Chapter 2: Overview of Supervised Learning

Yuan Yao

Department of Mathematics
Hong Kong University of Science and Technology

Most of the materials here are from Chapter 2 of Introduction to Statistical learning by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani.

Other related materials are listed in Reference.

Spring, 2018

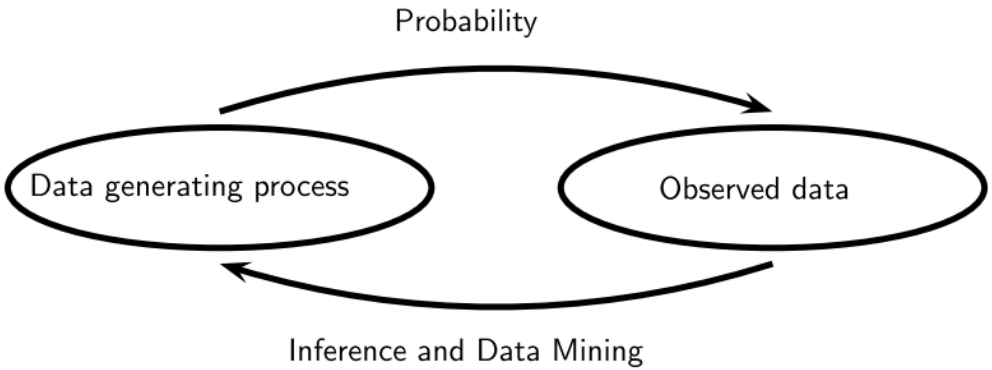
About the Course

- Course web: https://yuany-pku.github.io/2018_math4432/
- Basic and standard contents: supervised learning (regression and classification)
- (Slightly) more advanced topics: nonlinear models, tree methods, boosting, svm, neural networks...
- Emphasize model selection (such as regularization, validation) that are directly related with learning/prediction.
- Related courses:
  - Statistical Learning: Math 5470 by Bingyi JING
  - Unsupervised learning: CSIC5011 2017, Topological and Geometric Data Reduction (http://math.stanford.edu/~yuany/course/2017.fall/)
  - Deep learning: Math 6380o (https://deeplearning-math.github.io/)

Textbooks and Reference Books

- Textbook: An introduction to Statistical Learning (ISLR)
- Reference: Elements of Statistical Learning (ESL)
- Will stick with ISL and may cite ESL accassionally.
- Programming languages: R or Python
- Acknowledge the use of the graphics in the textbook/refernce for only the purpose of presentation.

Probability vs. Statistical Machine Learning



Forward problem: Probability is a language to quantify uncertainty.

Inverse Problem: Statistics or Machine Learning

Statistics/Data Mining Dictionary

Statisticians and computer scientists often use different language for the same thing. Here is a dictionary that the reader may want to return to throughout the course.

Statistics	Computer Science	Meaning
estimation	learning	using data to estimate an unknown quantity
classification	supervised learning	predicting a discrete Y from X
clustering	unsupervised learning	putting data into groups
data	training sample	$(X_1, Y_1), \ldots, (X_n, Y_n)$
covariates	features	the X_i's
classifier	hypothesis	a map from covariates to outcomes
hypothesis	—	subset of a parameter space Θ
confidence interval	—	interval that contains an unknown quantity with given frequency
directed acyclic graph	Bayes net	multivariate distribution with given conditional independence relations
Bayesian inference	Bayesian inference	statistical methods for using data to update beliefs
frequentist inference	—	statistical methods with guaranteed frequency behavior
large deviation bounds	PAC learning	uniform bounds on probability of errors

Figure 1: Larry Wasserman's classification of statistical learning vs. machine learning in Computer Science

Outline

Outline

Statistical learning

- Suppose that we observe a quantitative response Y and p different predictors, $X_1, X_2, \ldots, X_p$. We assume that there is some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$, which can be written in the very general form

$$Y = f(X) + \epsilon, \tag{1}$$

  where f is some fixed but unknown function of $X_1, \ldots, X_p$,and ϵ is a random *error* term, which is independent of X and has mean zero. In this formulation, f represents the *systematic* information that X provides about Y.
- In essence, statistical learning refers to a set of approaches for estimating f. In this chapter we outline some of the key theoretical concepts that arise in estimating f, as well as tools for evaluating the estimates obtained.
- There are two main reasons that we may wish to estimate f: *prediction* and *inference*.

Prediction

- In many situations, a set of inputs X are readily available, but the output Y cannot be easily obtained. In the context of time series analysis, X could correspond to $X_{t-1}, \ldots, X_{t-p}$, and Y corresponds to X_t.
- We can predict Y using

$$\hat{Y} = \hat{f}(X), \tag{2}$$

where $\hat{f}$ represents our estimate for f, and $\hat{Y}$ represents the resulting prediction for Y. The accuracy of $\hat{Y}$ as a prediction for Y depends on two quantities:

  - **reducible error**: $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce some error. This error is reducible because we can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique to estimate f.
  - **irreducible error**: Even if it were possible to form a perfect estimate for f, so that our estimated response took the form $\hat{Y} = f(X)$, our prediction would still have some error in it! This is because Y is also a function of ϵ, which, by definition, cannot be predicted using X. Therefore, variability associated with ϵ also affects the accuracy of our predictions. This is known as the irreducible error, because no matter how well we estimate f, we cannot reduce the error introduced by ϵ.

- Why is the irreducible error larger than zero? The quantity ϵ may contain unmeasured variables that are useful in predicting Y: since we don't measure them, f cannot use them for its prediction. The quantity ϵ may also contain unmeasurable variation (any thoughts here?).
- Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, it is easy to show that

$$\begin{aligned} \mathbb{E}(Y - \hat{Y})^2 &= \mathbb{E}[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\textbf{Reducible}} + \underbrace{Var(\epsilon)}_{\textbf{Irreducible}}, \end{aligned} \tag{3}$$

  where $\mathbb{E}(Y - \hat{Y})^2$ represents the expected value of the squared difference between the predicted and actual value of Y, and $Var(\epsilon)$ represents the variance associated with the error term ϵ.
- The focus of this course is on techniques for estimating f with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

Inference

We are often interested in understanding the way that Y is affected as $X_1, \ldots, X_p$ change. In this situation we wish to estimate f, but our goal is not necessarily to make predictions for Y. We instead want to understand the relationship between X and Y, or more specifically, to understand how Y changes as a function of $X_1, \ldots, X_p$.

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation, or is the relationship more complicated?

How to estimate f?

Depending on whether our ultimate goal is prediction, inference, or a combination of the two, different methods for estimating f may be appropriate.

- Assume that we have observed a set of n different data points. These observations are called the *training data* because we will use these observations to train, or teach, our method how to estimate f.
- Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f. In other words, we want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y).
- Most statistical learning methods for this task can be characterized as either *parametric* or *non-parametric*.

Parametric Methods

Parametric methods involve a two-step model-based approach.

- First, we make an assumption about the functional form, or shape, of f. For example, one very simple assumption is that f is linear in X:

$$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{4}$$

  Once we have assumed that f is linear, the problem of estimating f is greatly simplified. Instead of having to estimate an entirely arbitrary p-dimensional function $f(X)$, one only needs to estimate the $p + 1$ coefficients $\beta_0, \beta_1, \ldots, \beta_p$.
- After a model has been selected, we need a procedure that uses the training data to fit or train the model. That is, we want to find values of these parameters such that

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p. \tag{5}$$

  The most common approach to fitting the model (4) is referred to as *(ordinary) least squares*.

The model-based approach just described is referred to as *parametric*; it reduces the problem of estimating f down to one of estimating a set of

- Assuming a parametric form for f simplifies the problem of estimating f because it is generally much easier to estimate a set of parameters, such as $\beta_0, \beta_1, \ldots, \beta_p$ in the linear model (4), than it is to fit an entirely arbitrary function f.
- The potential disadvantage of a parametric approach is that the model we choose will usually not match the true unknown form of f. If the chosen model is too far from the true f,then our estimate will be poor.
- We can try to address this problem by choosing *flexible* models that can fit many different possible functional forms flexible for f. But in general, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to a phenomenon known as **overfitting** the data, which essentially means they follow the errors, or noise, too closely.

Non-parametric Methods

- Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly.
- Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f.
- Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well.
- In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made.
- But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

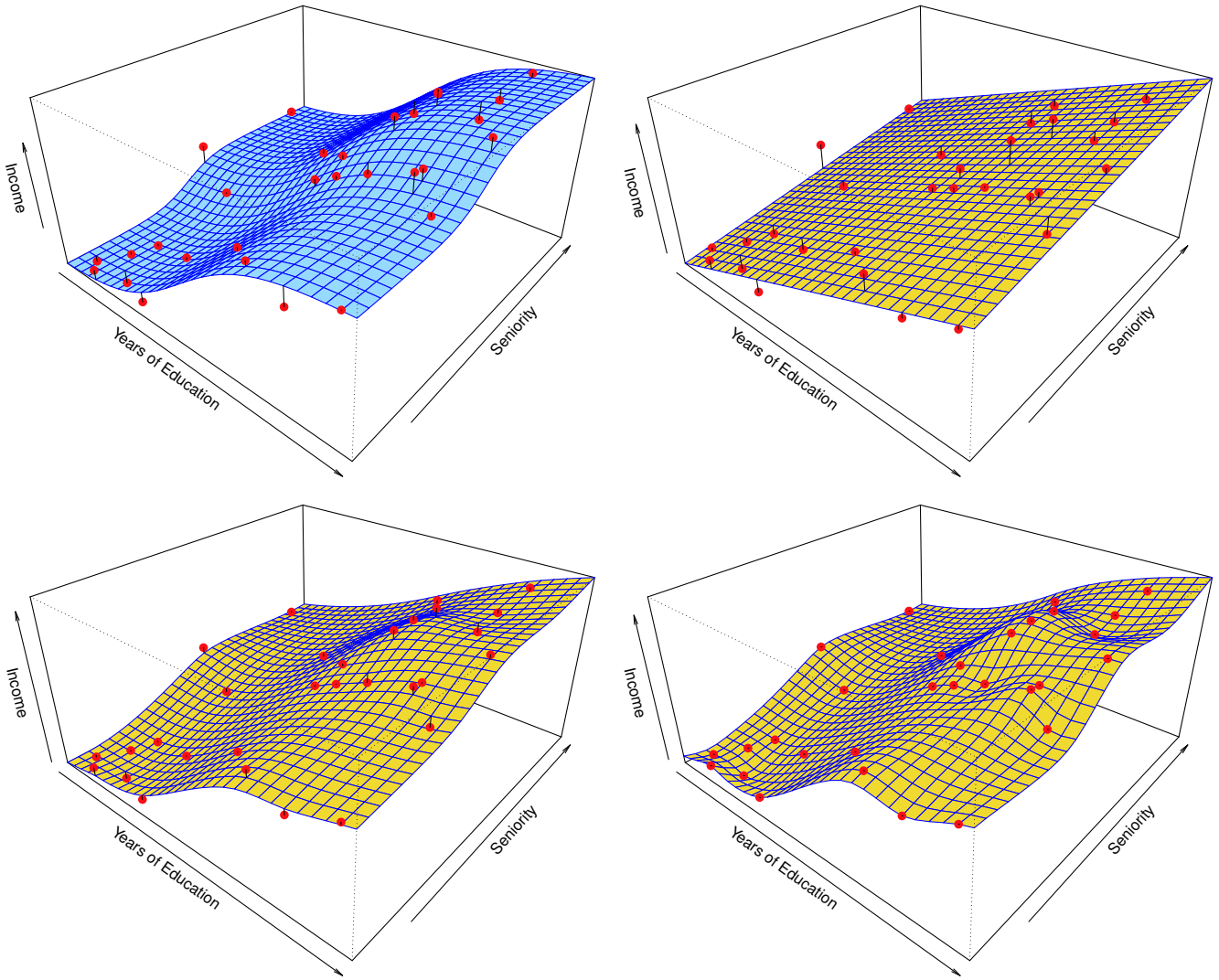


Figure 2: An illustrative example. Upper Left: A simulated *Income* data (red dots) with its true generative model (blue surface). Upper Right: A fitted Linear model (parametric). Lower Left: A fitted spline model (non-parametric)

Outline

"No free lunch in statistics"

- Why is it necessary to introduce so many different statistical learning approaches, rather than just a single best method? **There is no free lunch in statistics**: no one method dominates all others over all possible data sets. On a particular data set, one specific method may work best, but some other method may work better on a similar but different data set.
- Hence it is an important task to decide for any given set of data which method produces the best results. Selecting the best approach can be one of the most challenging parts of performing statistical learning in practice.
- In this section, we discuss some of the most important concepts that arise in selecting a statistical learning procedure for a specific data set.

Measuring the Quality of Fit

- In the regression setting, the most commonly-used measure is the *mean squared error* (MSE), given by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2, \tag{6}$$

  where $f(x_i)$ is the prediction that $\hat{f}$ gives for the i-th observation.
- The MSE in (6) is computed using the *training* data that was used to fit the model, and so should more accurately be referred to as the *training MSE*.
- But in general, we do not really care how well the method works training MSE on the training data. Rather, **we are interested in the accuracy of the predictions that we obtain when we apply our method to previously unseen test data**.
- Suppose that we are interested test data in developing an algorithm to predict a stock's price based on previous stock returns. We can train the method using stock returns from the past 6 months. But we don't really care how well our method predicts last week's stock price. We instead care about how well it will predict tomorrow's price or next month's price.

- To state it more mathematically, suppose that we fit our statistical learning method on our training observations $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ and we obtain the estimate $\hat{f}$.
- We can then compute $\hat{f}(x_1), \hat{f}(x_2), \ldots, \hat{f}(x_n)$. If these are approximately equal to $y_1, y_2, ..., y_n$, then the training MSE given by (6) is small.
- However, we are really not interested in whether $\hat{f}(x_i) \approx y_i$; instead, we want to know whether $\hat{f}(x_0)$ is approximately equal to y_0, where (x_0, y_0) is a **previously unseen test observation not used to train the statistical learning method**.
- We want to choose the method that gives the lowest **test MSE**, as opposed to the lowest training MSE. In other words, if we had a large number of test observations, we could compute

$$Ave(\hat{f}(x_0) - y_0)^2, \tag{7}$$

the average squared prediction error for these test observations (x_0, y_0). We'd like to select the model for which the average of this quantity-the test MSE-is as small as possible.

- How can we go about trying to select a method that minimizes the test MSE? In some settings, we may have a test data set available-that is, we may have access to a set of observations that were not used to train the statistical learning method. We can then simply evaluate (7) on the test observations, and select the learning method for which the test MSE is smallest.
- But what if no test observations are available? In that case, one might imagine simply selecting a statistical learning method that minimizes the training MSE (6). This seems like it might be a sensible approach, since the training MSE and the test MSE appear to be closely related.
- Unfortunately, there is a fundamental problem with this strategy: there is no guarantee that the method with the lowest training MSE will also have the lowest test MSE. Roughly speaking, the problem is that many statistical methods specifically estimate coefficients so as to minimize the training set MSE. For these methods, the training set MSE can be quite small, but the test MSE is often much larger.

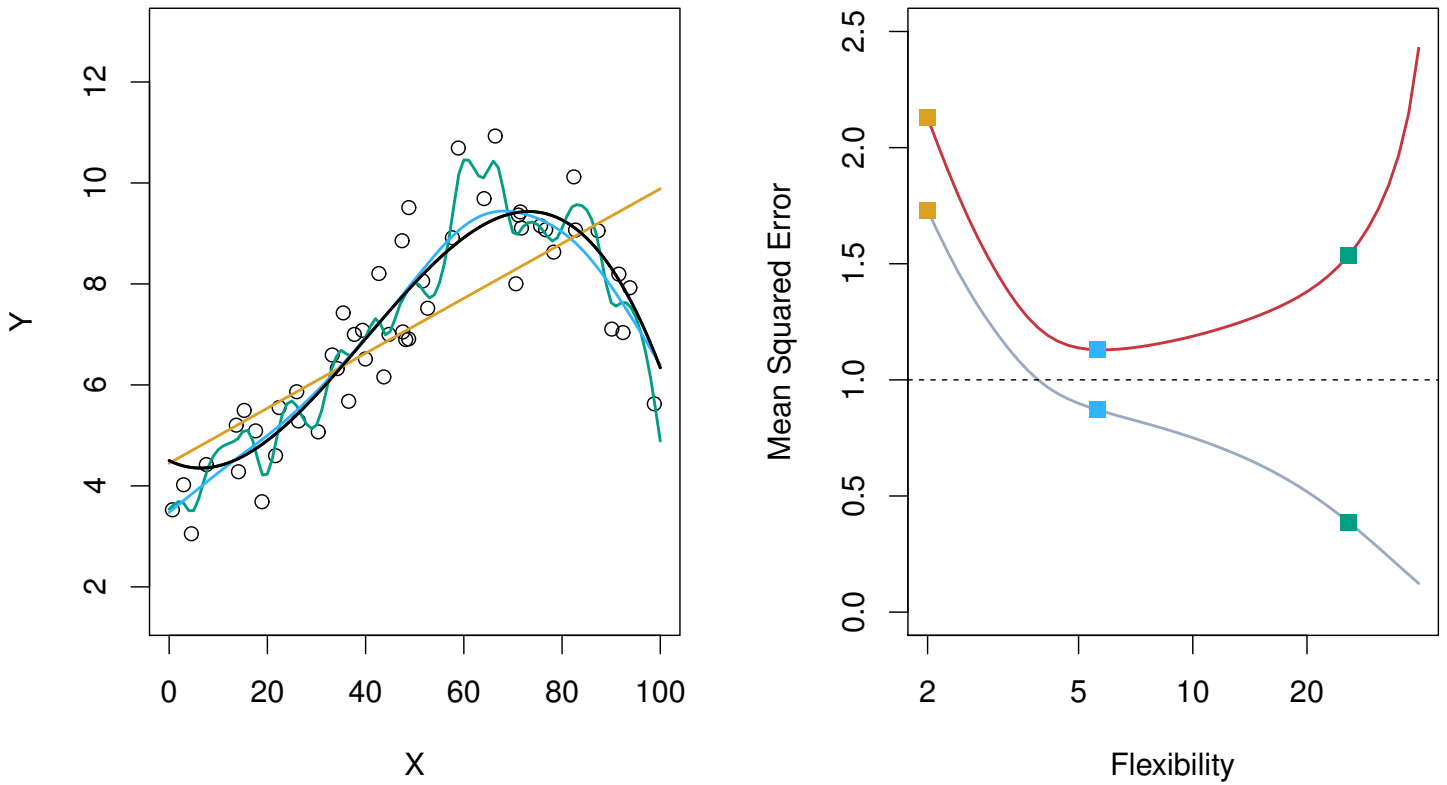


Figure 3: Illustration. Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the

Comments on Left panel of Figure 3

- The orange, blue and green curves illustrate three possible estimates for f obtained using methods with increasing levels of flexibility. The orange line is the linear regression fit, which is relatively inflexible. The blue and green curves were produced using smoothing splines with different levels of smoothness.

$$\sum_{i=1}^{n}(y_i - g(x_i))^2 + \lambda \int g''(t)^2 dt \tag{8}$$

  where λ is a nonnegative tuning parameter. The function g that minimizes (8) is known as a smoothing spline.
- As λ tends to ∞, the function g tends to linear because $\int g''(t)^2 dt$ has to tend to 0.
- It is clear that as the level of flexibility increases, the curves fit the observed data more closely. The green curve is the most flexible and matches the data very well; however, we observe that it fits the true f (shown in black) poorly because it is too wiggly. By adjusting the level of flexibility of the smoothing spline fit, we can produce many different fits to this data.

Comments on Right panel of Figure 3: Training MSE

- The grey curve displays the average training MSE as a function of flexibility, or more formally, the **degrees of freedom** which is a quantity that summarizes the flexibility of a model. The orange, blue and green squares indicate the MSEs associated with the corresponding curves in the left-hand panel.
- A more restricted and hence smoother curve has fewer degrees of freedom than a wiggly curve, linear regression is at the most restrictive end, with two degrees of freedom. The training MSE declines monotonically as flexibility increases. In this example the true f is non-linear, and so the orange linear fit is not flexible enough to estimate f well. The green curve has the lowest training MSE of all three methods, since it corresponds to the most flexible of the three curves fit in the left-hand panel.

Comments on Right panel of Figure 3: Test MSE

- In this example, we know the true function f, and so we can also compute the test MSE over a very large test set, as a function of flexibility. (Of course, in general f is unknown, so this will not be possible.)
- As with the training MSE, the test MSE initially declines as the level of flexibility increases. However, at some point the test MSE levels off and then starts to increase again. Consequently, the orange and green curves both have higher test MSE. The blue curve minimizes the test MSE, which should not be surprising given that visually it appears to estimate f the best.
- The horizontal dashed line indicates $Var(\epsilon)$, the irreducible error in (3), which corresponds to the lowest achievable test MSE among all possible methods.

Overfitting

- In the right-hand panel of Figure 3, as the flexibility of the statistical learning method increases, we observe a monotone decrease in the training MSE and a U-shape in the test MSE.
- This is a fundamental property of statistical learning that holds regardless of the particular data set at hand and regardless of the statistical method being used. **As model flexibility increases, training MSE will decrease, but the test MSE may not**.
- When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function f.
- 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.

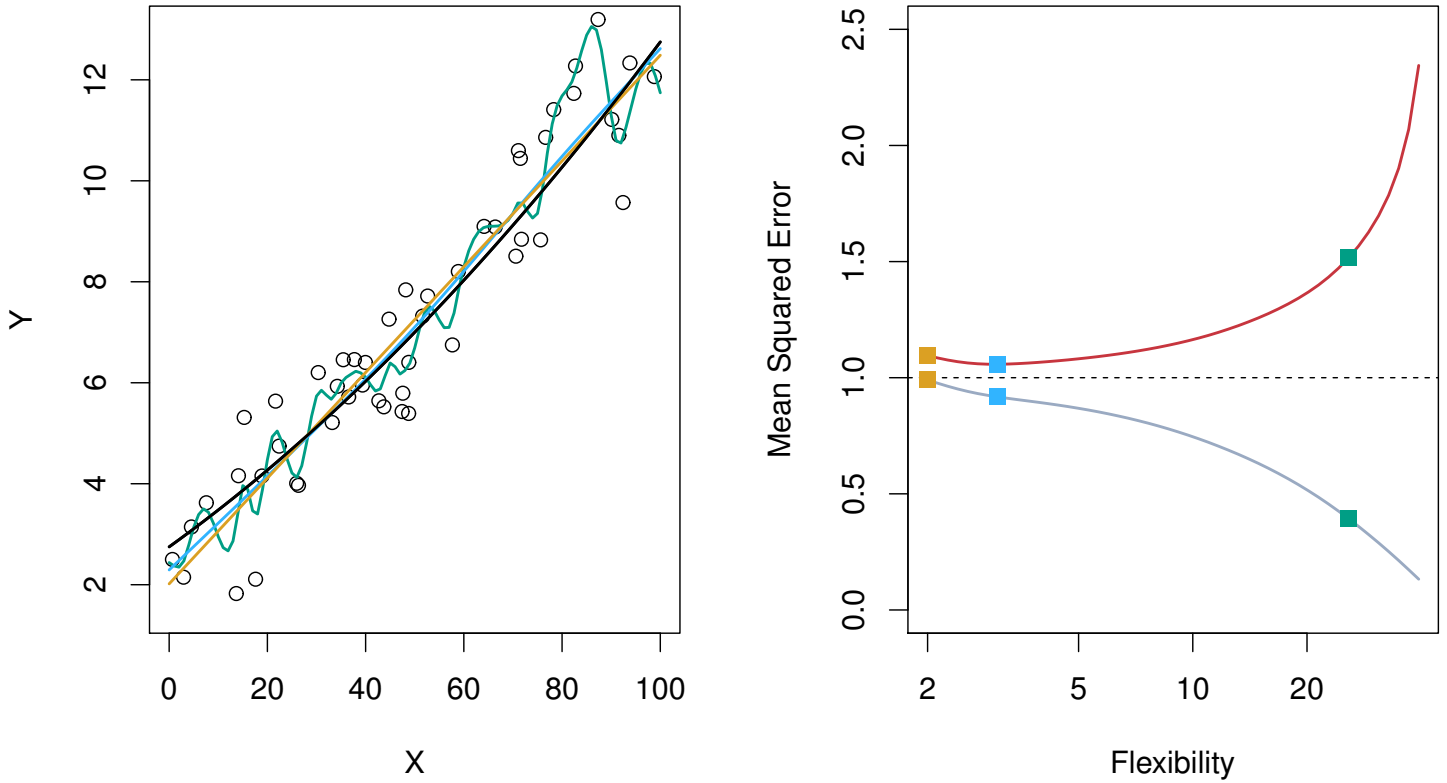


Figure 4: More illustration. Details are as in Figure 3, using a different **true** f **that is much closer to linear**. In this setting, linear regression provides a very good fit to the data.

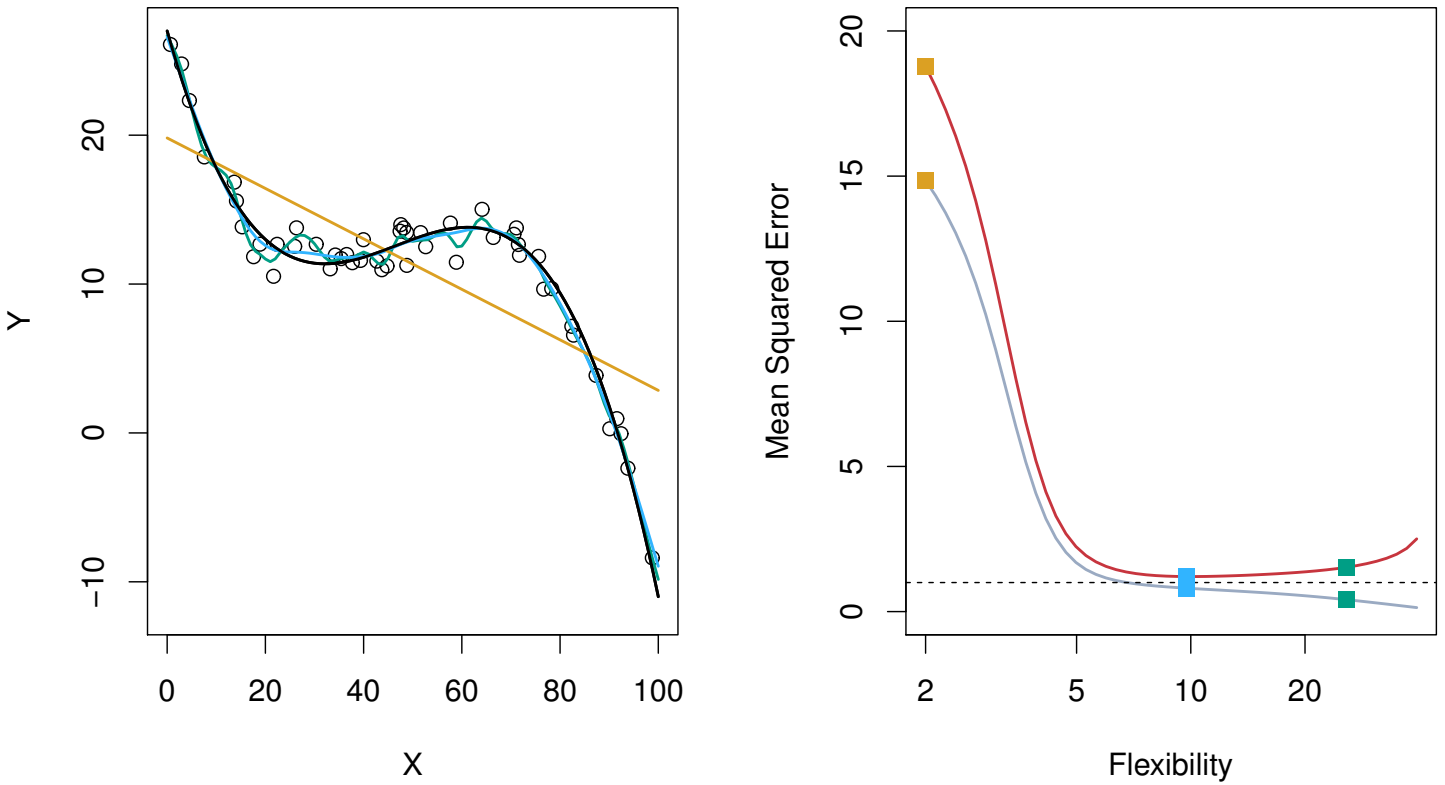


Figure 5: More illustration. Details are as in Figure 3, using a different **true** f **that is far from linear**. In this setting, linear regression provides a very poor fit to the data (**underfitting**).

- In practice, one can usually compute the training MSE with relative ease, but estimating test MSE is considerably more difficult because usually no test data are available.
- As the previous three examples illustrate, the flexibility level corresponding to the model with the minimal test MSE can vary considerably among data sets.
- In Chapter 3, we discuss some approaches that can be used in practice to estimate this minimum point, such as **Cross-validation** which is method for estimating test MSE using the training data.

Outline

The Bias-Variance Trade-Off

- Let $f(X)$ be the true function which we aim at estimating from a training data set $\mathcal{D}$.
- Let $\hat{f}(X;\mathcal{D})$ be the estimated function from the training data set $\mathcal{D}$.
- Are we really interested in

$$\min_{\hat{f}} \left[f(X) - \hat{f}(X;\mathcal{D}) \right]^2 ? \tag{9}$$

- **Fisher's view**: the measurements are a **random selection** from the set of all possible measurements which form the true distribution!
- What we really care is

$$\min_{\hat{f}} \mathbb{E}_{\mathcal{D}} \left[f(X) - \hat{f}(X;\mathcal{D}) \right]^2, \tag{10}$$

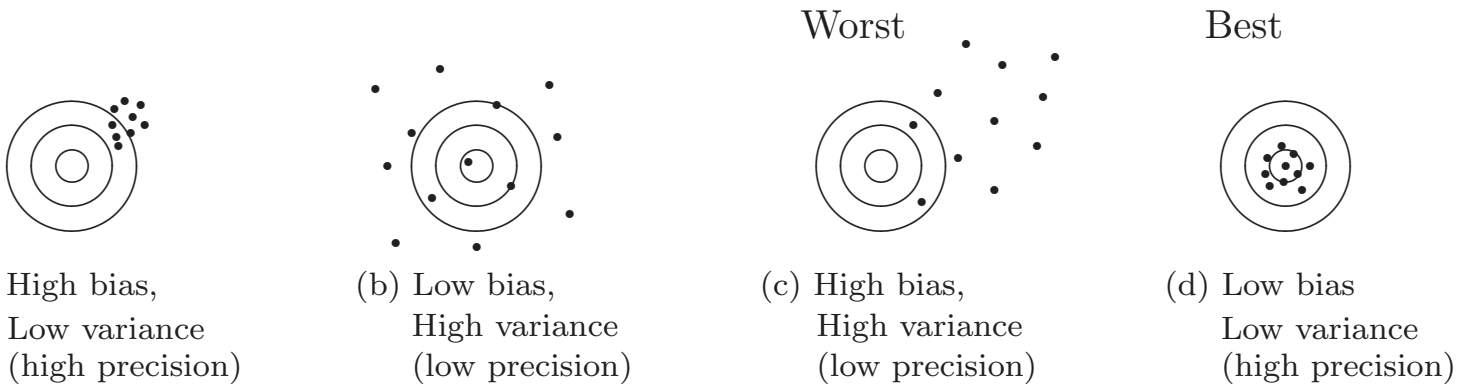
where randomness caused by **random selection** has been taken into account.

- If we add and subtract $\mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D}))$ inside the braces and then expand, we obtain

$$\begin{aligned}
&\left[f(X) - \hat{f}(X;\mathcal{D})\right]^2 \\
&= \left[f(X) - \mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D})) + \mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D})) - \hat{f}(X;\mathcal{D})\right]^2 \\
&= \left[f(X) - \mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D}))\right]^2 + \left[\mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D})) - \hat{f}(X;\mathcal{D})\right]^2 \\
&\quad + 2\left[f(X) - \mathbb{E}_{\mathcal{D}}[\hat{f}(X;\mathcal{D})]\right]\left[\mathbb{E}_{\mathcal{D}}[\hat{f}(X;\mathcal{D})] - \hat{f}(X;\mathcal{D})\right].
\end{aligned}$$

- Now we take the expectation of this expression with respect to $\mathcal{D}$ and note that the final term will vanish, giving

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}\left[f(X) - \hat{f}(X;\mathcal{D})\right]^2 \\
&= \underbrace{\left[f(X) - \mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D}))\right]^2}_{Bias^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left[\left[\mathbb{E}_{\mathcal{D}}(\hat{f}(X;\mathcal{D})) - \hat{f}(X;\mathcal{D})\right]^2\right]}_{Variance}
\end{aligned}$$



(a) High bias, Low variance (high precision)

(b) Low bias, High variance (low precision)

(c) High bias, High variance (low precision)

(d) Low bias Low variance (high precision)

- **Bias** refers to the error that is introduced by approximating a real-life problem, which may be extremely complicated, by a much simpler model.
- **Variance** refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set. Since the training data are used to fit the statistical learning method, different training data sets will result in a different $\hat{f}$. But ideally the estimate for f should not vary too much between training sets.
- **Bias and variance trade-off**: The optimal predictive acpability is the one that leads to balance between bias and variance.

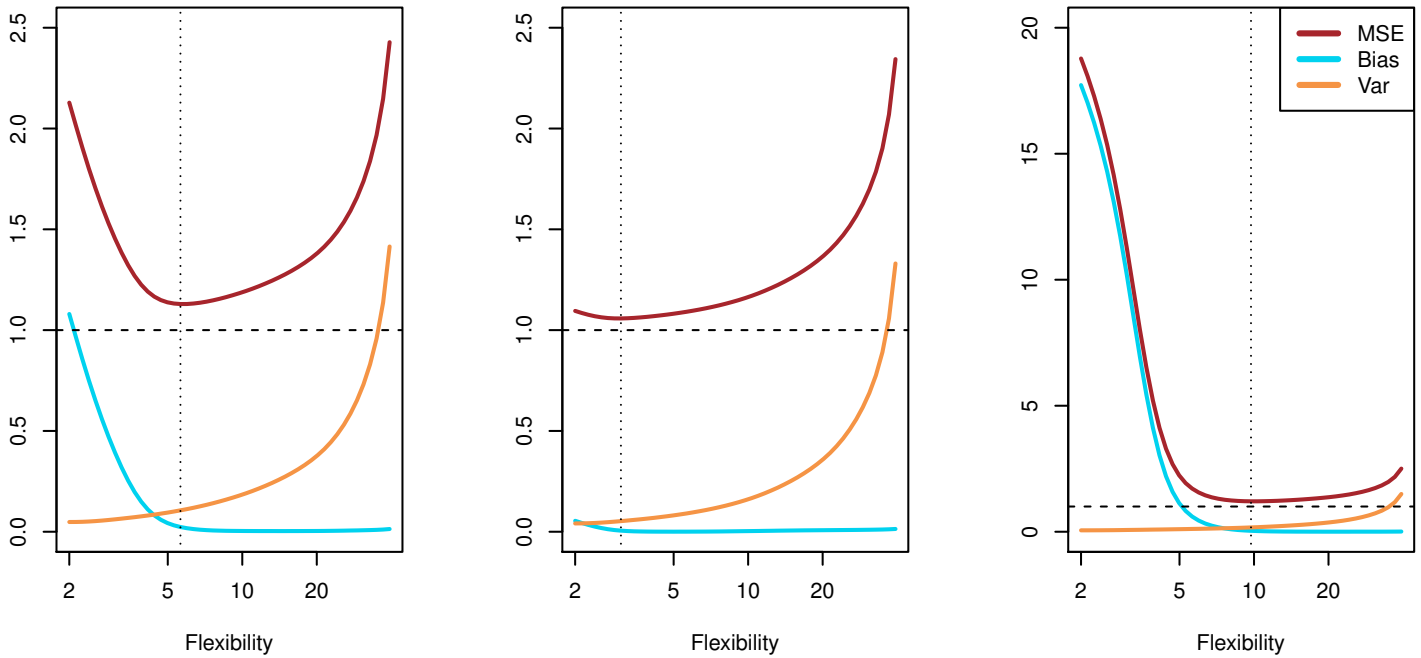


Figure 6: Squared bias (blue curve), variance (orange curve), Var(ϵ) (dashed line), and test MSE (red curve) for the three data sets in Figures 3-5. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

Bias-variance tradeoff

Elements of Statistical Learning (2nd Ed.) ©Hastie, Tibshirani & Friedman 2009 Chap 2

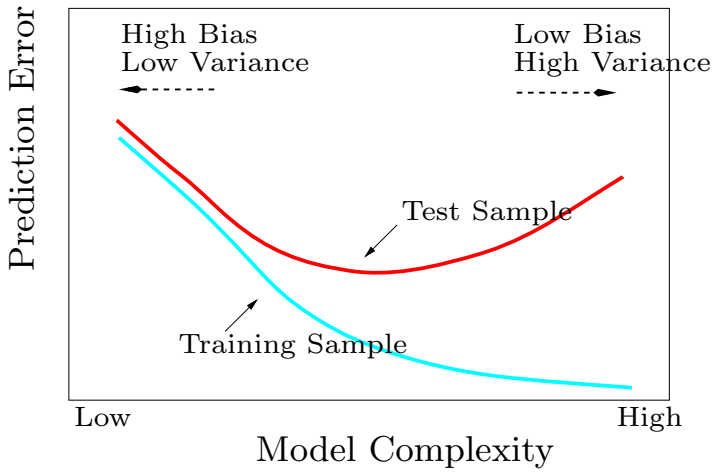


FIGURE 2.11. *Test and training error as a function of model complexity.*

Figure 7: George Box: "Essentially, all models are wrong, but some are useful."

References


Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani
An Introduction to Statistical Learning with Applications in R,
Springer, 2013.


Hastie, T., Tibshirani R., Friedman J.
The elements of statistical learning, 2nd
Springer, 2009.


Larry Wasserman.
All of Statistics.
http://www.stat.cmu.edu/~larry/all-of-statistics/