / Underfitting

Damien Benveniste

October 23, 2017

1 Bias-variance trade off

We assume that we can relate features x to a target y through the relation

$$y = f(x) + \epsilon \tag{1}$$

where f is a deterministic function and ϵ is a random noise $\epsilon \sim \mathcal{D}(0, \sigma)$. Machine learning is about finding an estimate $\hat{f}$ of f. Note that ϵ is independent of $\hat{f}(x)$

Lets look at the Mean Square Error metric (MSE). We have

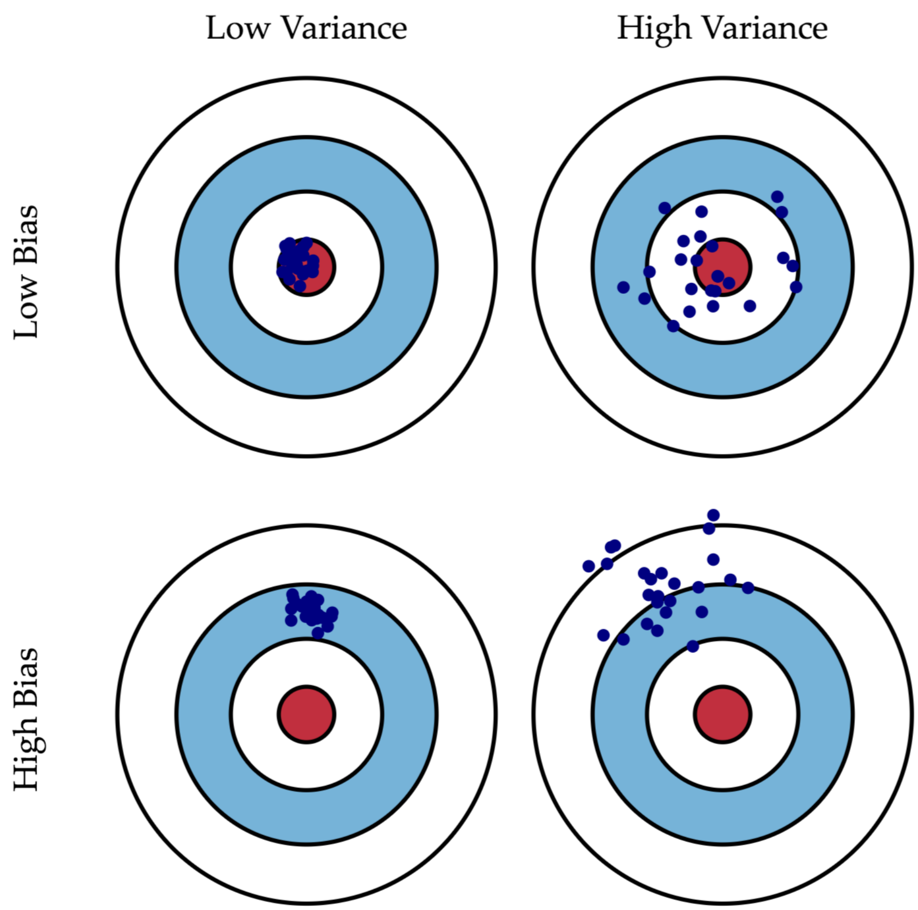
$$MSE = E\left[(y - \hat{f}(x))^2\right] \tag{2}$$

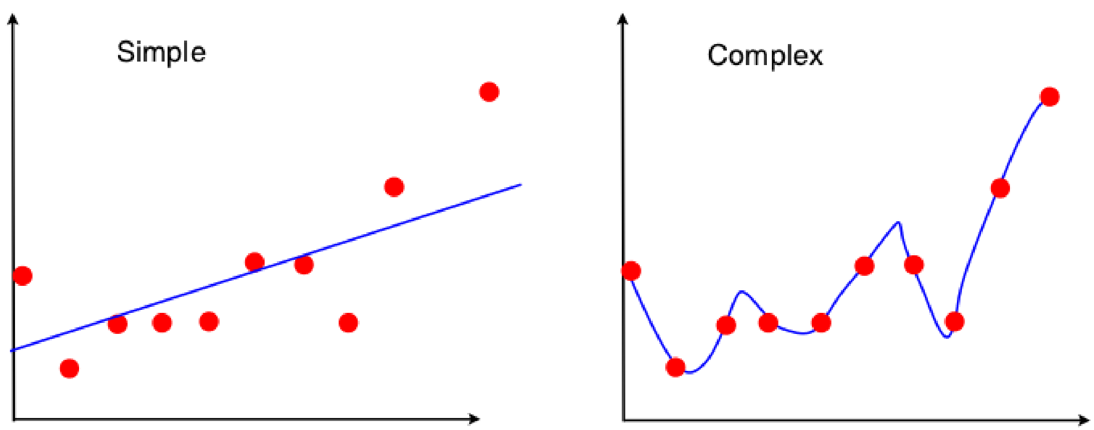
We are going to define

$$Var[\hat{f}(x)] = E\left[\hat{f}(x)^2\right] - E\left[\hat{f}(x)\right]^2 \tag{3}$$

and

$$Bias[\hat{f}(x)] = E\left[\hat{f}(x) - f(x)\right] \tag{4}$$


Low Variance
High Variance
Low Bias
High Bias


Simple
Complex

Therefore

$$
\begin{aligned}
MSE &= E\left[(y-\hat{f}(x))^2\right] \\
&= E\left[y^2\right] + E\left[\hat{f}(x)^2\right] - 2E\left[y\hat{f}(x)\right] \\
&= Var[y] + E\left[y\right]^2 + Var[\hat{f}(x)] + E\left[\hat{f}(x)\right]^2 - 2E\left[y\hat{f}(x)\right] \qquad (5)
\end{aligned}
$$

Because ϵ is independent of $\hat{f}(x)$ we have

$$
\begin{aligned}
E\left[y\hat{f}(x)\right] &= E\left[(f(x)+\epsilon)\hat{f}(x)\right] \\
&= E\left[f(x)\hat{f}(x)\right] + E\left[\epsilon\hat{f}(x)\right] \\
&= f(x)E\left[\hat{f}(x)\right] + E\left[\epsilon\right]E\left[\hat{f}(x)\right] \\
&= f(x)E\left[\hat{f}(x)\right] \qquad (6)
\end{aligned}
$$

also

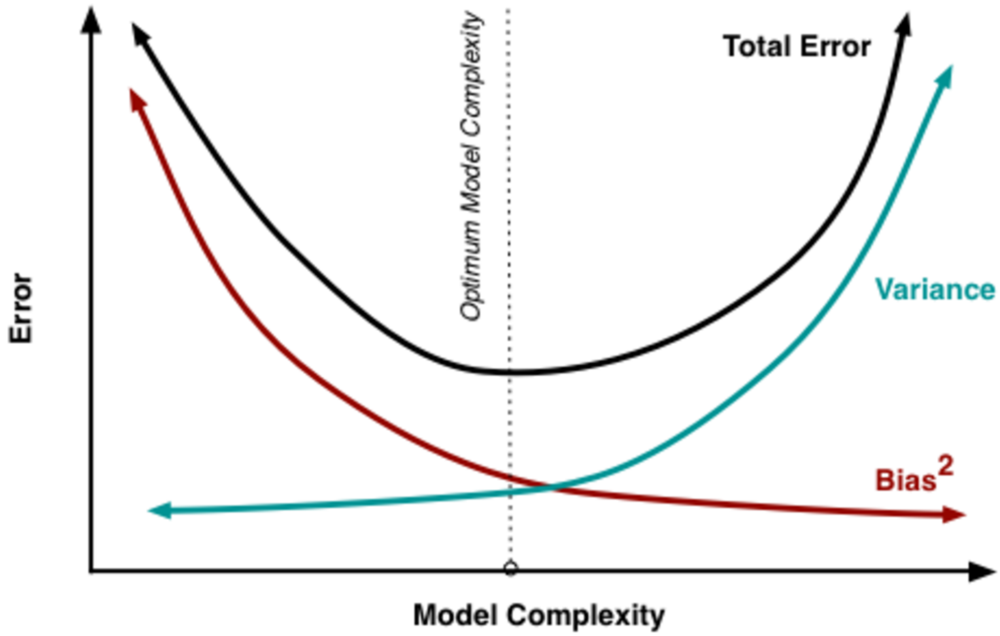
$$E\left[y\right] = f(x) \qquad (7)$$

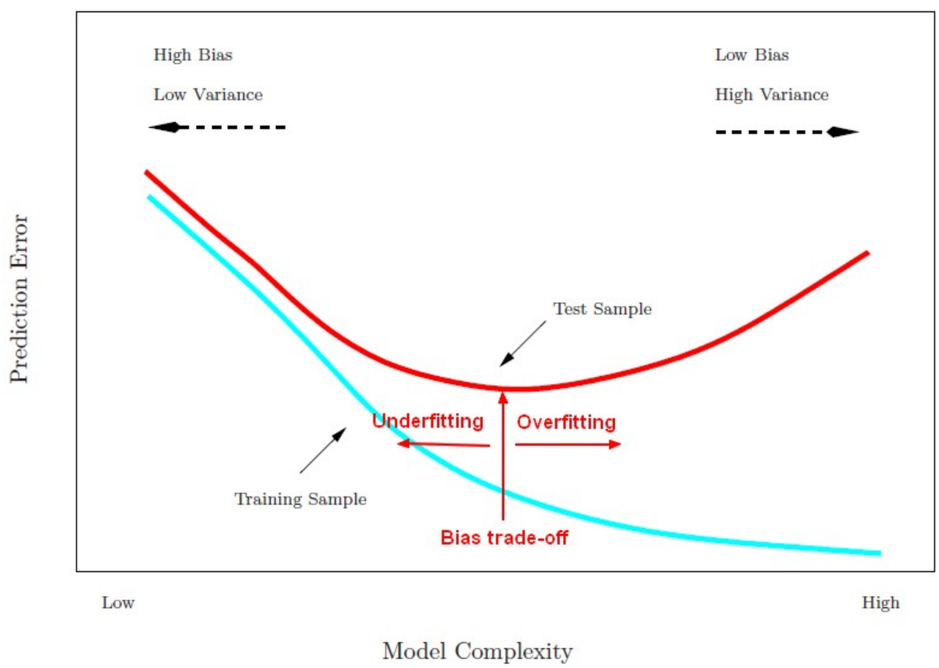
and

$$Var[y] = \sigma^2 \qquad (8)$$

Therefore

$$
\begin{aligned}
MSE &= \sigma^2 + Var[\hat{f}(x)] + f(x)^2 + E\left[\hat{f}(x)\right]^2 - 2f(x)E\left[\hat{f}(x)\right] \\
&= \sigma^2 + Var[\hat{f}(x)] + \left(f(x) - E\left[\hat{f}(x)\right]\right)^2 \\
&= \sigma^2 + Var[\hat{f}(x)] + E\left[f(x) - \hat{f}(x)\right]^2 \\
&= \sigma^2 + Var[\hat{f}(x)] + Bias[\hat{f}(x)]^2 \qquad (9)
\end{aligned}
$$


Total Error
Optimum Model Complexity
Variance
Bias2
Error
Model Complexity


High Bias
Low Variance
Low Bias
High Variance
Test Sample
Underfitting
Overfitting
Training Sample
Bias trade-off
Prediction Error
Low
High
Model Complexity

2 Regularization

3 Cross-validation