Contents

I Part: Prediction **1**

1 Bias and Variance trade-off **1**
- 1.1 No free lunch . . . 1
- 1.2 Assessing model accuracy . . . 1
- 1.3 Bias-Variance Trade-Off . . . 4
- 1.4 Test error computation . . . 4

Part I

Part: Prediction

1 Bias and Variance trade-off

1.1 No free lunch

- No one statistical learning method dominates all others over all possible problems.
- Hence it is an important task to decide for any given problem which method produces the best results.

1.2 Assessing model accuracy

- In order to evaluate the performance of a statistical learning method on a given data set, we need some way to measure *how well its predictions actually match the observed data.*
- In the regression setting, the most commonly-used measure is the **mean squared error** (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2$$

- The above MSE is computed from the data used for estimating the model, the so called *training data*, and so should be referred as the **training** MSE.

- But, in general, *we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data.*
- If (x_0, y_0) is a previously unseen test observation not used to train the statistical learning method, we'd like to select the model for which the average of this quantity

$$\text{Average}(y_0 - \hat{f}(x_0))^2$$

  the **test** MSE—is the lowest.
- But the training MSE often is quite different from the test MSE, and in particular the former can dramatically underestimate the latter.

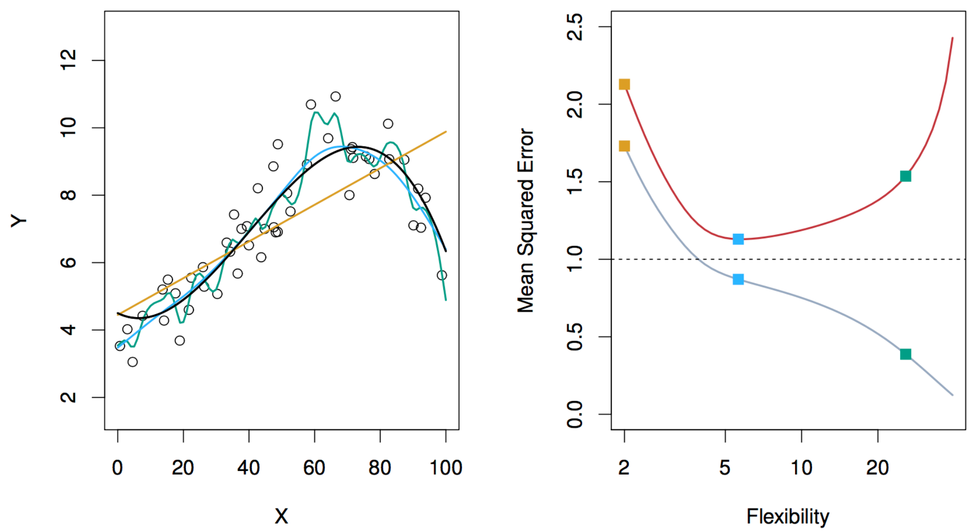

Figure 1: Training- versus Test-Set Performance: Linear regression and two smoothing splines at comparison.

1.2.1 Training Error versus Test error

- In the classification setting, we will see the distinction between the *test error* and the *training error*.
- The **test error** is the average error that results from using a statistical learning method to predict the response on a new observation, one that was not used in training the method.
- In contrast, the **training error** can be easily calculated by applying the statistical learning method to the observations used in its training.

Some trade-offs

Flexibility is measured by *degrees of freedom*.

Statistical learning techniques at comparison.

Overfitting

- When a given method yields a small training MSE but a large test MSE, we are said to be *overfitting* the data.
- When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply don't exist in the test data.
- Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE.
- Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE.

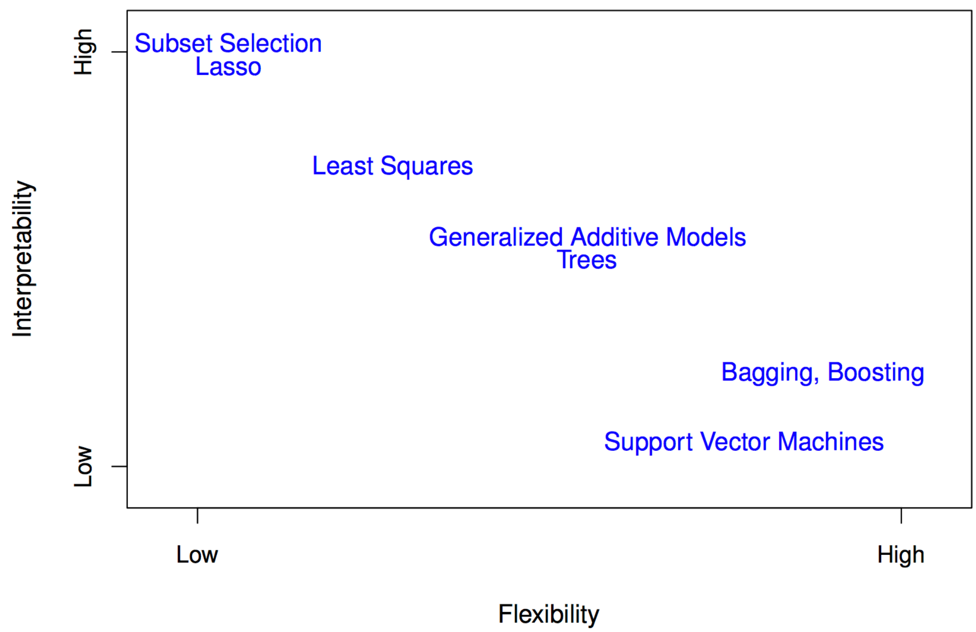

Figure 2: Interpretability vs Flexibility.

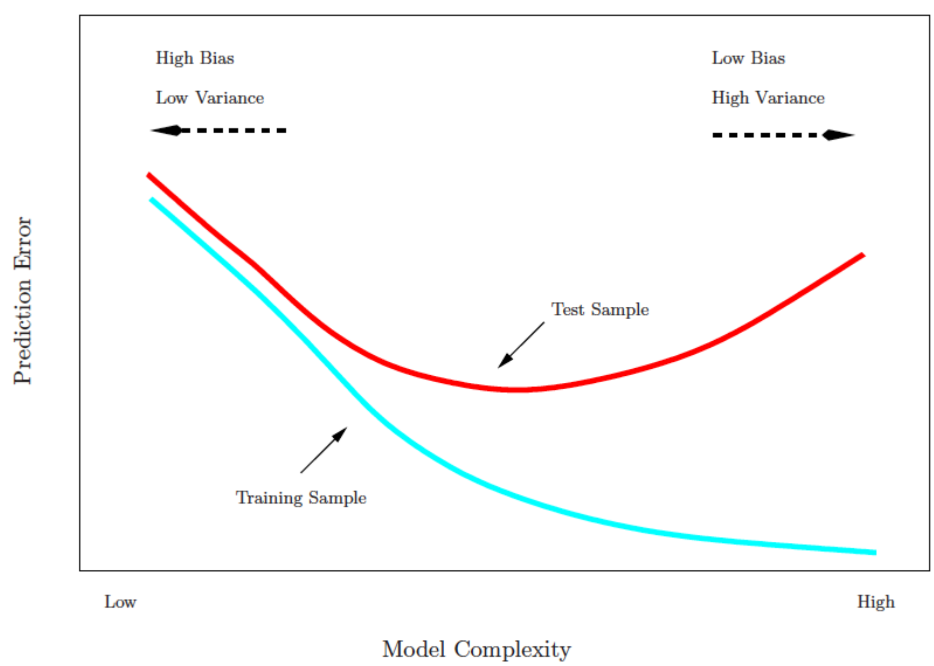

Figure 3: Training- versus Test-Set Performance.

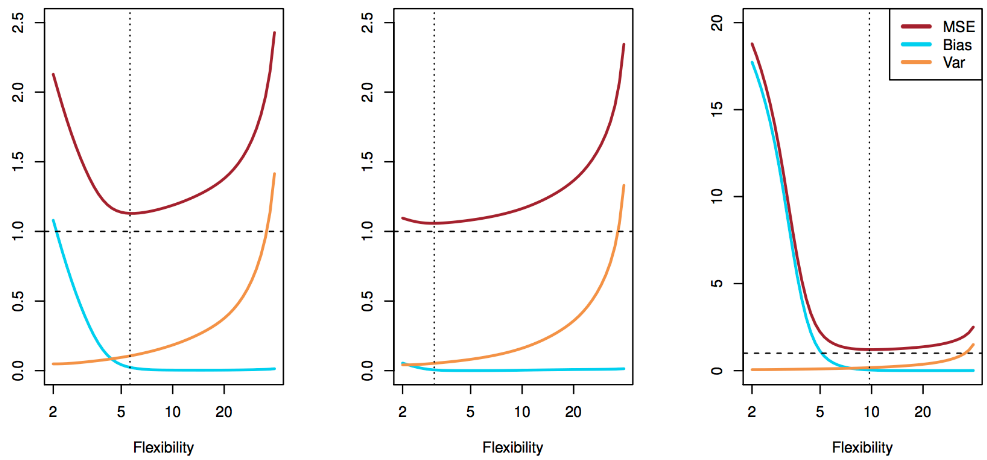

Figure 4: Bias-Variance Trade-Off in three examples.

1.3 Bias-Variance Trade-Off

- The **U-shape** observed in the test MSE curves turns out to be the result of two competing properties of statistical learning methods.
- it is possible to show that the *test MSE*, for a given value x_0, can always be *decomposed into the sum of three fundamental quantities*

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{V}ar(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{V}ar(\epsilon)$$

- **Variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set.
- In general, more flexible statistical methods have higher variance.
- **Bias** refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- Generally, more flexible methods result in less bias.
- Variance of the **irreducible** error: even if we knew $f(X)$, we would still make errors in prediction, since at each $X = x$ there is typically a distribution of possible Y values.

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \underbrace{E[f(x_0) - \hat{f}(x_0)]^2}_{\textbf{reducible}} + \underbrace{\mathrm{V}ar(\epsilon)}_{\textbf{irreducible}}$$

The relationship between bias, variance, and test set MSE is referred to as the **bias-variance trade-off**.

The challenge lies in finding a method for which both the variance and the squared bias are low.

1.4 Test error computation

- In real life problems in which f is unobserved, it is generally not possible to explicitly compute the test performance measure, bias, or variance for a statistical learning method.

- In the absence of a very large designated test set that can be used to *directly estimate the test performance* measure, a number of techniques can be used to *estimate this quantity using the available training data*.

- Best solution: a large designated test set. Often not available.
- Some methods make a *mathematical adjustment* to the training performance measure in order to estimate the test performance measure. These include the C_p statistic, **AIC** and **BIC**.
- A class of methods that estimate the test performance measure by *holding out* a subset of the training observations from the fitting process, and then applying the statistical learning method to those held out observations.