Section 2F. Bias-Variance Tradeoff

Statistics for Data Science

Victor M. Preciado, PhD MIT EECS
Dept of Electrical & Systems Engineering
University of Pennsylvania
preciado@seas.upenn.edu

Bias-Variance Tradeoff

We can explain the shape of the test MSE as follows:

- Consider a training dataset $\mathcal{D}_{\mathsf{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$, where the input/output data pairs $(\mathbf{x}_i, y_i)$ are drawn independently from an additive model:

$$\mathbf{x}_i \sim f_X \text{ and } y_i = f(\mathbf{x}_i) + \varepsilon \text{ with } \varepsilon \sim f_\varepsilon$$

  where ε is a measurement noise and f is an unknown regression function that "nature" is using to generate the dataset. Notice that the additive model induces a joint PDF, f_{XY}
- Assuming a parametric form of our regression function $\widehat{f}(\mathbf{x};\theta)$, our task is to find the parameters $\theta^\star$ that minimize the *training MSE*

$$\theta^\star = \arg\min_{\theta} \frac{1}{N} \sum_{(\mathbf{x}_i, y_i) \in \mathcal{D}_{\mathsf{Tr}}} \left(y_i - \widehat{f}(\mathbf{x}_i;\theta) \right)^2$$

Bias-Variance Tradeoff (cont.)

- Once we have a trained model $\widehat{f}(\mathbf{x};\theta^\star)$, we are interested in analyzing the *test MSE*. Considering a new datapoint $(\mathbf{x}_0, y_0) \sim f_{XY}$ (not used in the training process), we can *theoretically* write the test MSE as

$$\text{MSE}_{\text{Te}} = \mathbb{E}_{(\mathbf{x_0},y_\mathbf{0})\sim f_{XY}}\left[\left(y_0 - \widehat{f}(\mathbf{x}_0;\theta^\star)\right)^2\right]$$

- Notice that, since $\mathbf{x}_0$ is a r.v., the function $\widehat{f}(\mathbf{x}_0;\theta^\star)$ is also a r.v. The test MSE can be written as (proof omitted)

$$\text{MSE}_{\text{Te}} = \text{Var}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right] + \left(\text{Bias}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right]\right)^2 + \text{Var}[\varepsilon]$$

where

$$\text{Bias}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right] = \mathbb{E}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right] - f(\mathbf{x}_0)$$

$$\text{Var}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right] = \mathbb{E}\left[\left(\widehat{f}(\mathbf{x}_0;\theta^\star) - \mathbb{E}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right]\right)^2\right]$$

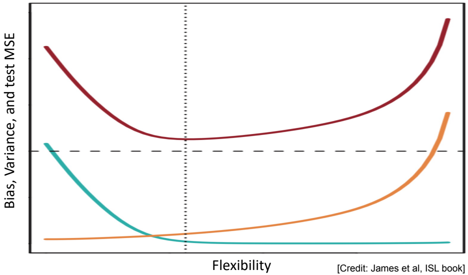
Bias-Variance Tradeoff (cont.)

- The sum of the terms $\mathrm{Var}\left[\widehat{f}\left(\mathbf{x}_0;\theta^\star\right)\right] + \left(\mathrm{Bias}\left[\widehat{f}\left(\mathbf{x}_0;\theta^\star\right)\right]\right)^2$ is called the *reducible error*, since it can be reduced by choosing a good parametric function $\widehat{f}(\mathbf{x};\theta)$
- The term $\mathrm{Var}\left[\varepsilon\right]$ is called *irreducible error*, since it is always there, even when your parametric function is exactly the same as the regression function f that "Nature" is using to generate the dataset you observe

Bias-Variance Tradeoff (cont.)

The two terms in the reducible error depend on the flexibility of the model $\widehat{f}(\mathbf{x};\theta)$

- The *bias term* $\text{Bias}\left[\widehat{f}(\mathbf{x}_0;\theta^\star)\right]$ decreases monotonically as the flexibility of the model increases (cyan plot)
- The variance term $\text{Var}\left[\widehat{f}(\mathbf{x};\theta)\right]$, increases monotonically as the flexibility of the model increases (orange plot)



[Credit: James et al, ISL book]

Bias-Variance Tradeoff (cont.)

Numerical interpretation: Run 10 different polynomial fits (with 10 different training datasets) when $d = 1, 3$ and 20.

- For $d = 1$ (rigid case), the variance is low, but the bias is high
- For $d = 3$ (cubic case), the variance and bias are both low
- For $d = 20$ (flexible case), the bias is still low, but the variance increases