STATISTICAL LEARNING

CHAPTER 2: STATISTICAL LEARNING

INSTRUCTOR: SEOKHO LEE

HANKUK UNIVERSITY OF FOREIGN STUDIES

2015 SPRING

What Is Statistical Learning?

- We want to model some relationship between Y and $X = (X_1, X_2, \ldots, X_p)$ as

$$Y = f(X) + \epsilon \tag{2.1}$$

  - Y : **output**, **response**, or **dependent variable**
  - X : **input**, **predictors**, **independent variables**, **features**, or just **variables**
  - ϵ : random **error term**, independent of X and has mean zero
  - f : **systematic** information that X provides about Y

- In essence, statistical learning refers to a set of approaches for estimating f
- Consider a simple regression framework with Advertising data set
  - output : sales
  - input : 3 media (TV, radio, and newspaper)

What Is Statistical Learning?

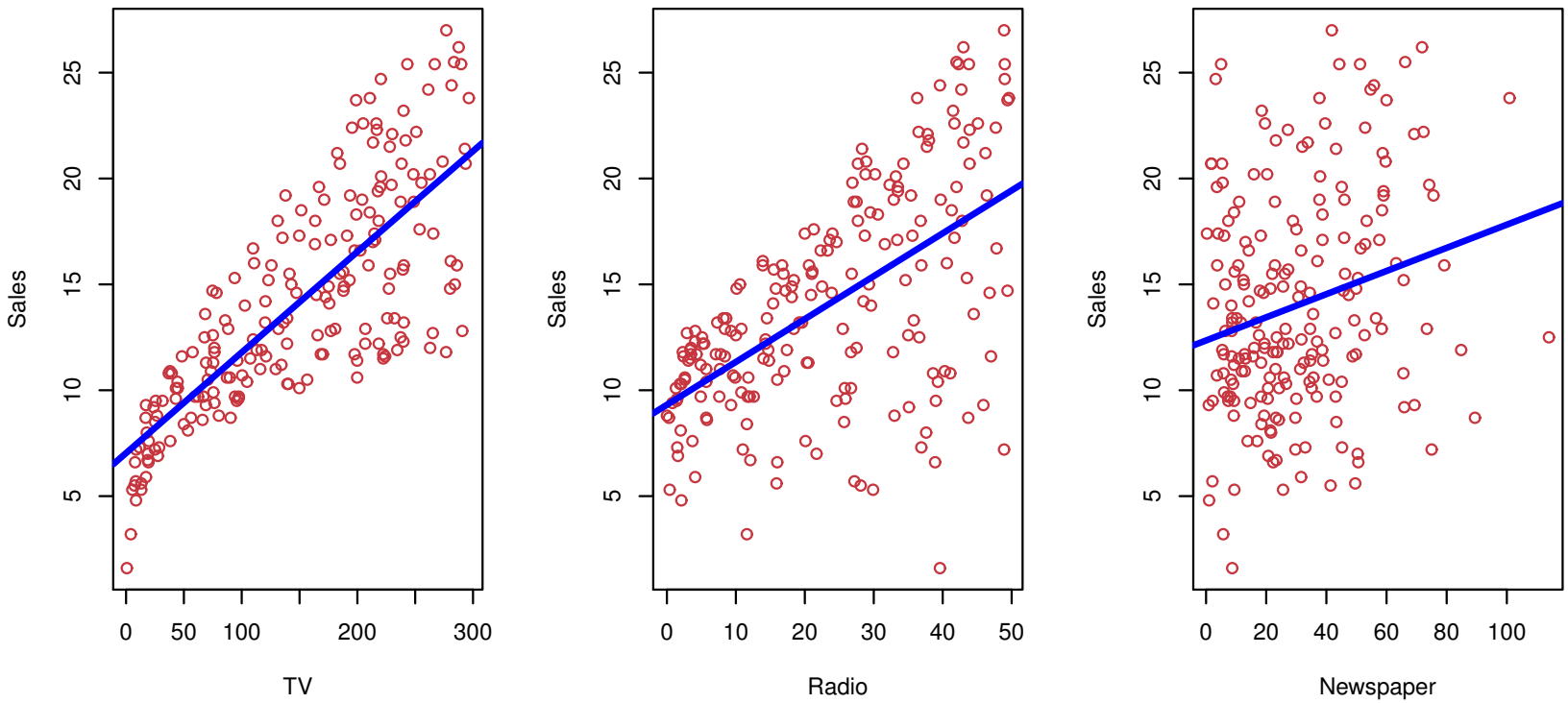

Figure 2.1: The Advertising data set. The plot displays sales, in thousands of units, as a function of TV, radio, and newspaper budgets, in thousands of dollars, for 300 different markets. In each plot we show the simple least squares fit of sales to that variable, as described in Chapter 3. In other words, each blue line represents a simple model that can be used to predict sales using TV, radio, and newspaper, respectively.

What Is Statistical Learning?

- Another example : Income data set (available in ISL webpage)
  - output : income
  - input : years of education

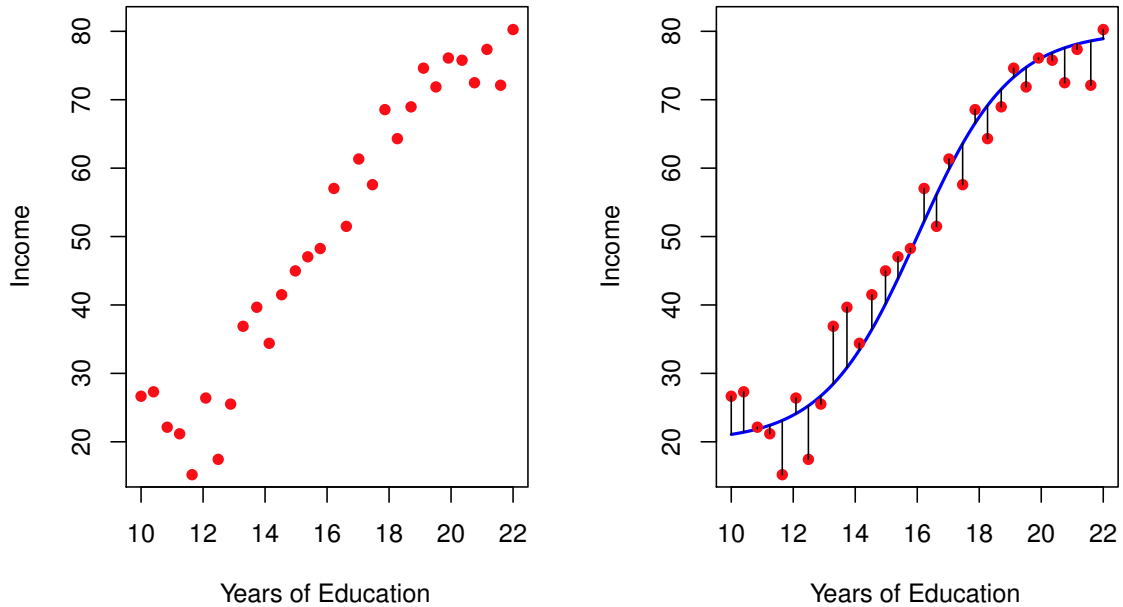

Figure 2.2: The Income data set. Left: The red dots are the observed values of income (in tens of thousands of dollars) and years of education for 30 individuals. Right: The blue curve represents the true underlying relationship between income and years of education, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Not that some errors are positive (if an observation lies above the blue curve) and some are negative (if an observation lies below the curve). Overall, these errors have approximately mean zero.

What Is Statistical Learning?

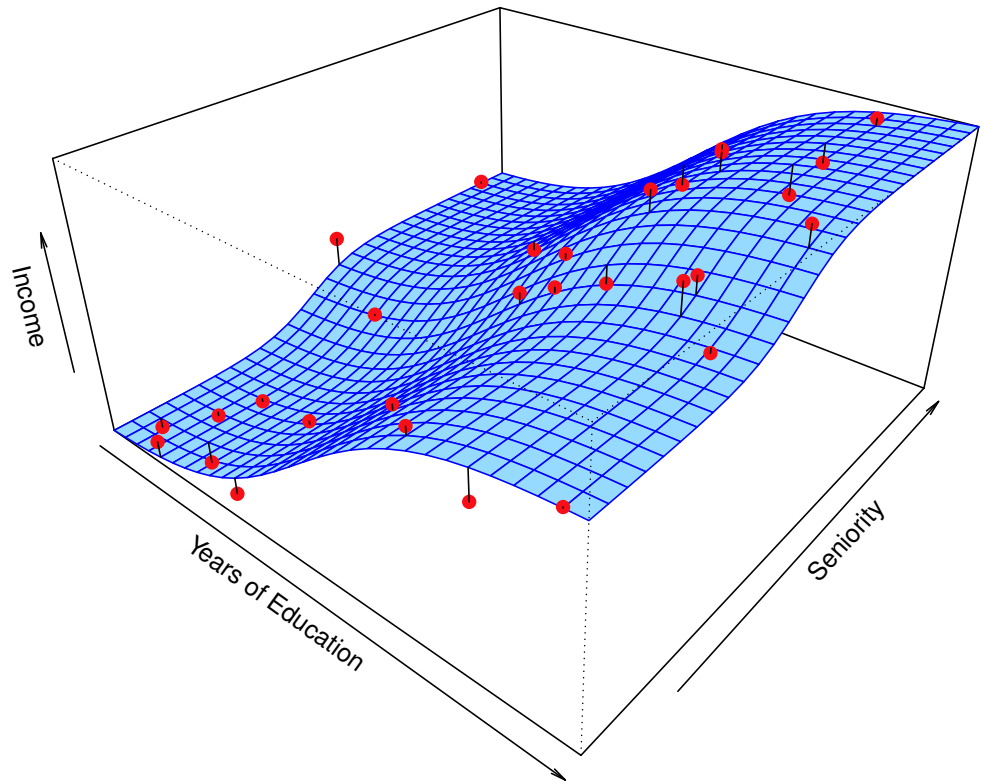

Figure 2.3: The plot displays income as a function of years of education and seniority in the Income data set. The blue surface represents the true underlying relationship between income and years of education and seniority, which is known since the data are simulated. The red dots indicate the observed values of these quantities for 30 individuals.

Why Estimate f?

- Two main reasons to estimate f: **prediction** and **inference**
- Prediction :

$$\hat{Y} = \hat{f}(X) \tag{2.2}$$

  - $\hat{Y}$: prediction of Y given X

- Accuracy of $\hat{Y}$ depends on two quantities:
  - **Reducible error** : $\hat{f}$ will not be a perfect estimate of f, and this inaccuracy will introduce some error. We can potentially improve the accuracy of $\hat{f}$ by using the most appropriate statistical learning technique
  - **Irreducible error** : No matter how well we estimate f, we cannot reduce the error introduced by ϵ

Prediction

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\
&= \underbrace{[f(X) - \hat{f}(X)]^2}_{\textit{Reducible}} + \underbrace{Var(\epsilon)}_{\textit{Irreducible}}
\end{aligned} \tag{2.3}
$$

- Aim : learn techniques for estimating f with the aim of minimizing the reducible error
  - 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

- Understand the way that Y is affected as $X_1, \ldots, X_p$ change
- Possible questions to be answered:
  - **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 Do We Estimate f?

- Training data : $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - Suppose we have n observations and p variables, and
  - x_{ij} represent the value of the jth predictor (input) for observation i $(i = 1, 2, \cdots, n; j = 1, 2, \cdots, p)$
  - y_i represent the response variable for the ith observation
- We train, or teach, our method how to estimate f based on the training data
- We want to find a function $\hat{f}$ such that $Y \approx \hat{f}(X)$ for any observation (X, Y)
- Two approaches for $\hat{f}$: **parametric** and **non-parametric**

Parametric Methods

- Two steps for model-based approach
  1. Make an assumption about the parametric functional form, or shape, of f
     - Linear model :
     $$f(X) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \tag{2.4}$$
  2. **Fit** or **train** the model using the training data
     - In linear model, estimate $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ such that
     $$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- "Estimating f" = "Estimating parameters"
- Disadvantage: The parametric model may not match the true unknown form of f
  - Selecting **flexible** models that can fit many different possible functional forms for f
  - Flexible parametric models involves many parameters
  - Complex model leads to **overfitting** to the (training) data. ("overfitting" means the models follow the errors, or **noise**, too closely)

Parametric Methods

- The Income data

$$\text{income} \approx \beta_0 + \beta_1 \times \text{education} + \beta_2 \times \text{seniority}$$

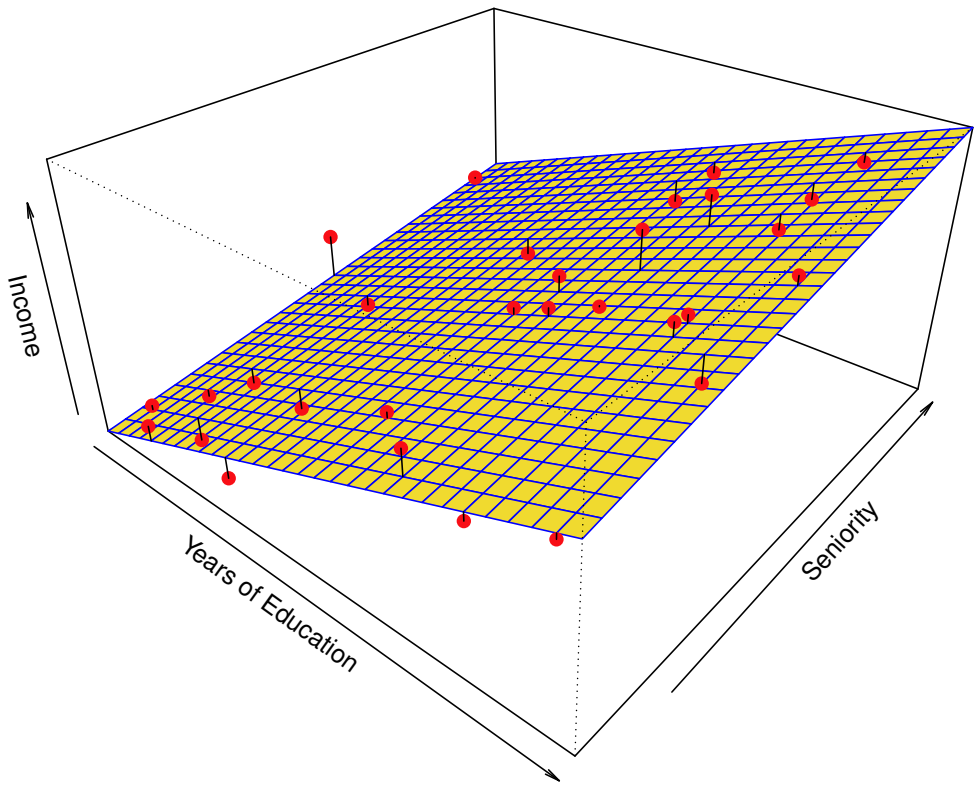

Figure 2.4: A linear model fit by least squares to the Income data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.

Non-parametric Methods

- Non-parametric methods do not make explicit assumption 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
  - **Advantage** : they have the potential to accurately fit a wider range of possible shapes for f
  - **Disadvantage** : a very large number of observation (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f
- Figure 2.5 and Figure 2.6 for Income data
  - **Thin-plate spline** is used to estimate f
  - Figure 2.5 is less smooth to Figure 2.6, showing a remarkably accurate estimate of the true f shown in Figure 2.3
  - The resulting estimate in Figure 2.6 fits the observed data perfectly, but far more variable than the true function

Non-parametric Methods

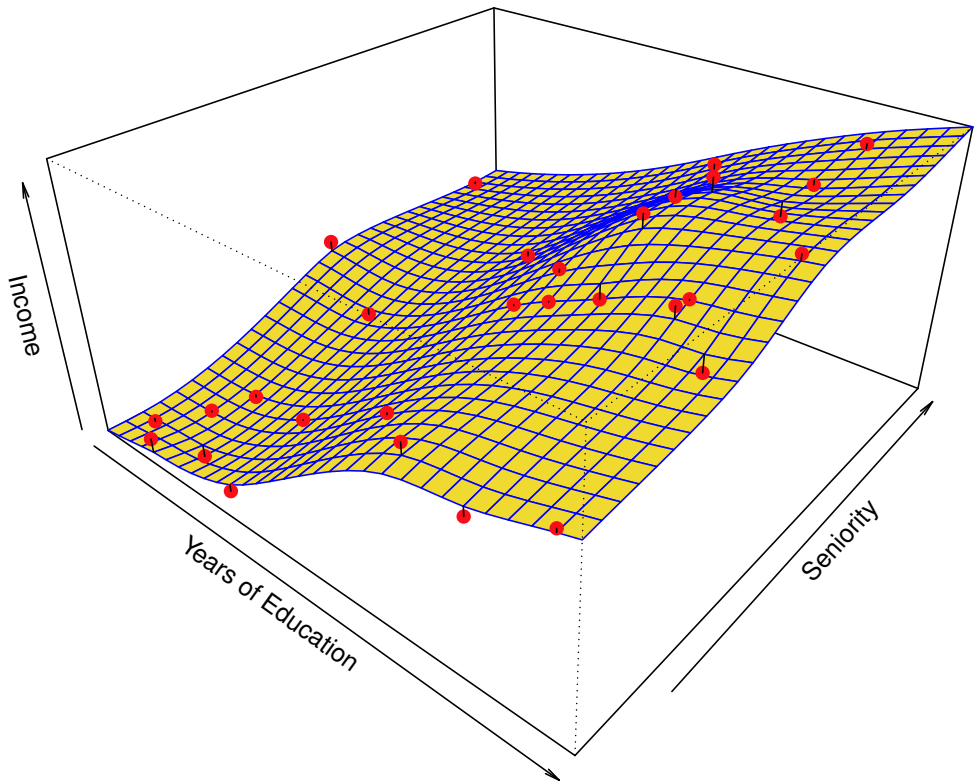

Figure 2.5: A smooth thin-plate spline fit to the Income data from Figure 2.3 is shown in yellow; the observations are displayed in red. Splines are discussed in Chapter 7.

Non-parametric Methods

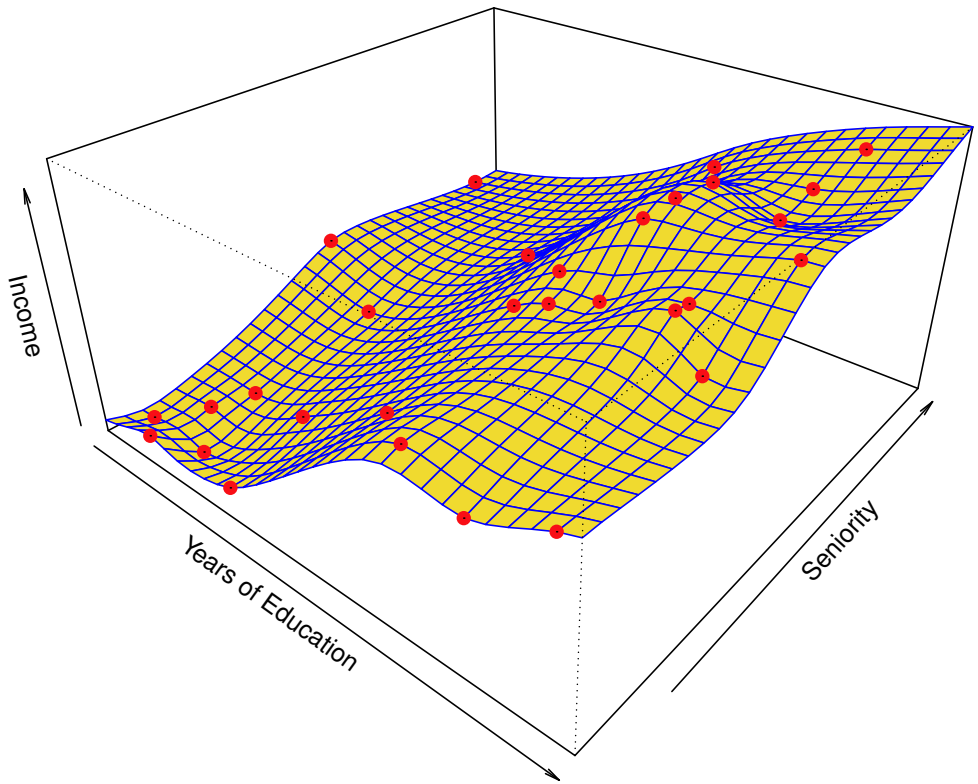

Figure 2.6: A rough thin-plate spline fit to the Income data from Figure 2.3. This fit makes zero errors on the training data.

The Trade-Off Between Prediction Accuracy and Model Interpretability

- Why would we ever choose to use a more restrictive method instead of a very flexible approach?
  - Restrictive models are much more interpretable
  - If interpretation (inference) is not of interest and we focus on prediction, then we might expect that it will be best to use the most flexible model available
  - Surprisingly, we will often obtain more accurate predictions using a less flexible method

The Trade-Off Between Prediction Accuracy and Model Interpretability

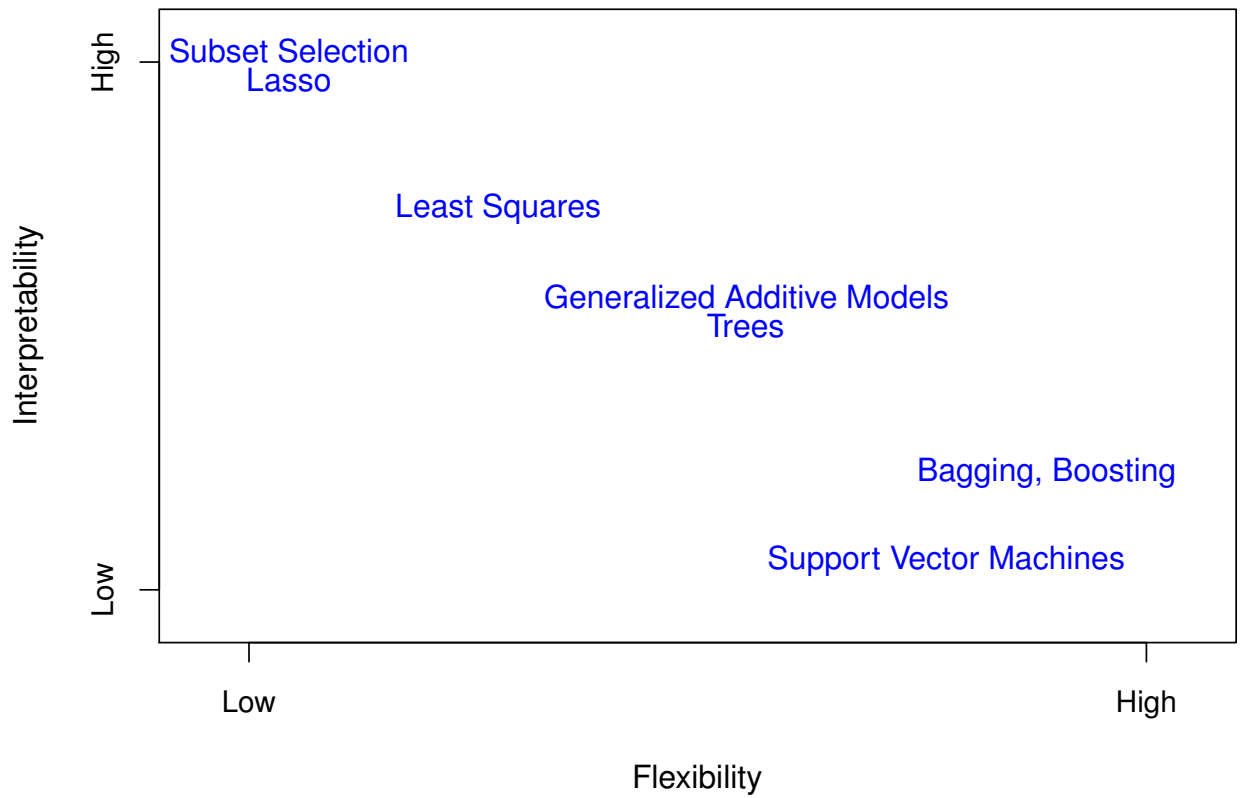

Figure 2.7: A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.

Supervised Versus Unsupervised Learning

- Most statistical learning problems fall into one of two categories: **supervised** or **unsupervised**
- Supervised learning
  - predictor $\mathbf{x}_i$ and response y_i $(i = 1, 2, \ldots, n)$
  - **linear regression**, **logistic regression**, **GAM**, **boosting**, **support vector machines**
  - goals:
    - Prediction - accurately predicting the response for future observations
    - Inference - better understanding the relationship between the response and the predictor
- Unsupervised learning
  - predictor $\mathbf{x}_i$ without response y_i (blind working)
  - **cluster analysis**
    - ascertaining, on the basis of $x_1, \ldots, x_n$, whether the observations fall into relatively distinct groups
    - Market segmentation study

Supervised Versus Unsupervised Learning

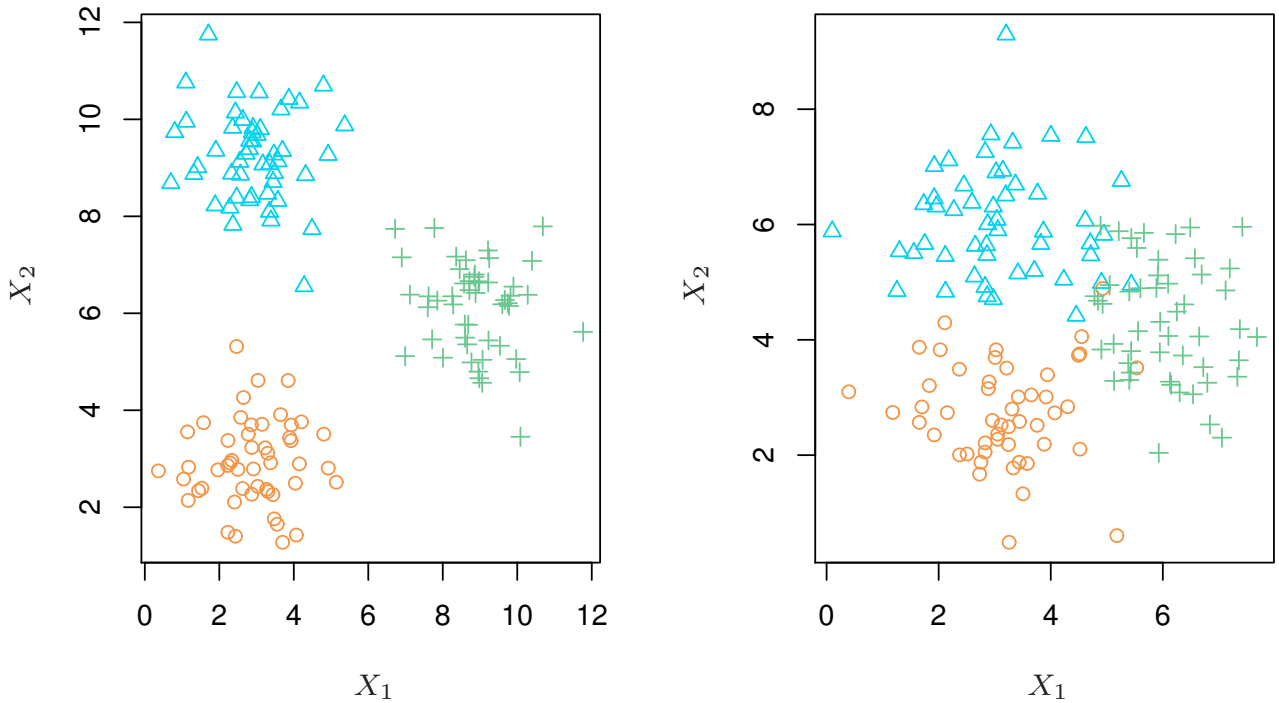

Figure 2.8: A clustering data set involving three groups. Each group is shown using a different colored symbol. Left: The three groups are well-separated. In this setting, a clustering approach should successfully identify the three groups. Right: There is some overlap among the groups. Now the clustering task is more challenging.

Regression Versus Classification Problems

- **Quantitative** variable
  - takes numeric values
  - a person's age, height, or income, the value of house, the price of stock, etc.
- **Qualitative** (also known as **categorical**) variable
  - takes on values in one of K different **classes**
  - a person's gender (male or female), the brand of product purchased (brand A, B, or C), whether a person defaults on a debt (yes or no), or a cancer diagnosis (Acute Myelogenous Leukemia, Acute Lymphoblastic Leukemia, or No Leukemia)
- We tend to refer to problems with a quantitative response as **regression** problems, while those involving a qualitative response are often referred to as **classification** problems.

Assessing Model Accuracy

- There is no best model for all data sets
- It is an important task to decide for any given set of data which method produces the best results
- Selecting the best approach is the most challenging parts of statistical learning in practice

Measuring the Quality of Fit

- We need some way to measure how well its predictions actually match the observed data
- **Mean squared error (MSE)**

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.5}$$

  - The smaller MSE, the better the fit
  - **Training MSE**
    - computed using the training data that was used to fit the model
    - not interested in training MSE (we already know the outcome in the training data!)
  - **Test MSE**
    - Average squared prediction error for test observations (x_0, y_0):

$$\text{Ave}(\hat{f}(x_0) - y_0)^2 \tag{2.6}$$

    - We want to choose the method that gives the lowest test MSE

Measuring the Quality of Fit

- If test data is available, test MSE can be evaluated straightforwardly
- If test data is not available,
  - you may tempt to use training MSE as a surrogate for test MSE $\Rightarrow$ WRONG!!!
  - training MSE can be very different from test MSE
  - smoothing spline example : see Figure 2.9 and Figure 2.10
- We will discuss a variety of approaches that can be used in practice to estimate the minimum of test MSE, including **cross-validation**
  - Cross-validation is a method for estimating test MSE using the training data

Measuring the Quality of Fit

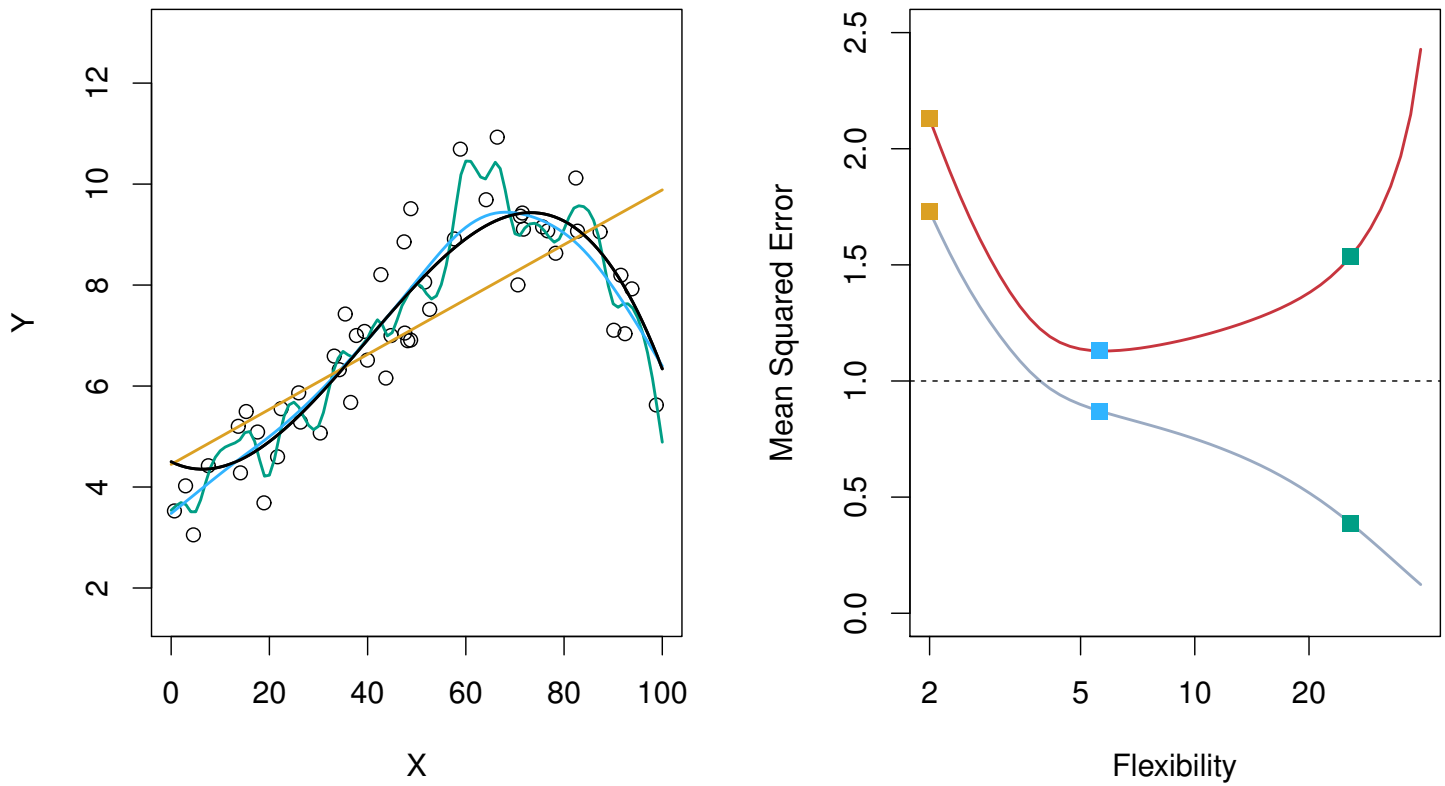

Figure 2.9: Left: Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves). Right: Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed line). Squares represent the training and test MSEs for the tree fits shown in the left-hand panel.

Measuring the Quality of Fit

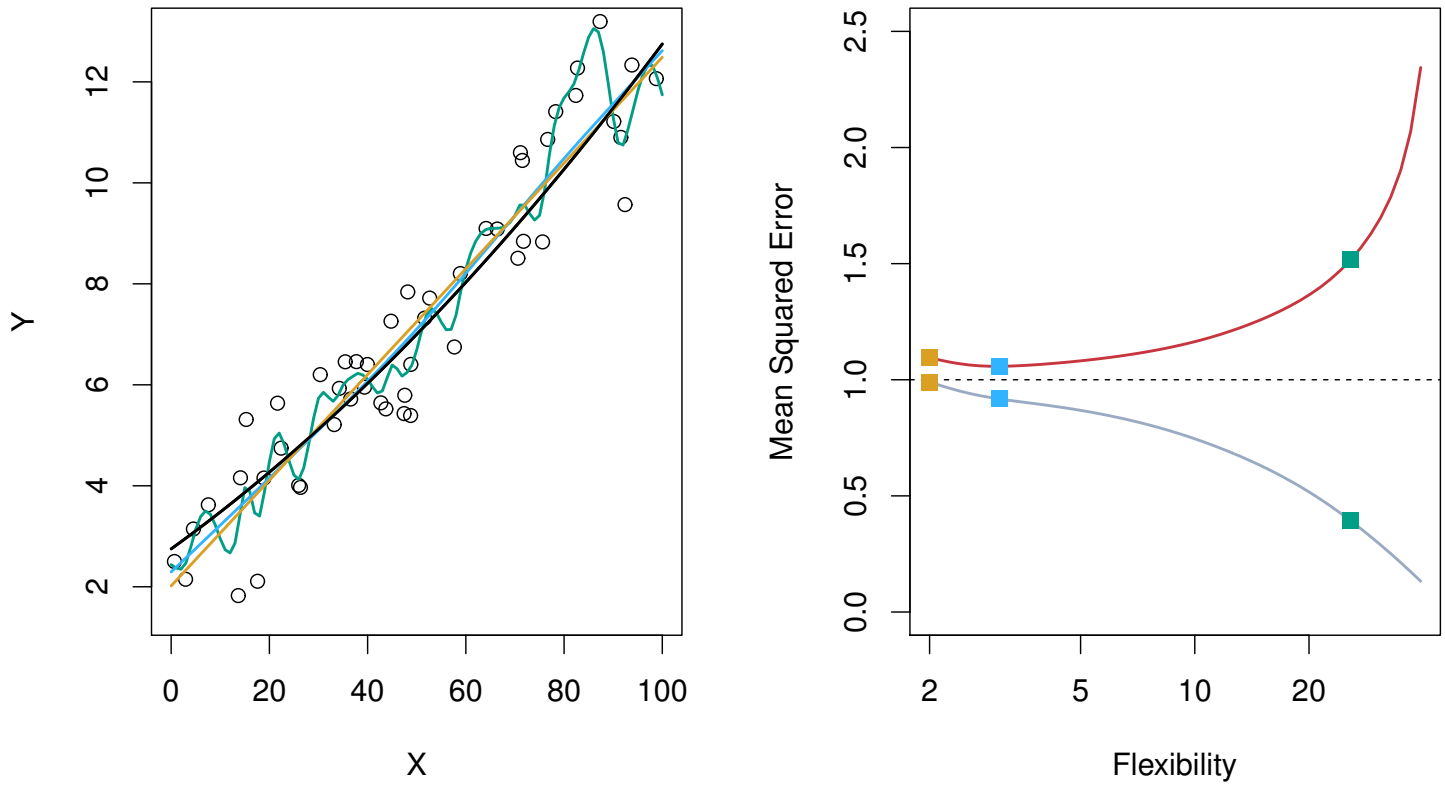

Figure 2.10: Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.

The Bias-Variance Trade-Off

- **Expected test MSE** is decomposed into three quantities:

$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \text{Var}(\hat{f}(x_0)) + \left[\text{Bias}(\hat{f}(x_0))\right]^2 + \text{Var}(\epsilon) \qquad (2.7)$$

  - the **variance** of $\hat{f}(x_0)$
  - the squared **bias** of $\hat{f}(x_0)$
  - the variance of the error

- To minimize the expected test error,
  - we need to select a statistical method that simultaneously achieves **low variance** and **low bias**
  - Note that we cannot reduce $\text{Var}(\epsilon)$

The Bias-Variance Trade-Off

- Meaning of variance and bias of a statistical learning method
  - Variance
    - Variance refers to the amount by which $\hat{f}$ would change if we estimated it using a different training data set
    - Ideally the estimate for f should not vary too much between training data sets
    - If a method has high variance, small changes in the training data can result in large changes in $\hat{f}$
  - Bias
    - Bias refers to the error that is introduced by approximating a real-life problem by a much simpler model
    - Obviously the small bias is preferred
  - See Figure 2.11 for linear regression model

The Bias-Variance Trade-Off

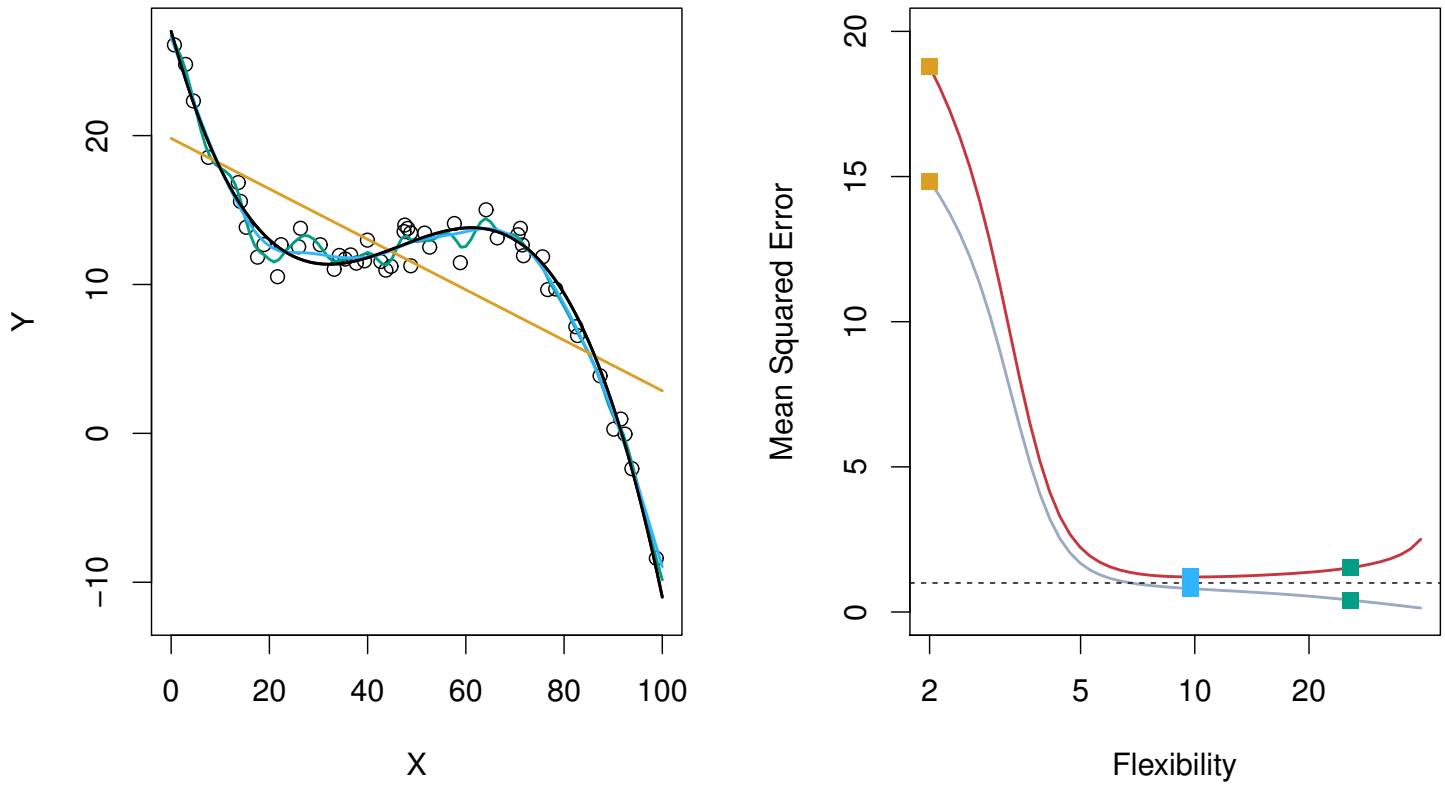

Figure 2.11: Details are as in Figure 2.9, using a different f that is far from linear. In this setting, linear regression provides a very poor fit to the data

The Bias-Variance Trade-Off

- General rule
  - More flexible model $\Rightarrow$ the variance will increase and the bias will decrease
  - The relative rate of change of variance and bias determines whether the test MSE increases or decreases
- The relationship between bias, variance, and test MSE in (2.7) and in Figure 2.12 is referred to as the **bias-variance trade-off**.
- Good test set performance of a statistical learning method requires low variance as well as low squared bias, leading to the smallest test MSE.

The Bias-Variance Trade-Off

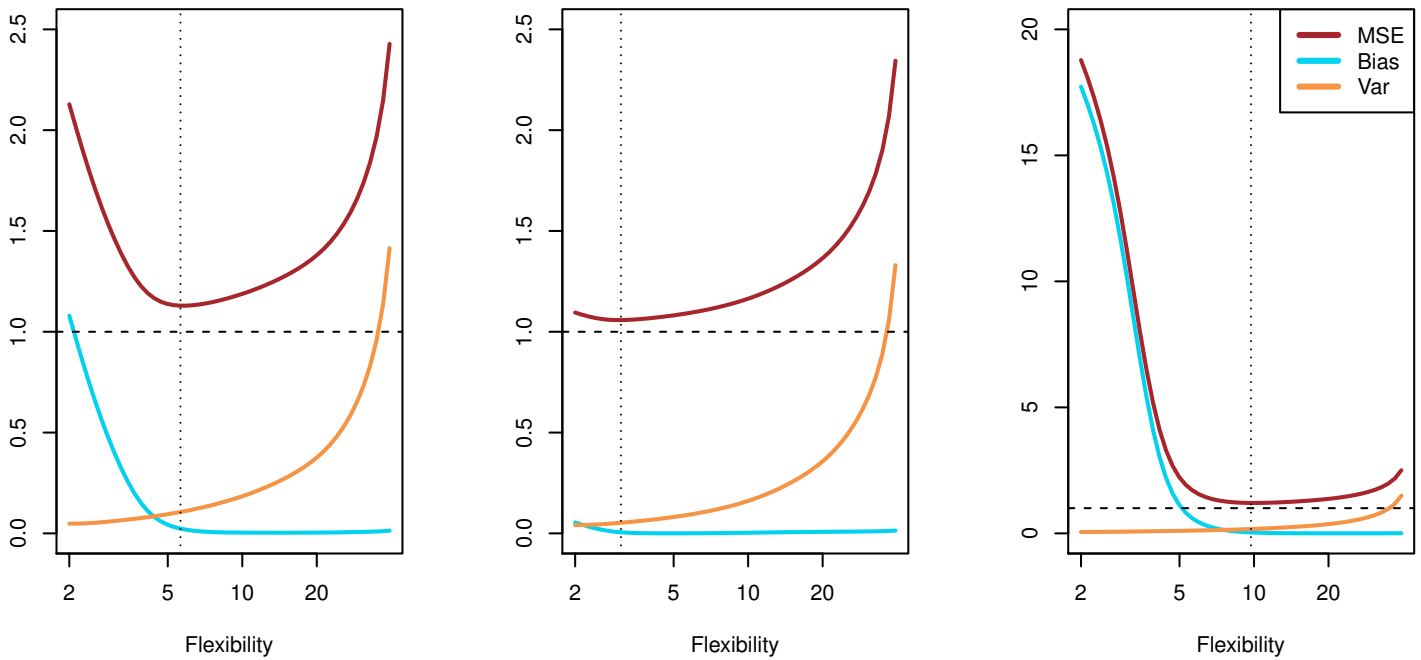

Figure 2.12: Squared bias (blue curve), variance (orange curve), $Var(\epsilon)$ (dashed line), and test MSE (red curve) for the three data sets in Figure 2.9-2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.

The Classification Setting

- Many of concepts in regression setting, such as the bias-variance trade-off, transfer over to the classification setting with some modifications.
  - Training set: $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$
  - y_i is no longer numerical
  - The objective of classification is to estimate the **classifier**, f, from the training data set
- **Training error rate**

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{2.8}$$

  - $\hat{y}_i$ is the predicted class label for the ith observation using $\hat{f}$
- **Test error rate** associated with test observations (x_0, y_0)

$$\text{Ave}(I(y_0 \neq \hat{y}_0)) \tag{2.9}$$

  - A good classifier is one for which the test error (2.9) is smallest

Bayes Classifier

- Conditional probability to assign the observation to the jth class

$$\Pr(Y = j|X = x_0) \tag{2.10}$$

- **Bayes classifier** is the classifier that assigns each observation to the most likely class, given its predictor values
  - In two-class classification ($j = 1, 2$), Bayes classifier corresponds to predicting class 1 if $\Pr(Y = 1|X = x_0) > 0.5$, and class 2 otherwise
  - **Bayes decision boundary** : the points where the probability is exactly 50%
  - The Bayes classifier's prediction is determined by the Bayes decision boundary

- **Bayes error rate**
  - The Bayes classifier produces the lowest possible test error rate, called the Bayes error rate
  - The error rate at $X = x_0$ will be $1 - \max_j \Pr(Y = j|X = x_0)$
  - Bayes error rate is given by

$$1 - E\left(\max_j \Pr(Y = j|X)\right) \tag{2.11}$$

Bayes Classifier

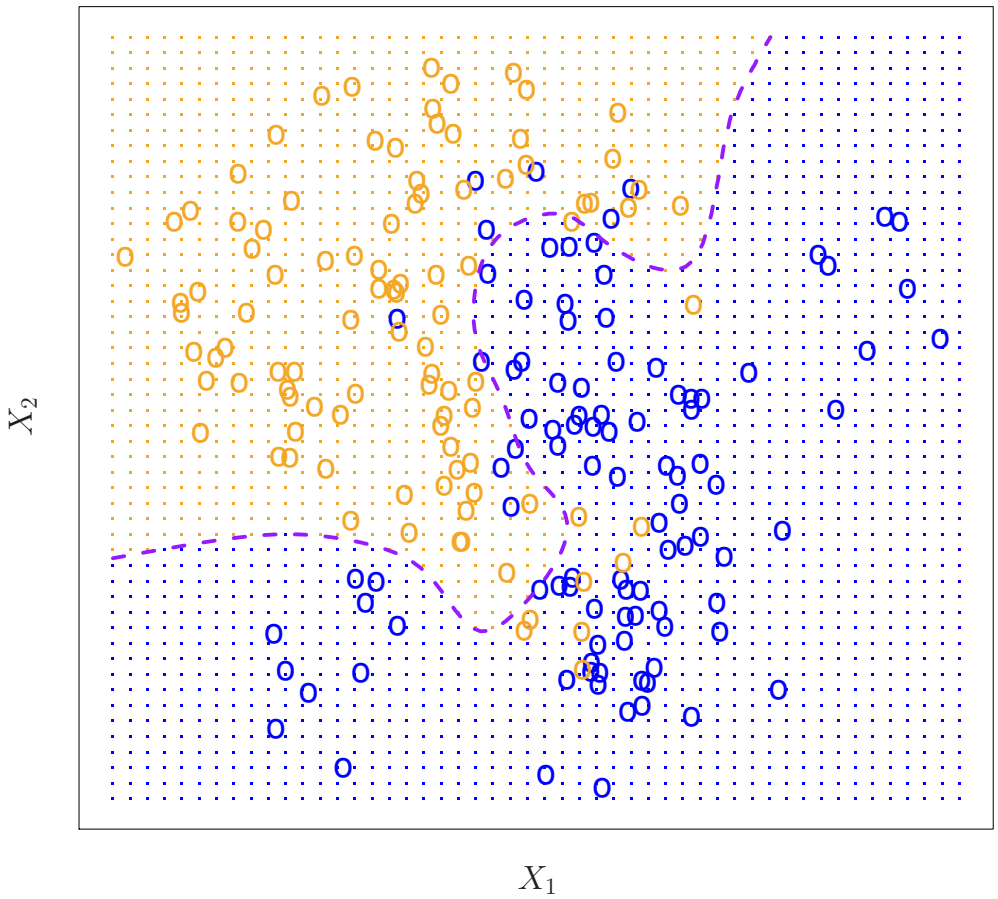

Figure 2.13: A simulated data set consisting of 100 observations in each of two groups, indicated in blue and in orange. The purple dashed line represents tow Bayes decision boundary. The orange background grid indicates the region in which a test observation will be assigned to the orange class, and the blue background grid indicates the region in which a test observation will be assigned to the blue class.

K-Nearest Neighbors

- For real data, computing the Bayes classifier is impossible because we do not know the conditional distribution of Y given X
- **K-nearest neighbors** (KNN) classifier classifies a given observation to the class with highest *estimated* probability

$$\Pr(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j) \tag{2.12}$$

  - $\mathcal{N}_0$ is the set of K points in the training data that are closest to x_0
  - KNN classifies the test observation x_0 to the class with the largest probability

Bayes Classifier

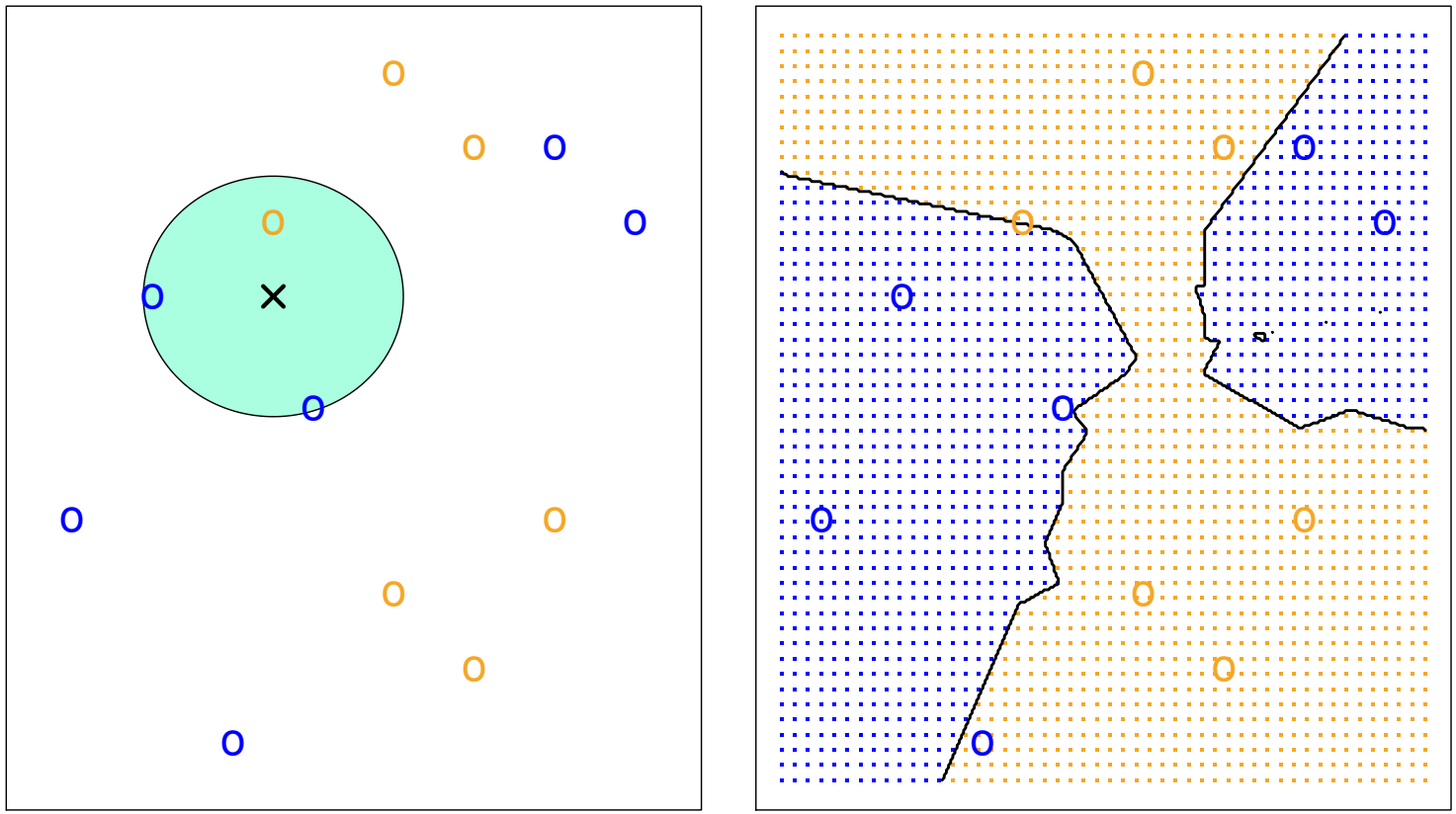

Figure 2.14: The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations. Left: a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified, and it is predicted that the test observation belongs to the most commonly-occurring class, in this case blue. Right: KNN decision boundary for this example is shown in black. The blue grid indicates the region in which a test observation will be assigned to the blue class, and the orange grid indicates the region in which it will be assigned to the orange class.

K-Nearest Neighbors

- With a good choice of K, the KNN classifier is surprisingly close to the optimal Bayes classifier (See Figure 2.15)
- The choice of K has a drastic effect on the KNN classifier obtained (See Figure 2.16)
  - When $K = 1$, the decision boundary is overly flexible and finds patterns in the data that don't correspond to the Bayes decision boundary (low bias, high variance)
  - When $K = 100$, the method becomes less flexible and produces a decision boundary that is close to linear (low variance, high bias)
  - Neither $K = 1$ or $K = 100$ give good prediction: they have test error rates of 0.1695 and 0.1925, respectively
- In general, as we use more flexible classification methods, the training error rate will decline. However, the test error exhibits a characteristic U-shape (See Figure 2.17)

K-Nearest Neighbors

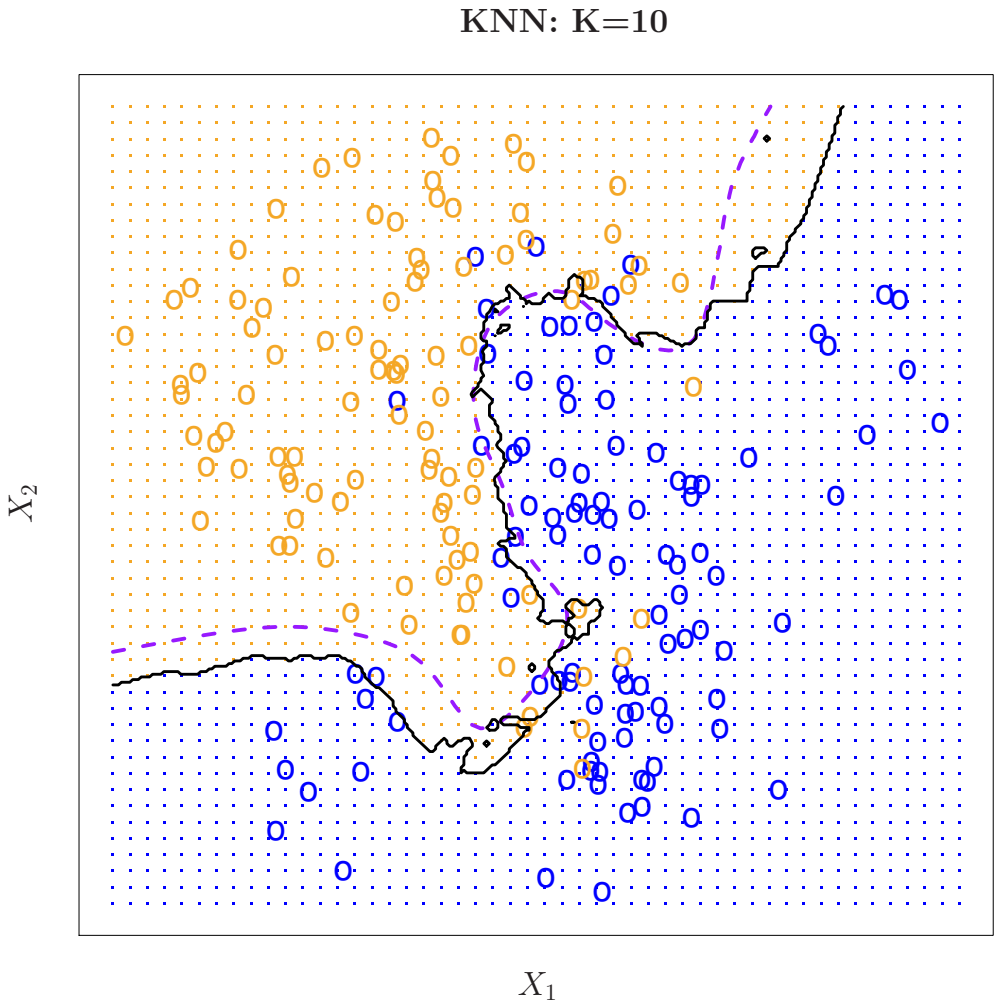

Figure 2.15: The black curve indicates the KNN decision boundary on the data from Figure 2.12, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.

K-Nearest Neighbors

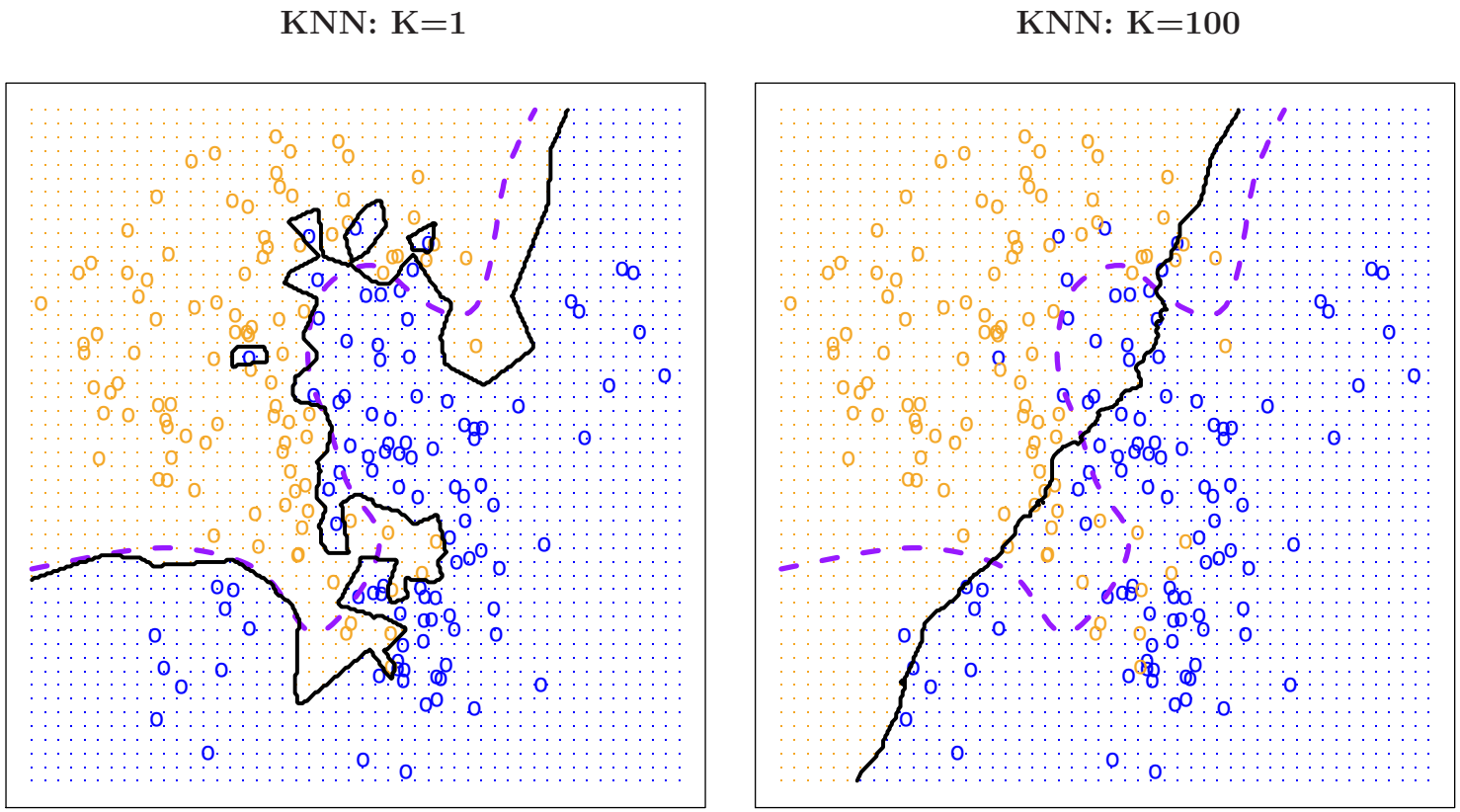

Figure 2.16: A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple dashed line.

K-Nearest Neighbors

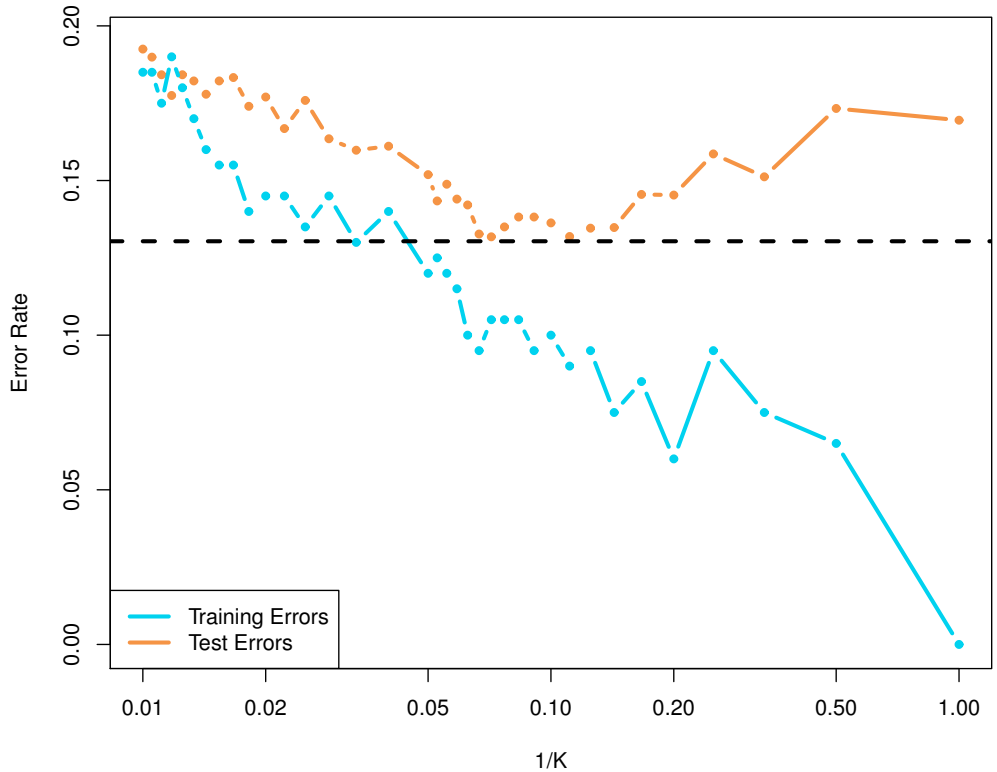

Figure 2.17: The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes error rate. The jumpiness of the curves is due to the small size of the training data set.

Lab: Introduction to R