Bias Variance Tradeoff

Fraida Fund

Contents

In this lecture

- Quantifying prediction error
- Bias-variance tradeoff

Prediction error

Model class

General ML estimation problem: given data (x_i, y_i), want to learn $y \approx \hat{y} = f(x)$

The **model class** is the **set** of possible estimates:

$$\hat{y} = f(\mathbf{x}, \boldsymbol{\beta})$$

parameterized by $\boldsymbol{\beta}$

Model class vs. true function

Our learning algorithm *assumes* a model class

$$\hat{y} = f(\mathbf{x}, \boldsymbol{\beta})$$

But the data has a *true* relation

$$y = f_0(\mathbf{x}) + \epsilon, \quad \epsilon \sim N(0, \sigma_\epsilon^2)$$

Sources of prediction error

- Noise: ϵ is fundamentally unpredictable, occurs because y is influenced by factors not in $\mathbf{x}$
- Assumed model class: maybe $f(\mathbf{x}, \boldsymbol{\beta}) \neq f_0(\mathbf{x})$ for *any* $\boldsymbol{\beta}$ (**under-modeling**)
- Parameter estimate: maybe $f(\mathbf{x}, \boldsymbol{\beta}) = f_0(\mathbf{x})$ for some true $\boldsymbol{\beta}_0$, but our estimate $\hat{\boldsymbol{\beta}} \neq \boldsymbol{\beta}_0$

Quantifying prediction error

Given

- parameter estimate $\hat{\boldsymbol{\beta}}$ (computed from a fixed training set)
- a *test point* $\mathbf{x}_{test}$ (was not in training set)

Then

- predicted value $\hat{y} = f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})$
- true value $y = f_0(\mathbf{x}_{test}) + \epsilon$

Output mean squared error (1)

Definition: output MSE given $\hat{\boldsymbol{\beta}}$:

$$MSE_y(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) := E[y - \hat{y}]^2$$

$$= E[f_0(\mathbf{x}_{test}) + \epsilon - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2$$

Output mean squared error (2)

Noise ϵ on test sample is independent of $f_0(\mathbf{x}_{test}), f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})$ so

$$= E[f_0(\mathbf{x}_{test}) + \epsilon - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2$$

$$= E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2 + E[\epsilon]^2$$

$$= E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2 + \sigma_\epsilon^2$$

Irreducible error (1)

Irreducible error σ_ϵ^2 is a fundamental limit on ability to predict y (lower bound on MSE).

$$MSE(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) \geq \sigma_\epsilon^2$$

Irreducible error (2)

Best case scenario: if

- true function is in model class: $f(\mathbf{x}, \boldsymbol{\beta}) = f_0(\mathbf{x})$ for a true $\boldsymbol{\beta_0}$, and
- our parameter estimate is perfect: $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta_0}$

then $E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2 = 0$ so output error = σ_ϵ^2.

Function MSE (1)

We had output MSE, error on predicted value:

$$MSE_y(\mathbf{x}_{test}) := E[y - \hat{y}]^2 = E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2 + \sigma_\epsilon^2$$

Now we will define function MSE, error on underlying function:

$$MSE_f(\mathbf{x}_{test}) := E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2$$

Function MSE (2)

Which can be decomposed into two parts:

$$MSE_f(\mathbf{x}_{test}) := E[f_0(\mathbf{x}_{test}) - f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]^2$$

$$\begin{aligned} MSE_f(\mathbf{x}_{test}) = & \\ & (f_0(\mathbf{x}_{test}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})])^2 + \\ & E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]]^2 \end{aligned} \tag{1}$$

Function MSE (3)

Note: cancellation of the cross term - Let $\bar{f}(\mathbf{x}_{test}) = E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]$. The cross term

$$E[(f_0(\mathbf{x}_{test}) - \bar{f}(\mathbf{x}_{test}))(f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) - \bar{f}(\mathbf{x}_{test}))]$$

$$= (f_0(\mathbf{x}_{test}) - \bar{f}(\mathbf{x}_{test}))E[(f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) - \bar{f}(\mathbf{x}_{test}))]$$

$$= (f_0(\mathbf{x}_{test}) - \bar{f}(\mathbf{x}_{test}))(\bar{f}(\mathbf{x}_{test}) - \bar{f}(\mathbf{x}_{test})) = 0$$

A hypothetical (impossible) experiment

Suppose we would get many independent training sets (from same process).

For each training set,

- train our model (estimate parameters), and
- use this model to estimate value of test point

Bias in function MSE

Bias: How much the average value of our estimate differs from the true value:

$$Bias(\mathbf{x}_{test}) := f_0(\mathbf{x}_{test}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]$$

Variance in function MSE

Variance: How much the estimate varies around its average:

$$Var(\mathbf{x}_{test}) := E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})]]^2$$

Bias and variance

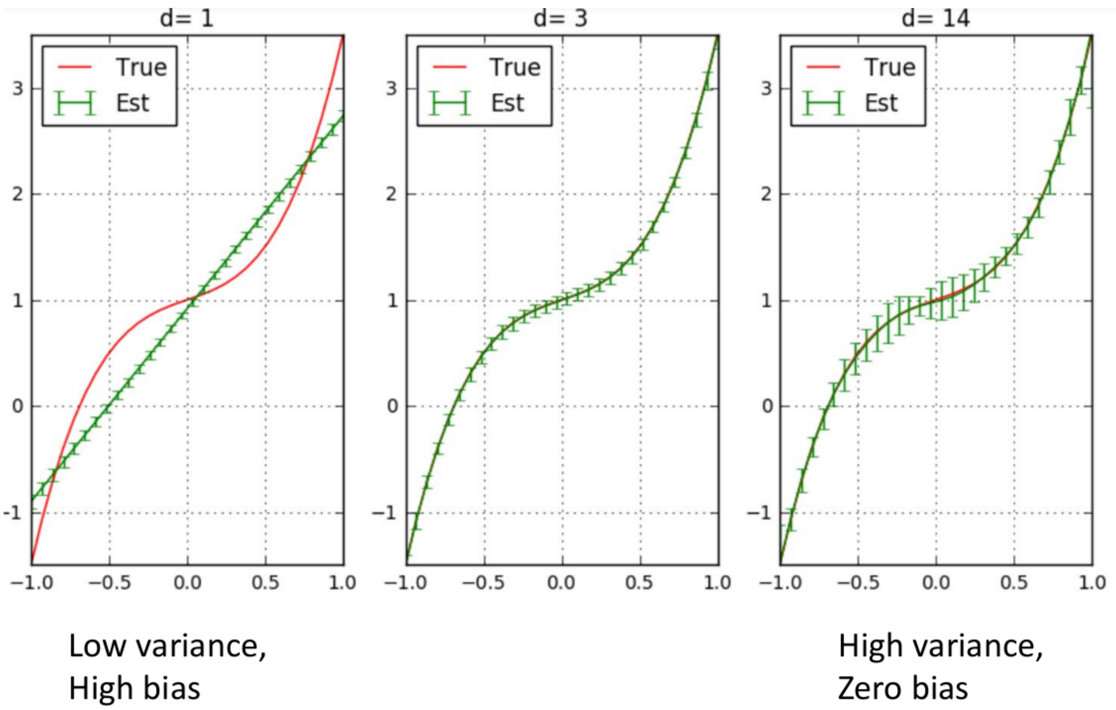

Figure 1: Example: 100 trials, mean estimate and standard deviation.

Summary: decomposition of MSE

Output MSE is the sum of squared bias, variance, and irreducible error:

$$
\begin{aligned}
MSE(\mathbf{x}_{test}) = & \\
& (f_0(\mathbf{x}_{test}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}})])^2 + \\
& E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}) - E[f(\mathbf{x}_{test}, \hat{\boldsymbol{\beta}}]]^2 + \\
& \sigma_\epsilon^2
\end{aligned}
\tag{2}
$$

What does it indicate?

Bias:

- Model "not flexible enough" - true function is not in model class (under-modeling or underfitting)

Variance:

- Model is very different each time we train it on a different training set
- Model "too flexible" - model class is too general and also learns noise (overfitting)

How to get small error?

- Get model selection right: not too flexible, but flexible enough (**how?**)
- Have enough data to constrain variability of model
- Other ways?

Bias variance tradeoff

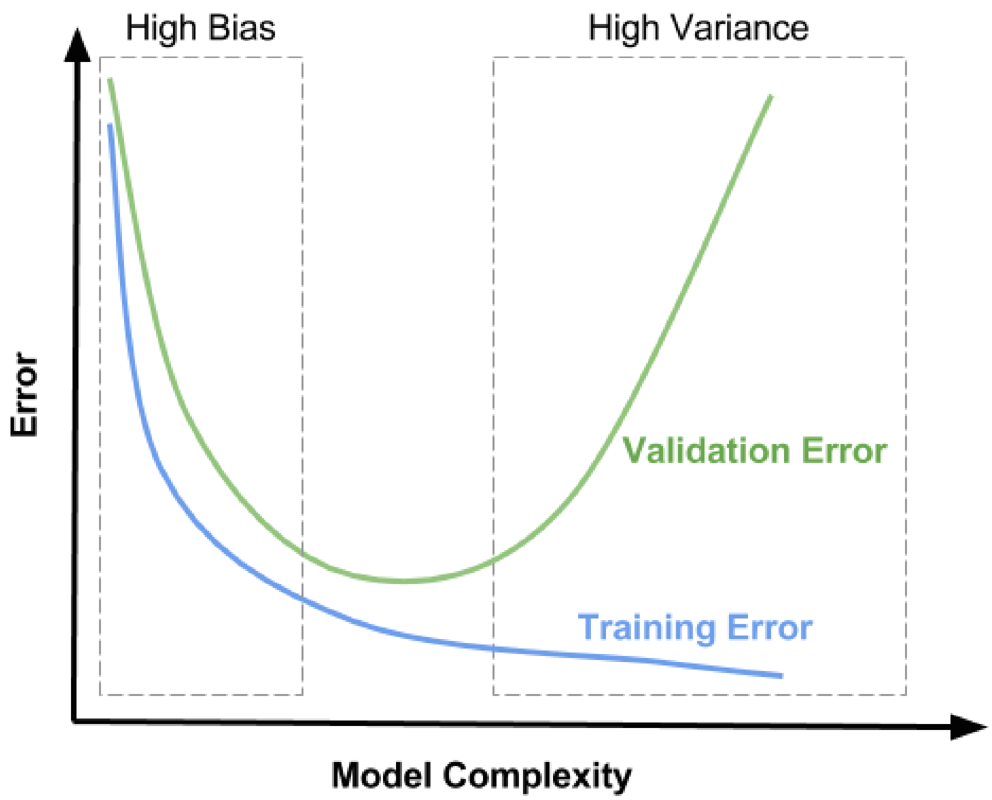

Figure 2: Bias variance tradeoff