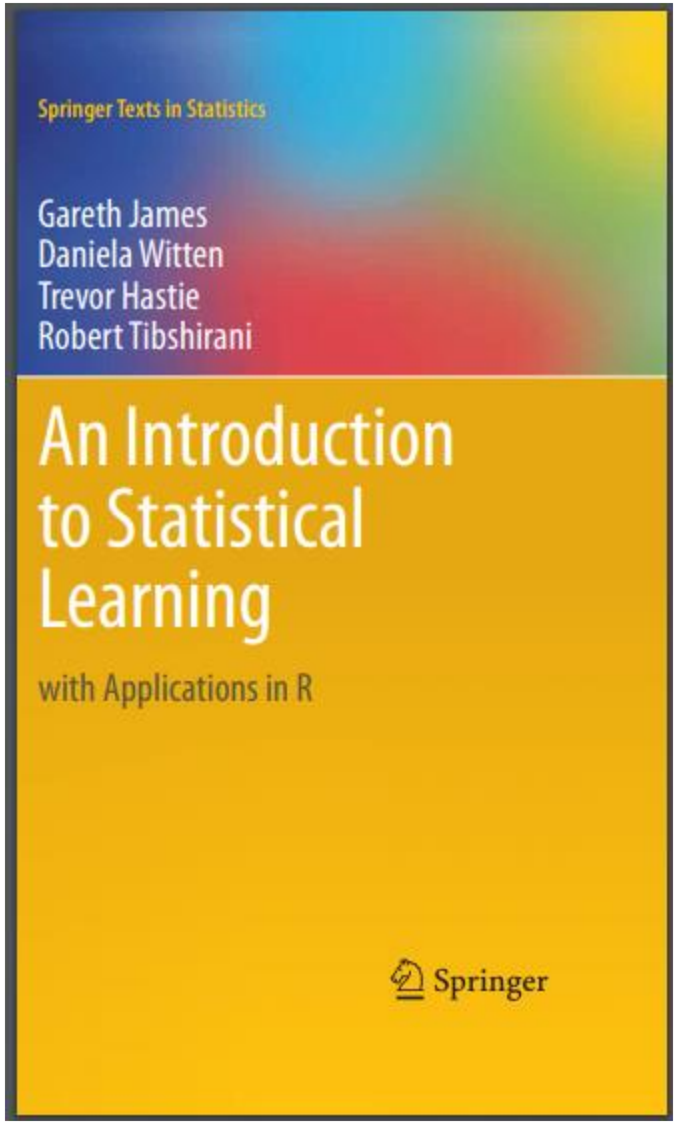
Statistical Learning

https://github.com/ggorr/Machine-Learning/tree/master/ISLR

교재

- An Introduction to Statistical Learning
  with Applications in R
  - G. James, D. Witten, T. Hastie and R. Tibshirani

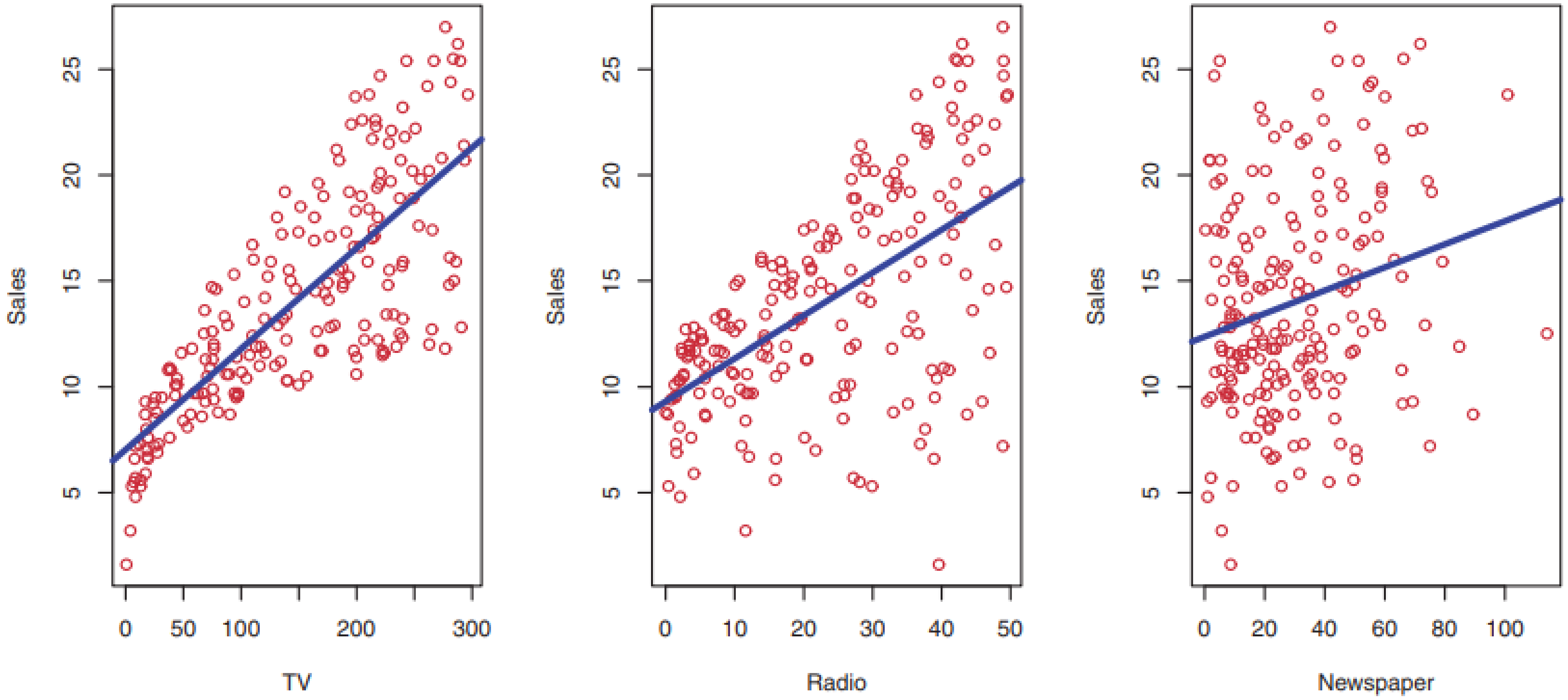


2 Statistical Learning

- 2.1 What Is Statistical Learning?
- 2.2 Assessing Model Accuracy
- 2.3 Lab: Introduction to R
- 2.4 Exercises

2.1 What Is Statistical Learning?

- Advertising data set



- Input variables
  = predictors, independent variables, features, variables
  - Notation: X, X_i
- Output variable
  = response, dependent variable
  - Notation: Y
- Example: Advertising
  - X_1 = TV, X_2 = radio, X_3 = newspaper
  - Y = sales

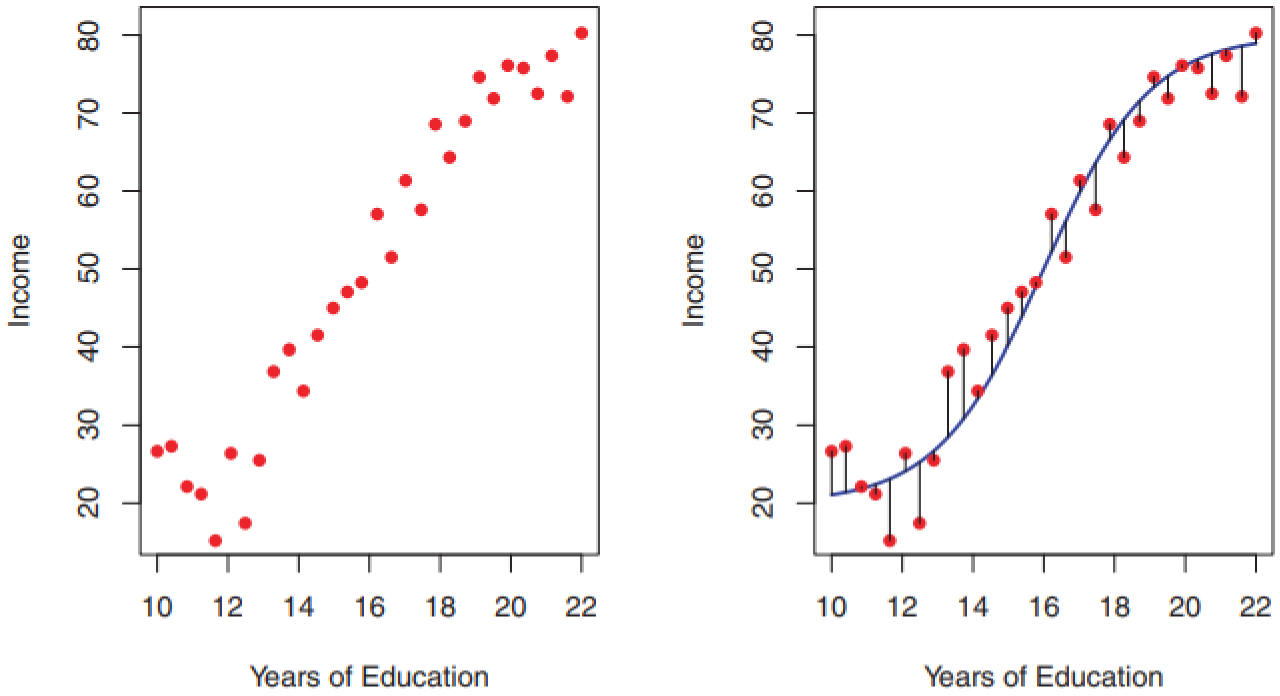
Relationship between Predictors and Response

- We assume that

$$Y = f(X) + \epsilon$$

  - $X = (X_1, X_2, \dots, X_p)$ - predictors
  - ϵ : random error
    independent of X
    $E[\epsilon] = 0$, i.e. zero mean

Example


Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education
Income
80
70
60
50
40
30
20
10
12
14
16
18
20
22
Years of Education

What Is Statistical Learning?

- Statistical learning refers to
  a set of approaches for estimating f.

2.1.1 Why Estimate f ?

- Reason
  - Prediction
  - Inference

Prediction

- Prediction
  - $\hat{Y} = \hat{f}(X)$
    - $\hat{Y}$ is an estimate of Y
  - $\hat{f}$ is treated as a black box
    - Not concerned with the exact form
    - Interested in the accuracy

Example

- $X_1, \dots, X_p$: characteristics of a blood sample
- Y: risk for an adverse reaction

Accuracy of $\hat{Y}$

- Error

$$\mathbb{E}\left[\left(Y-\hat{Y}\right)^2\right] = \mathbb{E}\left[\left(f(X)+\epsilon-\hat{f}(X)\right)^2\right]$$
$$= \mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right] + \mathrm{Var}(\epsilon)$$

- $\mathbb{E}\left[\left(f(X)-\hat{f}(X)\right)^2\right]$ - Reducible error
    - Inaccuracy of $\hat{f}$
    - We can potentially improve the accuracy of $\hat{f}$
- $\mathrm{Var}(\epsilon)$ - Irreducible error
    - ϵ cannot be predicted using X

Inference

- 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?

Example

- Advertising data set
  - Which media contribute to sales?
  - Which media generate the biggest boost in sales
  - How much increase in sales is associated with a given increase in TV advertising?

2.1.2 How Do We Estimate f?

- Training data set: observations
  - Input variables
    - p : number of predictors
    - n : number of data
    - x_{ij} for $i = 1, \dots, n$ and $j = 1, \dots, p$
    - $x_i = (x_{i1}, \dots, x_{ip})$
  - Output variable
    - y_i for $i = 1, \dots, n$

Training

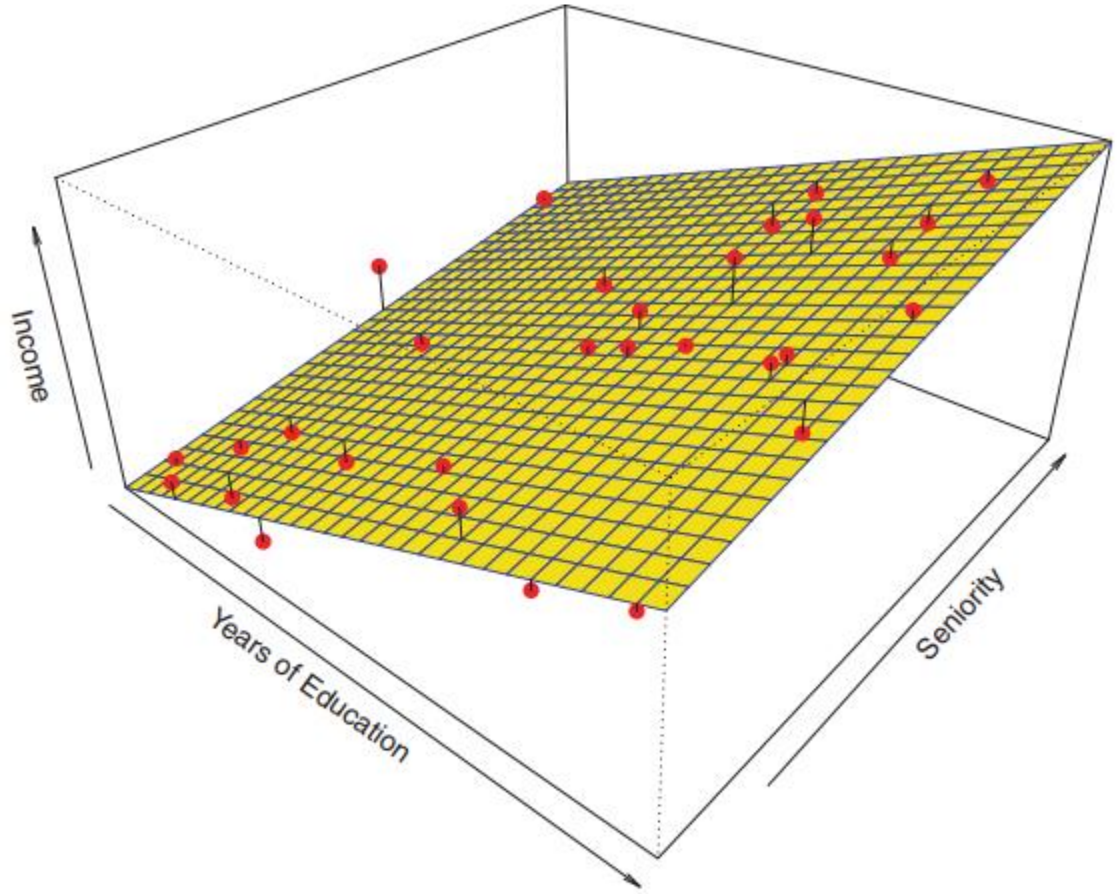
- Parametric method
- Non-parametric method

Parametric Method

- To make an assumption about the functional form of f
- Example: linear model
  - Assume
    $$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$
    and estimate the parameters $\beta_0, \ldots, \beta_p$

Example

- income $\approx \beta_0 + \beta_1 \times$ education $+ \beta_2 \times$ seniority



Model

- A model is a form of f
  - Example: linear model assumes that f is linear

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Model selection

- A model which is too far from true f
  - Estimation is poor
- A flexible model
  - Can fit many different possible functional forms
  - Requires estimating many parameters
  - Overfitting

Non-parametric Methods

- Do not make explicit assumptions about f
- A very large number of observations is required
- Example: thin-plate spline

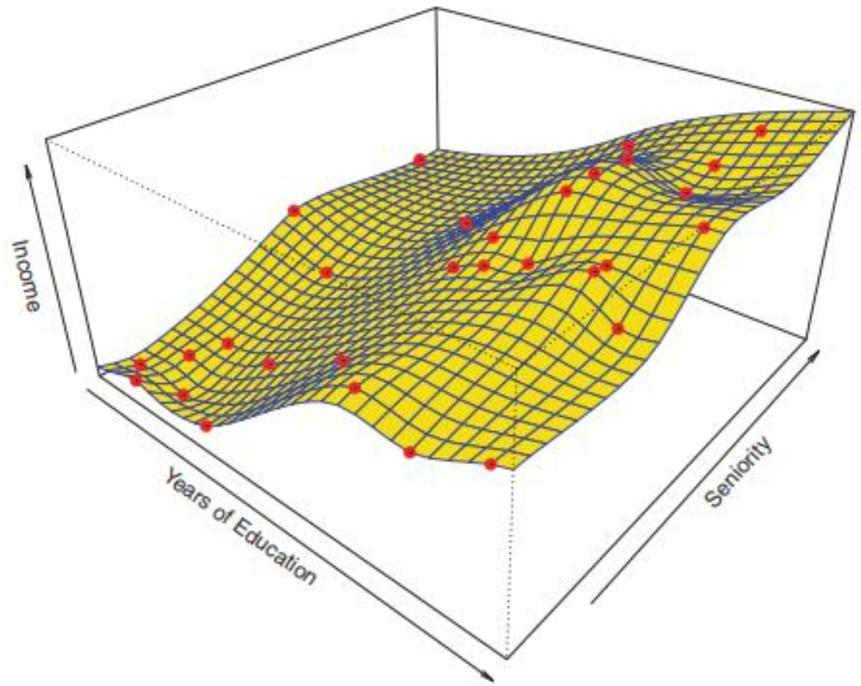


FIGURE 2.6. *A rough thin-plate spline fit to the* `Income` *data from Figure 2.3. This fit makes zero errors on the training data.*

2.1.3 The Trade-Off Between Prediction Accuracy and Model Interpretability

- Why would we choose a more restrictive method instead of a very flexible approach?
- In inference, restrictive models are much more interpretable

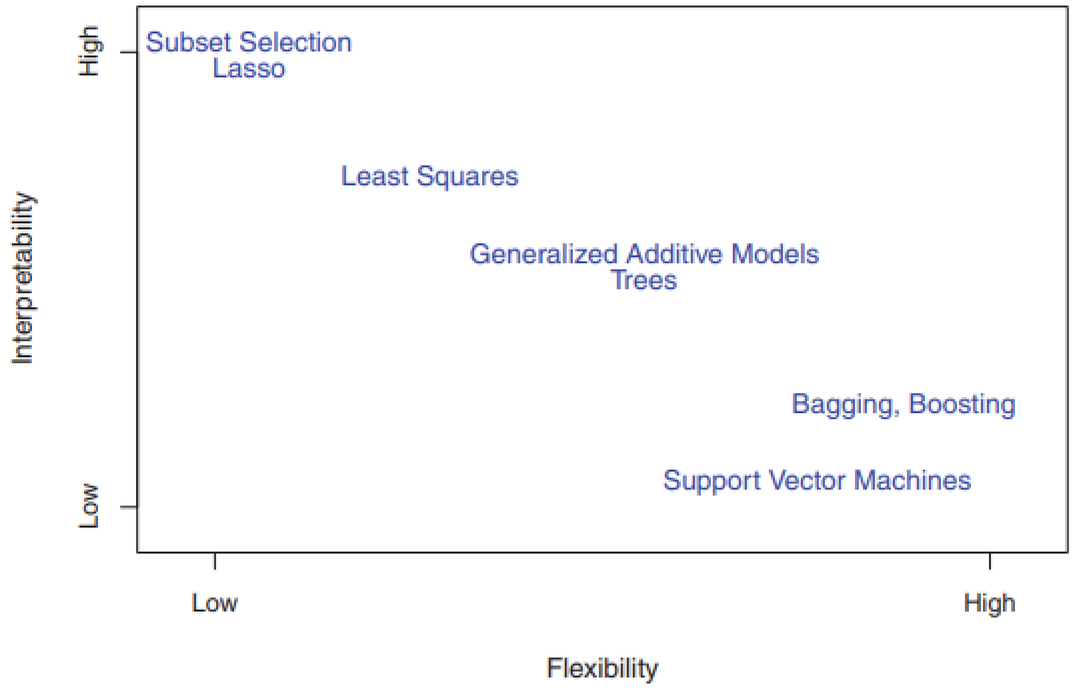
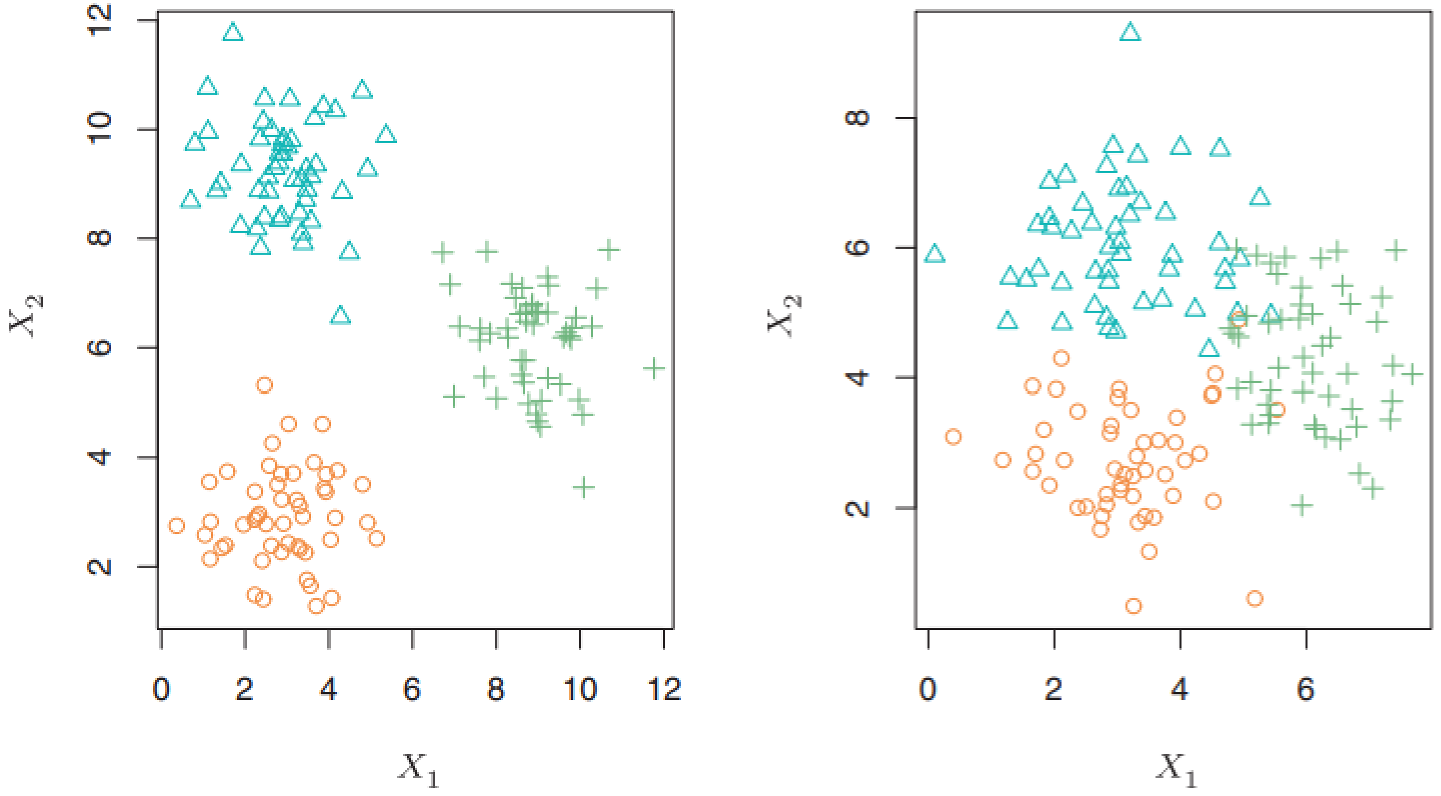


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

2.1.4 Supervised Versus Unsupervised Learning

- Supervised learning
  - Use responses for training data
- unsupervised learning
  - No response
  - Example: cluster analysis

Clustering



2.1.5 Regression Versus Classification Problems

- Regression
  - Quantitative
  - Example: a person's age, height, or income, the value of a house, the price of a stock
- Classification
  - Qualitative a.k.a. categorical
  - Example: a person's gender (male or female), the brand of product purchased (brand A, B, or C)

2.2 Assessing Model Accuracy

- 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
- To decide the best method for given data set

2.2.1 Measuring the Quality of Fit

- Mean squared error (MSE)

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

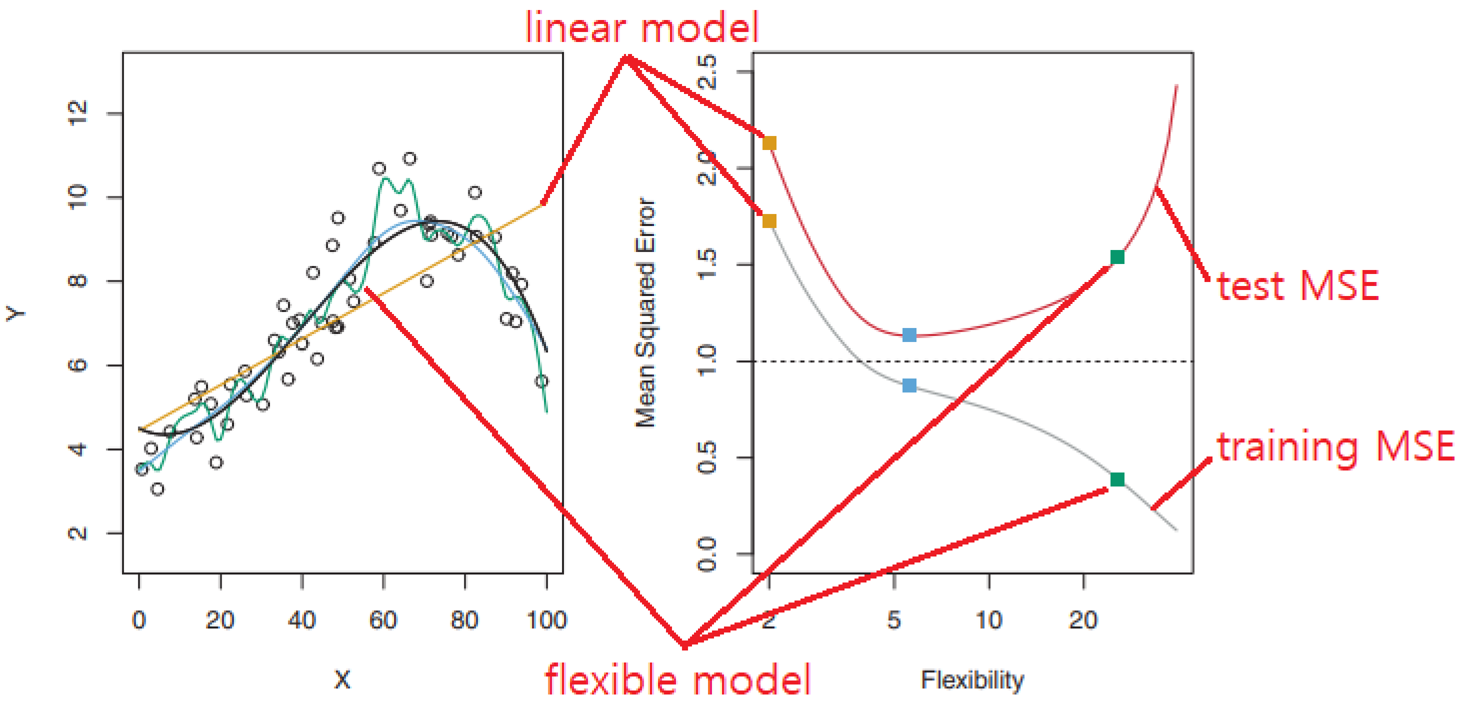
Training and Test

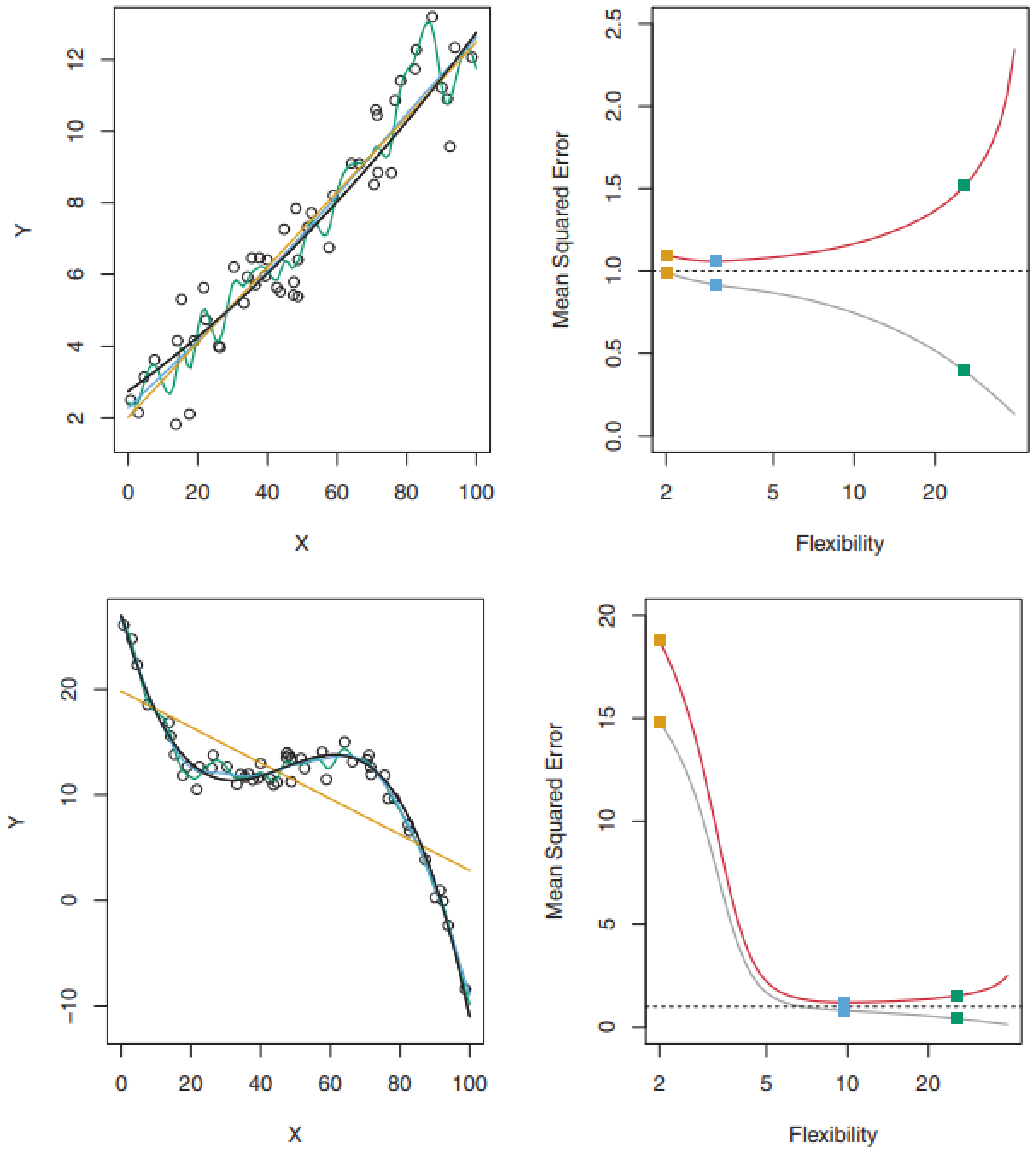
- Training MSE
  - MSE for training data
- Test MSE
  - MSE for test data
- Test data = test observations
  - Unseen data during training process
  - We are interested in the accuracy of the predictions on previously unseen test data

- Training data
  - $\{(x_1, y_1), \ldots, (x_n, y_n)\}$
- Test data
  - (x_0, y_0)

- If we have a large number of test data, we could compute

$$\mathrm{Ave}\left(\hat{f}(x_0) - y_0\right)^2$$


linear model
flexible model
test MSE
training MSE
Y
X
Mean Squared Error
Flexibility



Y
X
Mean Squared Error
Flexibility
Y
X
Mean Squared Error
Flexibility

2.2.2 The Bias-Variance Trade-Off

- For test data (x_0, y_0),
$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right]$$
  is called the expected test MSE
- Expected test MSE
$$\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] = \mathrm{Var}\left(\hat{f}(x_0)\right) + \mathrm{Bias}\left(\hat{f}(x_0)\right)^2 + \mathrm{Var}(\epsilon)$$
  where
$$\mathrm{Bias}\left(\hat{f}(x_0)\right) = \mathbb{E}\left[f(x_0) - \hat{f}(x_0)\right] = \mathbb{E}[y_0 - \hat{y}_0]$$

$$
\begin{aligned}
\mathbb{E}\left[\left(y_0 - \hat{f}(x_0)\right)^2\right] &= \mathbb{E}\left[\left(f(x_0) + \epsilon - \hat{f}(x_0)\right)^2\right] \\
&= \mathbb{E}\left[\left(f(x_0) - \hat{f}(x_0)\right)^2\right] + \mathrm{Var}(\epsilon) \\
&= \mathbb{E}\left[\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)] + \mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)^2\right] + \mathrm{Var}(\epsilon) \\
&= \mathbb{E}\left[\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right] + \mathbb{E}\left[\left(E[\hat{f}(x_0)] - \hat{f}(x_0)\right)^2\right] \\
&\qquad + 2\mathbb{E}\left[\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\left(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)\right] + \mathrm{Var}(\epsilon) \\
&= \left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2 + \mathrm{Var}\left(\hat{f}(x_0)\right) \\
&\qquad + 2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x_0)] - \hat{f}(x_0)\right)\right] + \mathrm{Var}(\epsilon) \\
&= \mathrm{Bias}\left(\hat{f}(x_0)\right)^2 + \mathrm{Var}\left(\hat{f}(x_0)\right) \\
&\qquad + 2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\left(\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)]\ \right) + \mathrm{Var}(\epsilon) \\
&= \mathrm{Var}\left(\hat{f}(x_0)\right) + \mathrm{Bias}\left(\hat{f}(x_0)\right)^2 + \mathrm{Var}(\epsilon)
\end{aligned}
$$

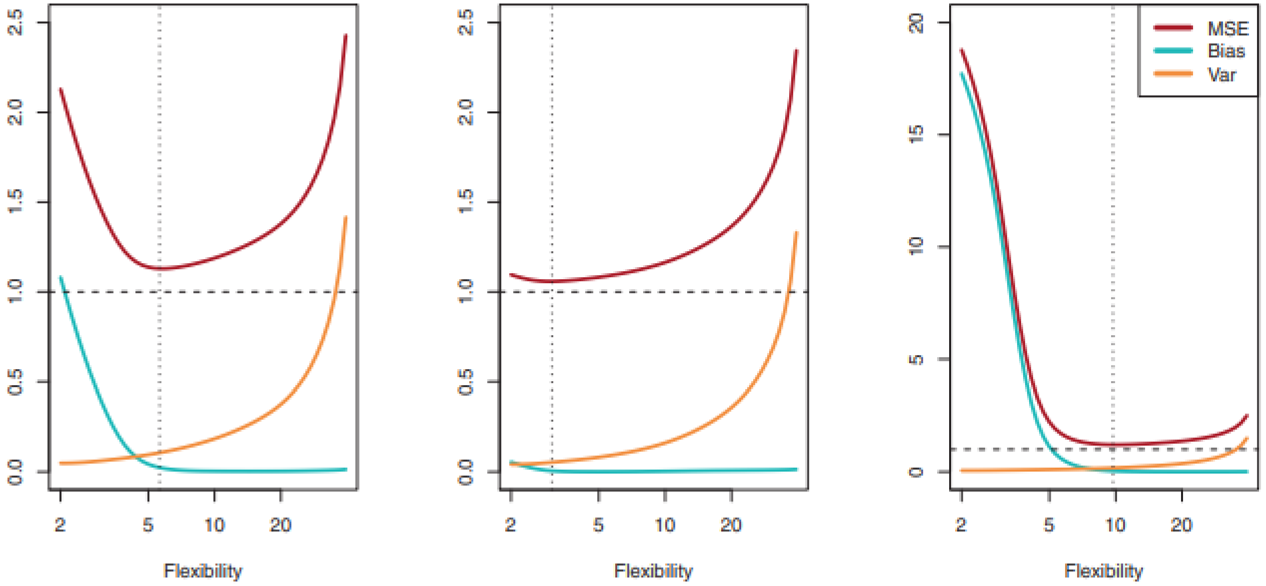


FIGURE 2.12. *Squared bias (blue curve), variance (orange curve),* $Var(\epsilon)$ *(dashed line), and test MSE (red curve) for the three data sets in Figures 2.9–2.11. The vertical dotted line indicates the flexibility level corresponding to the smallest test MSE.*

2.2.3 The Classification Setting

- Training error

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

where $I(y_i \neq \hat{y}_i) = \begin{cases} 1, & if\ y_i \neq \hat{y}_i \\ 0, & if\ y_i = \hat{y}_i \end{cases}$, called an indicator variable

- Test error

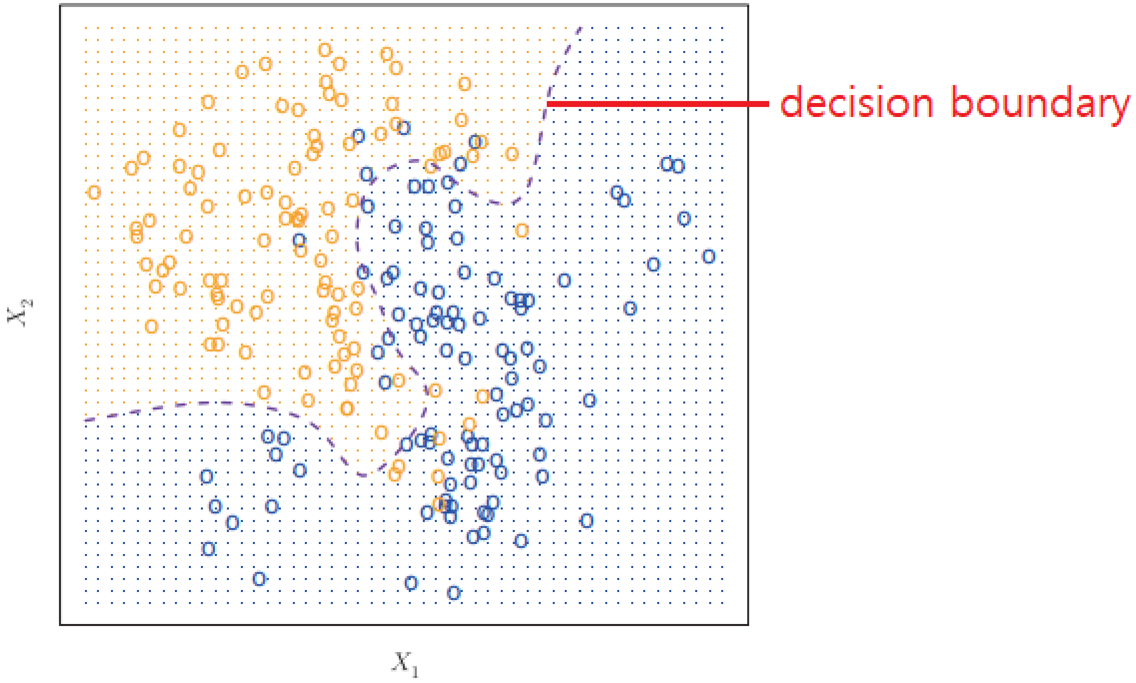
$$\text{Ave}\big(I(y_0 \neq \hat{y}_0)\big)$$

The Bayes Classifier

- Bayes classifier
  - assigns each observation to the most likely class

$$\hat{y}_i = \underset{j}{\operatorname{argmax}} \Pr(Y = j \mid X = x_i)$$

- Bayes decision boundary
  - the boundary of area



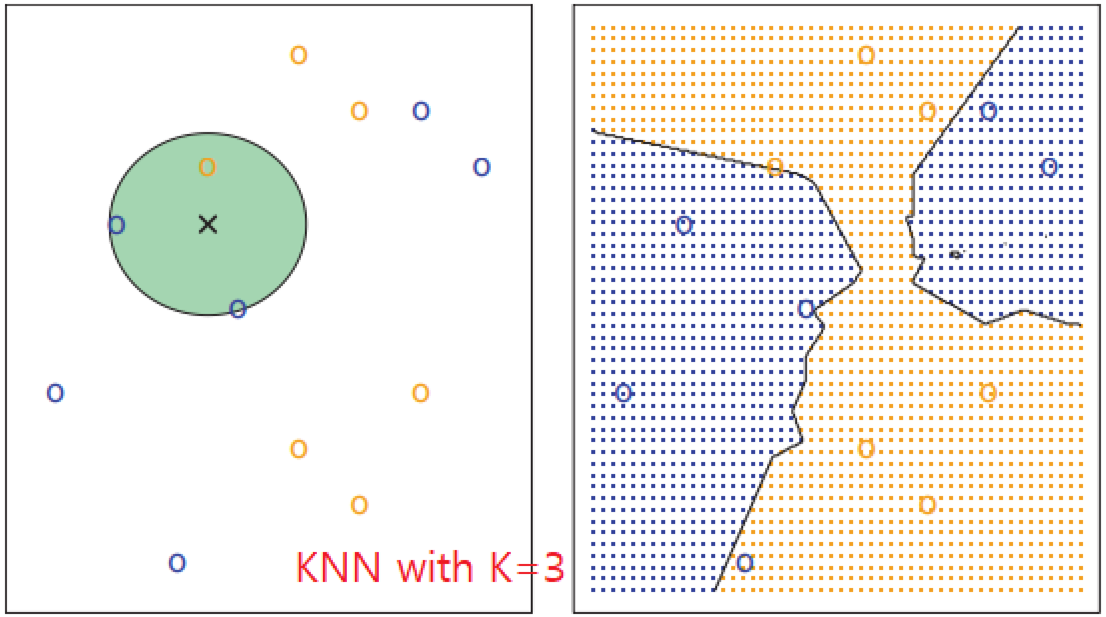
- Bayes error rate

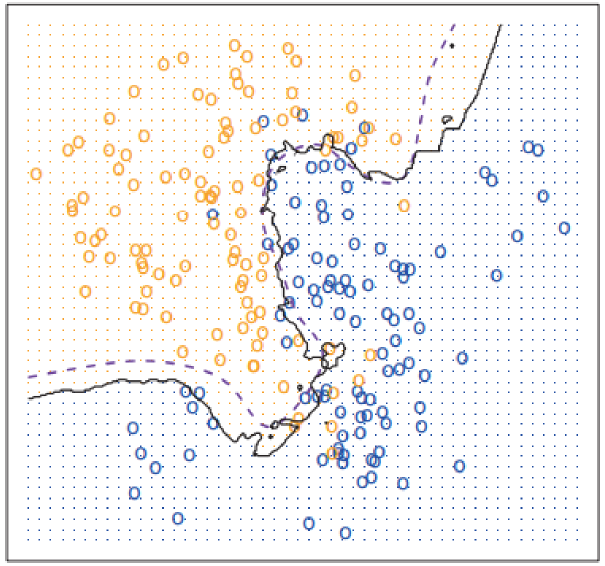
$$1 - \mathbb{E}\left[\max_j \Pr(Y = j|X)\right]$$

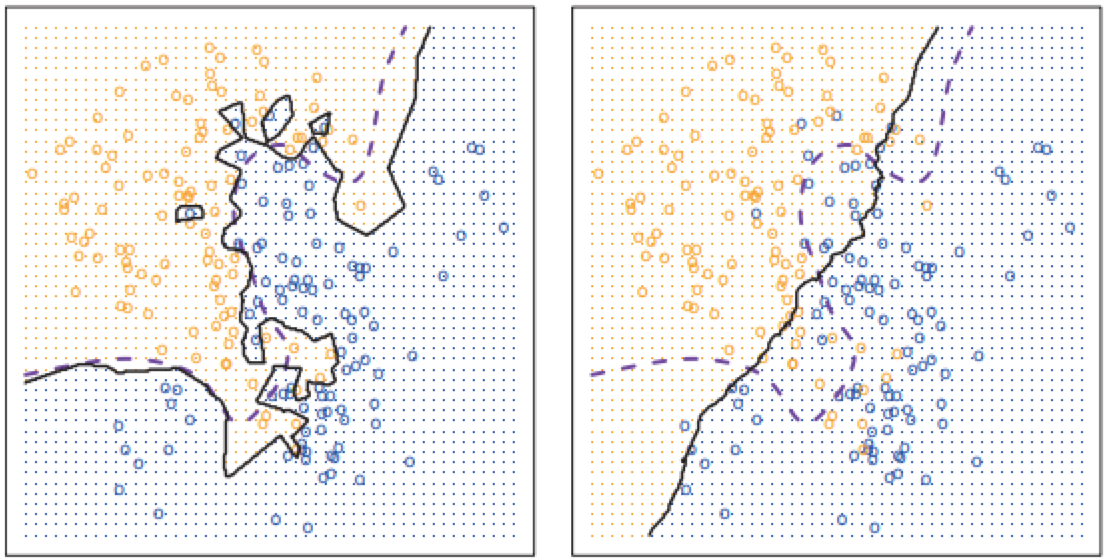
K-Nearest Neighbors

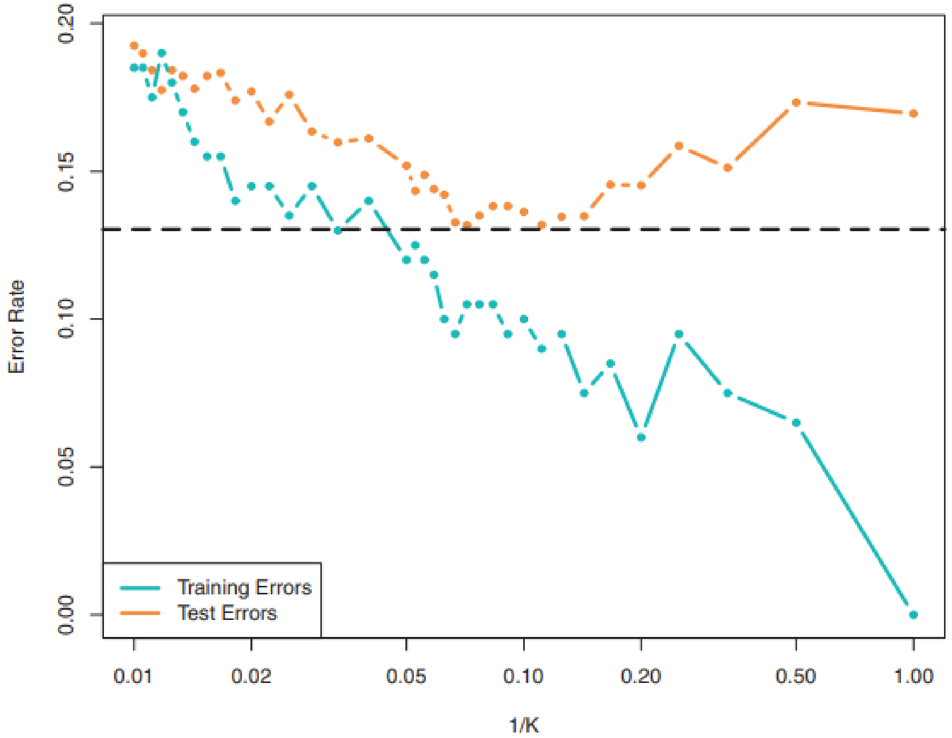
- K-Nearest Neighbors(KNN)

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j)$$




KNN: K=10


KNN: K=1
KNN: K=100


Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

2.4 Exercise

- Solve with pen
- Solve with Python
- Compute training error rate
- Compute test error rate

7. The table below provides a training data set containing six observations, three predictors, and one qualitative response variable.

Obs.	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Suppose we wish to use this data set to make a prediction for Y when $X_1 = X_2 = X_3 = 0$ using K-nearest neighbors.

(a) Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

(b) What is our prediction with $K = 1$? Why?

(c) What is our prediction with $K = 3$? Why?

(d) If the Bayes decision boundary in this problem is highly non-linear, then would we expect the *best* value for K to be large or small? Why?