Lecture 6 notes

Bias-variance tradeoff in prediction algorithms

Julian Reif
Fall 2025

What is MSE of a prediction algorithm?

$$Y = f(X) + \epsilon$$

where ϵ is irreducible error, i.e. $E[\epsilon] = 0$.

Goal: Use a prediction algorithm $\hat{f}$ to predict outcome Y when $X = x_0$.

Question: What is the mean-squared error of $\hat{f}$?

What does "mean" mean? It is the average outcome from repeating a process.

Squared error process

1. Draw training data sample and estimate $\hat{f}_i$

2. Draw test data sample and form predictions
    - $Y_i = f(x_0) + \epsilon$
    - $\hat{f}_i = \hat{f}_i(x_0)$

3. Calculate squared error: $\left(Y_i - \hat{f}_i\right)^2$

Repeating this process many times and averaging leads to a "mean" squared error

Key assumption: ϵ has mean zero and is independent of X

Review: Properties of expected values

Let X, Y be random variables and α, β constants.

1. Linearity: $E\left[\alpha X + \beta Y\right] = \alpha E\left[X\right] + \beta E\left[Y\right]$

2. If X and Y are independent: $E\left[X \cdot h(Y)\right] = E\left[X\right] \cdot E\left[h(Y)\right]$ for any $h()$

3. Variance decomposition: $Var(\alpha + X) = E\left[X^2\right] - \left(E[X]\right)^2$

MSE decomposition 1

$$
\begin{aligned}
MSE &= E\left[\left(Y - \hat{f}\right)^2\right] \\
&= E\left[\left(Y \underbrace{- f + f}_{=0} - \hat{f}\right)^2\right] \\
&= E\left[\left(\epsilon + f - \hat{f}\right)^2\right] \text{ (recall: } Y = f(x) + \epsilon) \\
&= E\left[\epsilon^2\right] + E\left[\left(f - \hat{f}\right)^2\right] + 2E\left[\epsilon\left(f - \hat{f}\right)\right]
\end{aligned}
$$

where the last equality follows because $(A+B)^2 = A^2 + B^2 + 2AB$

MSE decomposition 2

$$MSE = E\left[\epsilon^2\right] + \underbrace{E\left[\left(f - \hat{f}\right)^2\right]}_{\text{B}} + \underbrace{2E\left[\epsilon\left(f - \hat{f}\right)\right]}_{\text{A}}$$

Let's simplify term A:

$E\left[\epsilon(f - \hat{f})\right] = \underbrace{E[\epsilon]}_{=0} E\left[f - \hat{f}\right]$, by property (2) of expectations

MSE decomposition 3

$$MSE = E\left[\epsilon^2\right] + \underbrace{E\left[\left(f - \hat{f}\right)^2\right]}_{\text{B}} + 0$$

Let's simplify term B:

$$E\left[\left(f - \hat{f}\right)^2\right] = \left(E\left[f - \hat{f}\right]\right)^2 + Var\left(f - \hat{f}\right)$$, by property (3) of expectations

$$= \left(f - E\left[\hat{f}\right]\right)^2 + Var(\hat{f})$$ since f is fixed and thus drops out of the variance

MSE decomposition 4

$$MSE = \underbrace{E\left[\epsilon^2\right]}_{\text{Noise}} + \underbrace{\left(f - E\left[\hat{f}\right]\right)^2}_{\text{Bias}^2} + \underbrace{Var(\hat{f}\,)}_{\text{Variance}}$$

- Noise: Irreducible (at least with the current predictors)
- Bias: How far off the mark the prediction algorithm is, on average
- Variance: How much prediction results vary from different training data samples