STA 5820
Chapter 2
Statistical Learning

Kazuhiko Shinki

Wayne State University

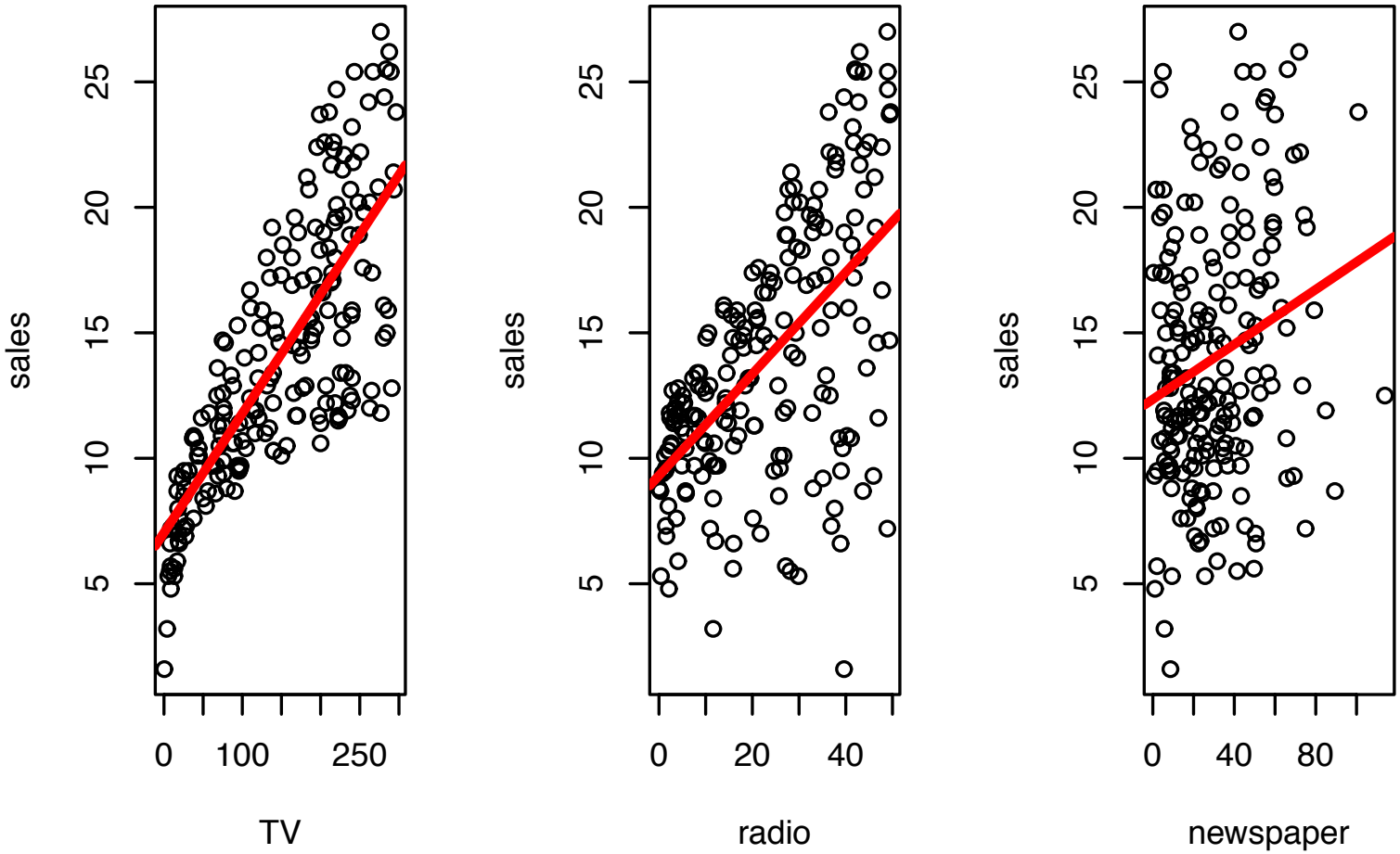
2.1 What is statistical learning?

Example:

- Objective: We want to increase sales ($\boldsymbol{Y}$) of a particular product.
- Data: `sales`, advertising budgets for `TV` ($\boldsymbol{X_1}$), `radio` ($\boldsymbol{X_2}$) and `newspaper` ($\boldsymbol{X_3}$) for 200 markets.
- Action: Which advertising budget should we increase: TV, radio or newspaper?

2.1.1 Why estimating *f*?

Figure 2-1: Estimate α_i and β_i in a linear model $Y = \alpha_i + \beta_i X_i + \epsilon$ ($i = 1, 2, 3$). i with the largest β_i is the most effective advertising channel.



2.1.1 Why estimating f?

Other models:

A multiple linear model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{1}$$

The estimated value of β_i as $\hat{\beta}_i$ ($i = 0, 1, 2, 3$) is denoted by $\hat{\beta}_i$.

More generally,

$$Y = f(X_1, X_2, X_3) + \epsilon \tag{2}$$

An estimated function of f is denoted by $\hat{f}$.

Conventionally $X = (X_1, X_2, X_3)$ is assumed deterministic, ϵ is a random variable (so is Y).

2.1.1 Why estimating f?

Objectives:

- **Forecasting**: When a new observation (x_1, x_2, x_3) is available but Y-value is unknown, Y is estimated by

$$\hat{Y} = \hat{f}(x_1, x_2, x_3) \tag{3}$$

- **Inference**: In the model (1), we may want to:
  - estimate the size of β_i $(i = 1, 2, 3)$. For example, we want to see how much sales increases as we spend \$1,000 additional dollars on advertisements on newspaper.
  - test if $\beta_i = 0$ $(i = 1, 2, 3)$. For example, we want to see if advertisements on newspaper is statistically meaningful.

2.1.1 Why estimating f?

A measure of error:

Let $(x_1, y_1), \cdots, (x_n, y_n)$ be observations and $\hat{y}_i = \hat{f}(x_i)$ be the predicted value for y_i $(1 \leq i \leq n)$.

The mean square error (MSE) is defined by

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{4}$$

The MSE is the sample version of the expected size of squared error $E[(Y - \hat{Y})^2]$.

2.1.1 Why estimating f?

Reducible and irreducible errors:

Assume that (i) ϵ's are independent for different observations.

Consider the situation that $\boldsymbol{f}$ is estimated by observations $(\boldsymbol{X_n}, \boldsymbol{Y_n}) = (\boldsymbol{X_{n,1}}, \cdots, \boldsymbol{X_{n,3}}, \boldsymbol{Y_n})$ $(\boldsymbol{n} = \boldsymbol{1}, \cdots, \boldsymbol{N})$.

Want to evaluate the size of the error (i.e., variance of the error) when $\boldsymbol{Y}$ is forecasted by $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{f}}(\boldsymbol{X})$ for a new observation $(\boldsymbol{X}, \boldsymbol{Y})$.

$$\begin{aligned} \boldsymbol{E}(\boldsymbol{Y} - \hat{\boldsymbol{Y}})^2 &= \boldsymbol{E}(\boldsymbol{f}(\boldsymbol{X}) + \epsilon - \hat{\boldsymbol{f}}(\boldsymbol{X}))^2 \\ &= \boldsymbol{E}(\boldsymbol{f}(\boldsymbol{X}) - \hat{\boldsymbol{f}}(\boldsymbol{X}))^2 + 2\boldsymbol{E}\epsilon(\boldsymbol{f}(\boldsymbol{X}) - \hat{\boldsymbol{f}}(\boldsymbol{X})) + \boldsymbol{E}\epsilon^2 \\ &= \boldsymbol{E}(\boldsymbol{f}(\boldsymbol{X}) - \hat{\boldsymbol{f}}(\boldsymbol{X}))^2 + \mathrm{Var}\epsilon^2 \end{aligned}$$

Note that the second term in the second line vanishes because ϵ is independent of $\hat{\boldsymbol{f}}(\boldsymbol{X})$ since $\hat{\boldsymbol{X}}$ is estimated by old observations.

2.1.1 Why estimating f?

In the most right hand side of the equation,

- $E(f(X) - \hat{f}(X))^2$ is called a reducible error since this term gets smaller if $\hat{f}$ is estimated accurately.
- $E\epsilon^2$ is called an irreducible error since this does not change regardless of our estimate of f.

In real world problems, the irreducible error exists because Y is partly detemined by factors not included in X. For example, sales (Y) may also be influenced by temperature, business cycle, rival companies' products, etc.

2.1.2 How do we estimate f?

- Parameteric methods: The number of parameters is finite and fixed .
  - Example: $\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$.
- Non-parameteric Methods: The model can not be represented by finite and fixed number of parameters.
  - Example: $\hat{Y}$ is defined by the mean Y-value of 5 nearest observations among $(X_1, Y_1), \cdots, (X_N, Y_N)$ (in terms of location of X). (Nearest neighborhood method).

2.1.3 The trade-off b/w Accuracy and Interpretability

More complicated model is often hard to justify by logic.

Example:

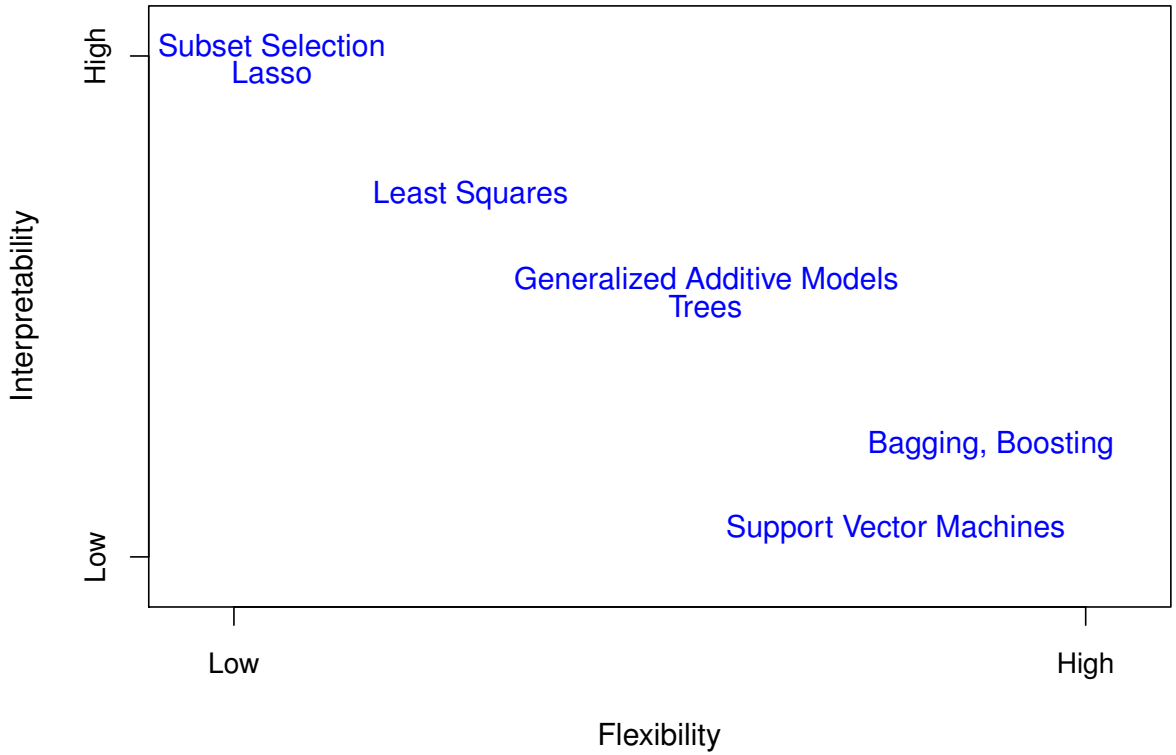
Let X be the diameter of a shell and Y be the weight of the shell.

$Y = \alpha X^3$ is justifiable.

$Y = \alpha X^{2.1}$ is hard to justify even if it is the statistically best model.

2.1.3 The trade-off b/w Accuracy and Interpretability

Figure 2-7: As the flexibility of the method increases, its interpretability decreases.



2.1.5 Regression vs Classification

As in chapter 1,

- Regression for quantitative response (typically real-valued).
- Classification for qualitative response (e.g., binary such as success/failure, and multi-class such as species { A, B, C }).

2.1.5 Regression vs Classification

In some cases, classification problem is regarded as a regression problem by considering probability.

Example: logistic regression

Want to explain gender (male or female) by some physical variables $\boldsymbol{X} = (\boldsymbol{X_1}, \cdots, \boldsymbol{X_p})$ such as weight, height, body fat percentage, body water percentage and age. When the probability of being female given $\boldsymbol{X}$ is considered, it can be modeled by a regression model.

$$\boldsymbol{P(female|X)} = \boldsymbol{f}(\beta_0 + \beta_1 \boldsymbol{X_1} + \cdots + \beta_p \boldsymbol{X_p}) \tag{5}$$

where $\boldsymbol{f}$ is called a link function which maps all real numbers to an interval $(0, 1)$.

2.1.5 Regression vs Classification

Predictors may be quantitative or qualitative (categorical), but the difference is not very important.

Example: dummy variables

Want to explain house sale price by square footage (quantitative), age (quantitative), and architecture style (qualitative; ranch, bungalow or colonial).
You can set up dummy variables for styles (bungalow and colonial). Then, usual regression models can be used.

Price	Sqft	Age	Style	Bungalow	Colonial
325000	1950	9	Ranch	0	0
349000	3120	31	Colonial	0	1
271000	2070	35	Colonial	0	1
197000	1470	50	Ranch	0	0
149000	1210	58	Bungalow	1	0

2.2 Assessing Model Accuracy

For regression problems,

- Measuring the Quality of fit (2.2.1).
- Bias-Variance trade-off (2.2.2).

For classification problems, both of the measuring and trade-off issues are discussed in

- Classification setting (2.2.3).

2.2.1 Measuring the Quality of Fit

Training MSE

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2. \quad (6)$$

is a good measure of fit. However, our ultimate objective is to minimize the expected error size for new observations, say (x_0, y_0), and not for training data $(x_1, y_1), \cdots, (x_n, y_n)$.

(6) is called a training MSE.

2.2.1 Measuring the Quality of Fit

A problem on training MSE

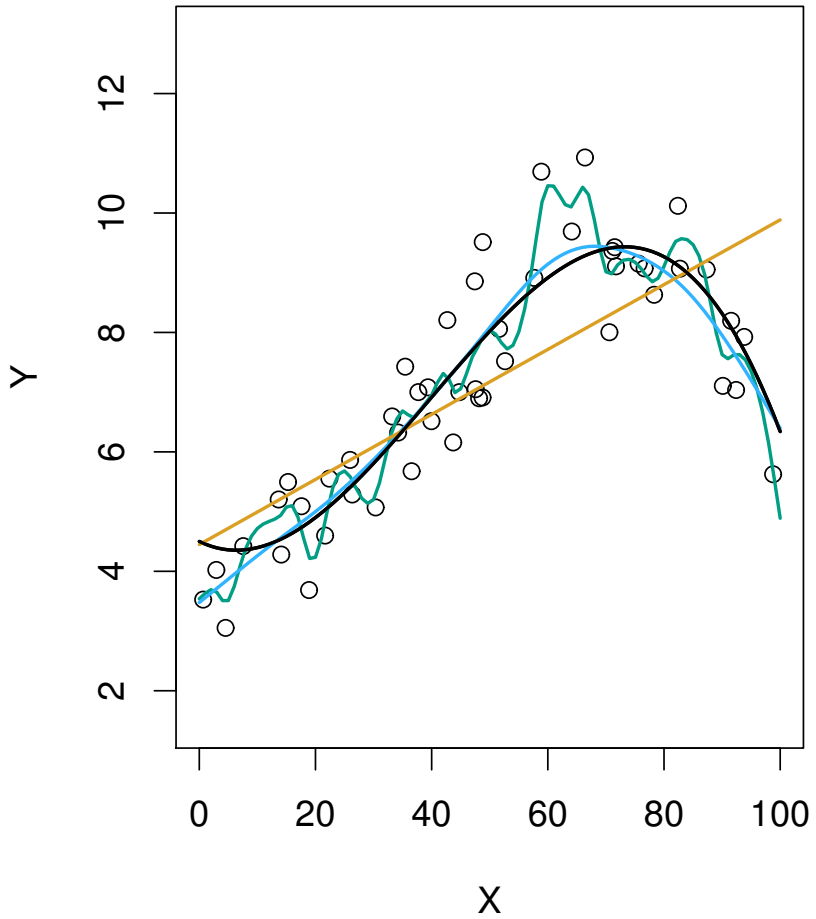
The training MSE gets (often arbitrarily) smaller as the model is more complex, but it does not mean that the expected error size for new observations gets smaller.

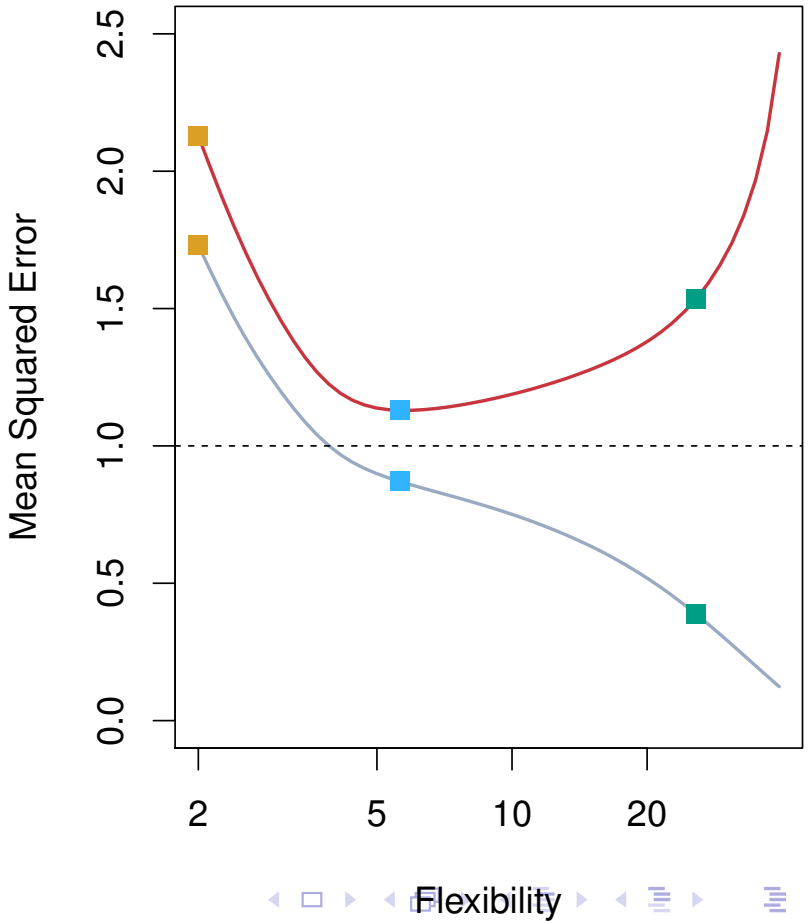
For example, consider a data set: $(x_1, y_1), \cdots, (x_{100}, y_{100})$.

If we use a polynomial of degree 99 (so that there are 100 parameters), we can perfectly fit the data set by a curve. However, it makes poor sense.

2.2.1 Measuring the Quality of Fit

Figure 2-9: (left) Training data (circles) were generated by black curve plus noise. As a function becomes more complex (green), it fits better for training data. However, it does not mean the green curve is better than the simpler blue curve.





2.2.1 Measuring the Quality of Fit

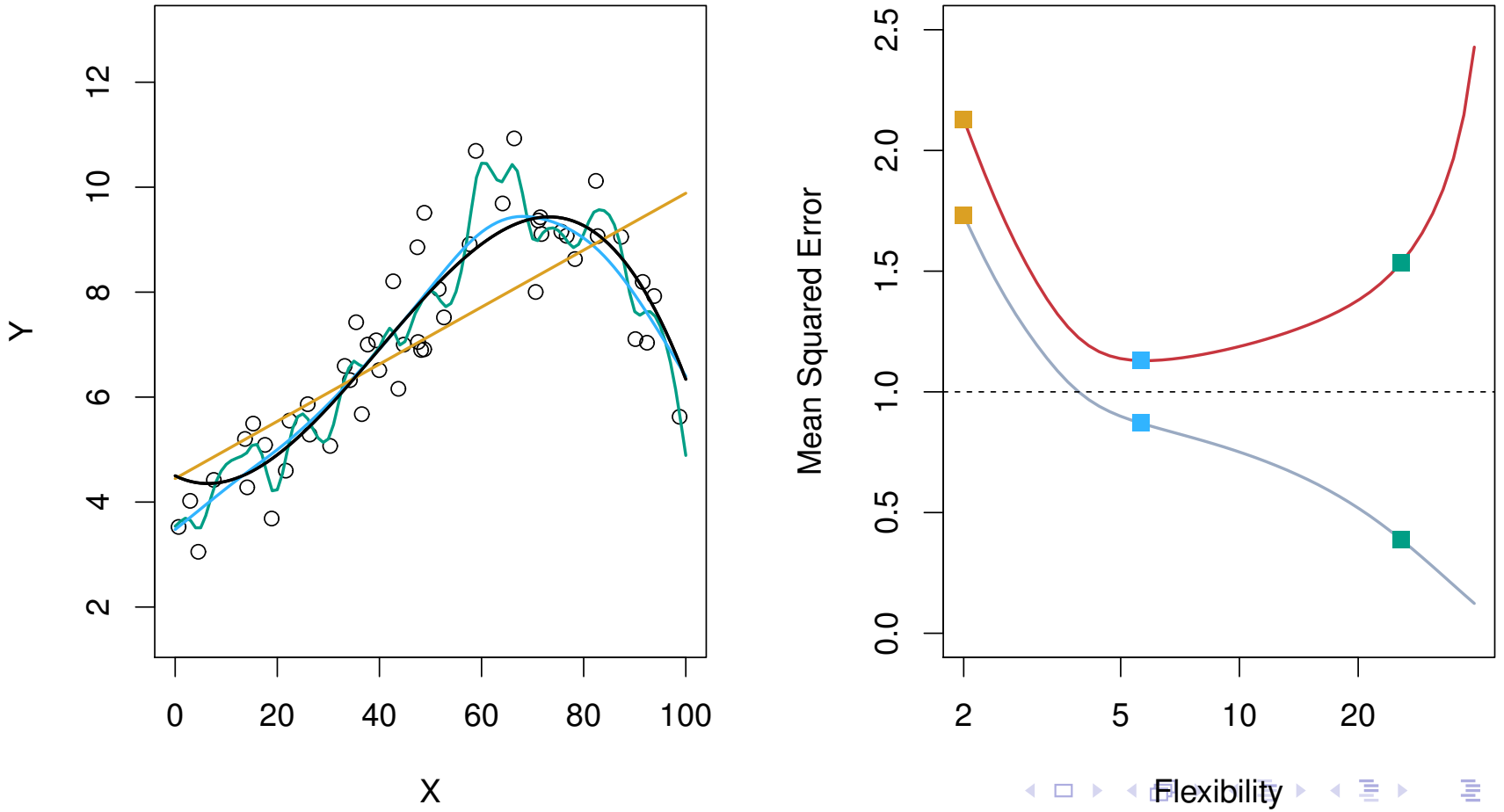
Test MSE

Consider we have (or divide a data set into) a training data set $\boldsymbol{D_1}$ and a test data set $\boldsymbol{D_2}$. When we estimate the function $\boldsymbol{f}$ with $\boldsymbol{D_1}$ and calculate MSE by $\boldsymbol{D_2}$, the MSE is called a test MSE.

The test MSE is a more accurate measure of the expected error.

2.2.1 Measuring the Quality of Fit

Figure 2-9: (right) The gray curve shows training MSE (vertical) given the flexibility (i.e., the number of parameters, horizontal). The red curve shows test MSE. The green curve is better in training MSE, but the blue curve is optimal in test MSE.



2.2.2 Bias-Variance trade-off

Example: Bias and Variance

Predict the height ($\boldsymbol{Y}$) of a male by foot size ($\boldsymbol{X}$).

Training data (foot size in cm, height in cm):

(27.2, 176), (27.5, 180), (28.0, 176), (28.7, 184), (29.0, 183).

A new observation: the foot size is 28.6cm with height unknown.

Compare two choices for prediction:

(a) Use the data point $\mathbf{(28.7, 184)}$ only. $\hat{\boldsymbol{Y}} = \mathbf{184}$.
- $\hat{\boldsymbol{Y}}$ has a smaller bias since only the closest observation is used.

(b) Use the mean height of 5 data points. $\hat{\boldsymbol{Y}} = \mathbf{179.8}$.
- $\hat{\boldsymbol{Y}}$ has a smaller variance since more observations are used.

2.2.2 Bias-Variance trade-off

A mathematical result

Let $y = f(x) + \epsilon$ be the true model, and $\hat{f}$ is the estimated model from data. Denote $f_0 := f(x_0)$, $\hat{f}_0 := \hat{f}(x_0)$.

The expected error size is decomposed into three components:

$$
\begin{aligned}
E\left(y_0 - \hat{f}_0\right)^2 &= E\left(f_0 + \epsilon - \hat{f}_0\right)^2 && (7)\\
&= E\left(f_0 - E\hat{f}_0 + E\hat{f}_0 - \hat{f}_0 + \epsilon\right)^2 && (8)\\
&= (E\hat{f}_0 - f_0)^2 + E(\hat{f}_0 - E\hat{f}_0)^2 + E\epsilon^2 && (9)\\
&= (\textit{Bias of } \hat{f}_0)^2 + \mathrm{Var}\hat{f}_0 + E\epsilon^2 && (10)
\end{aligned}
$$

where the last term is an irreducible error.

Note that these three components are conceptual, and not observed in real data analysis as f_0 and ϵ are unknown. The decomposition is possible for simulated examples only.

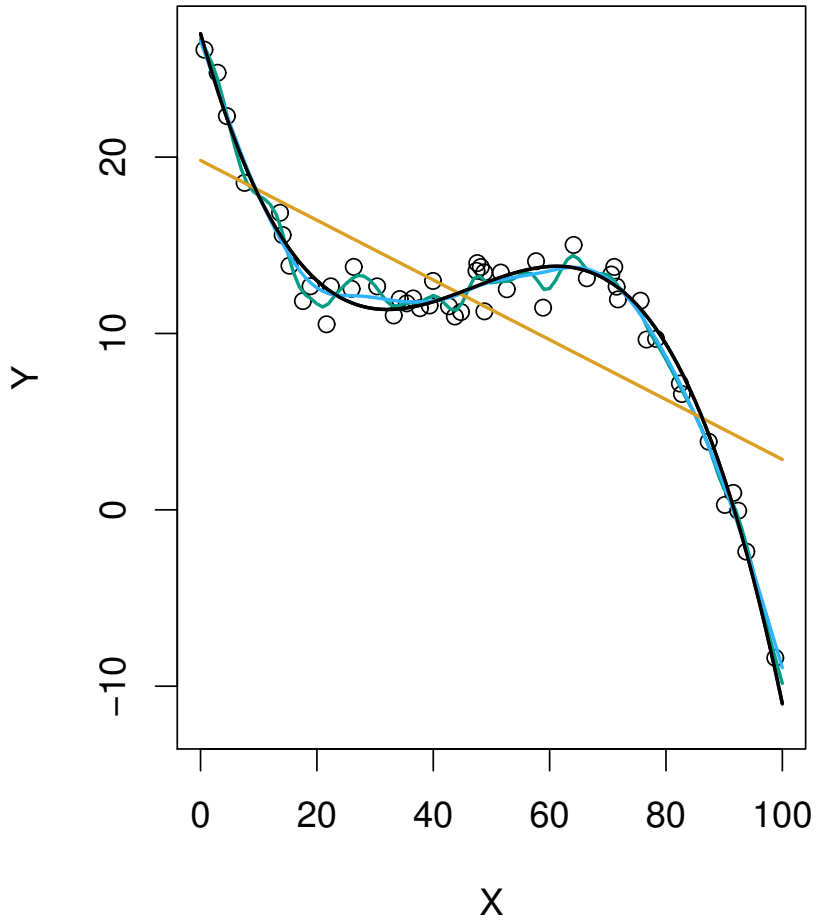
2.2.2 Bias-Variance trade-off

Implications of the mathematical result:

- To minimize expected squared error, the model has to balance between bias and variance.
- There is a trade off between bias and variance.

2.2.2 Bias-Variance trade-off

Fig 2-11: The fitted line (yellow) does not change much when the data set changes (small variance), but for example $\boldsymbol{y}$ is underestimated at $\boldsymbol{x = 70}$ due to too much simplification (large bias).



2.2.3 Classification setting

A measure of error for classification problems

The error rate is a natural measure of training error for classification problems.

Suppose that observations are $(x_1, y_1), \cdots, (x_n, y_n)$ are observations. $y_i \in \{1, 2, 3, \cdots, G\}$ if there are G classes.

We estimate y_i by x_i. Let $\hat{y}_i$ be the estimated class. Then,

$$\text{error rate} := \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}_i) \tag{11}$$

where $I(y_i \neq \hat{y}_i) = 1$ if $y_i \neq \hat{y}_i$, and $I(y_i \neq \hat{y}_i) = 0$ if $y_i = \hat{y}_i$.

2.2.3 Classification setting

Test error

The expected error size $\boldsymbol{E[I(y_0 \neq \hat{y}_0)]}$ is estimated by the test error rate, which is also defined by (11) but for the test data set.

A good classifier is one for which the ~~text~~ error is smallest.

test

2.2.3 Classification setting

Bayes Classifier

Suppose that the conditional probability of Y given x_0, i.e., $P(Y = j|X = x_0)$, is calculated for $j = 1, \cdots, G$ and all x_0.

Then, the best classifier is to assign each observation to the most likely class. This is called the Bayes classifier.

The error rate for this classifier is called the Bayesian error rate. Given X,

$$\text{Bayesian error rate} = 1 - E\left(\max_j P(Y = j|X)\right). \tag{12}$$

2.2.3 Classification setting

Example 1:

Suppose that $\boldsymbol{X}$ is height of a person, and $\boldsymbol{Y}$ is either 'child', 'adult female' or 'adult male', and that we have following conditional probabilities.

Height (cm)	Child	Adult Female	Adult Male
$\boldsymbol{X < 145}$	0.7	0.2	0.1
$\boldsymbol{145 \leq X < 165}$	0.2	0.6	0.2
$\boldsymbol{165 \leq X}$	0.1	0.3	0.6

Then, the red class is assigned for each observation. For example, if a person has height between 145cm and 165cm, 'adult female' is assigned since it is the most likely class.

2.2.3 Classification setting

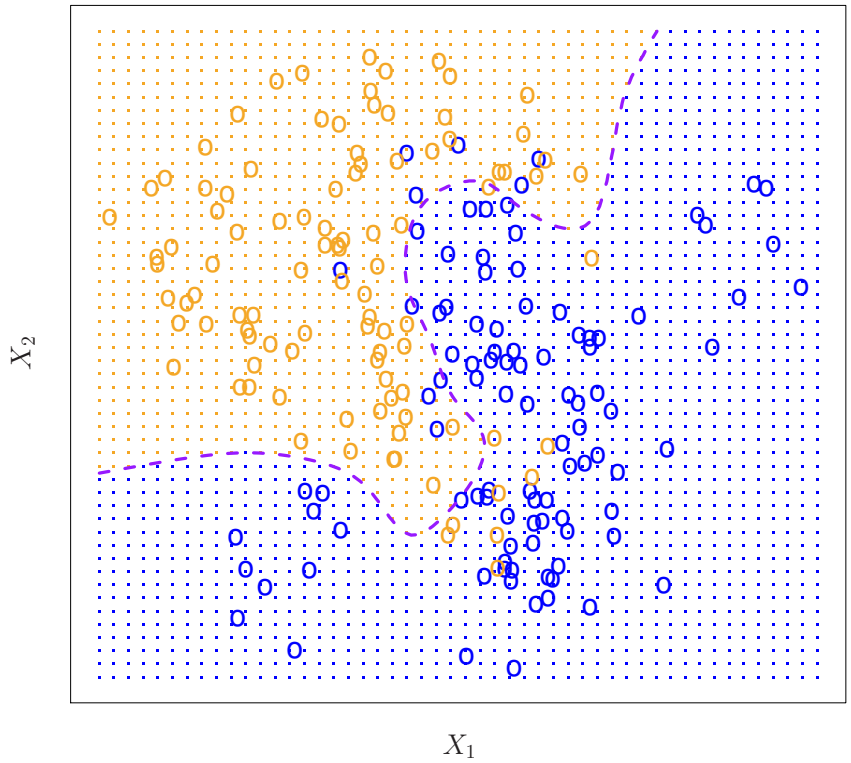
Example 2: Binary case

Consider a binary classification problem. Y is either 0 or 1. Then, assign $Y = 1$ if $P(Y = 1|X = x_0) > 0.5$ and $Y = 0$ if $P(Y = 1|X = x_0) \leq 0.5$.

The conditional probability is given by a statistical model (cf. logistic regression).

2.2.3 Classification setting

Fig 2-13: An example of Bayesian classifier. $\boldsymbol{X = (X_1, X_2)}$ is 2-dimensional. There are 2 classes: orange and blue. Circles are observations. Dots show classification for each possible value of $\boldsymbol{X}$. The black dashed curve is the Bayesian decision boundary, i.e., the set of $\boldsymbol{x}$ satisfying $\boldsymbol{P(Y = 1|X = x) = 0.5}$.



2.2.3 Classification setting

K-nearest neighbors

Let K be a natural number $(1, 2, 3, \cdots)$. The K-nearest neighbors (KNN) is a non-parametric approach to estimate conditional probabilities by

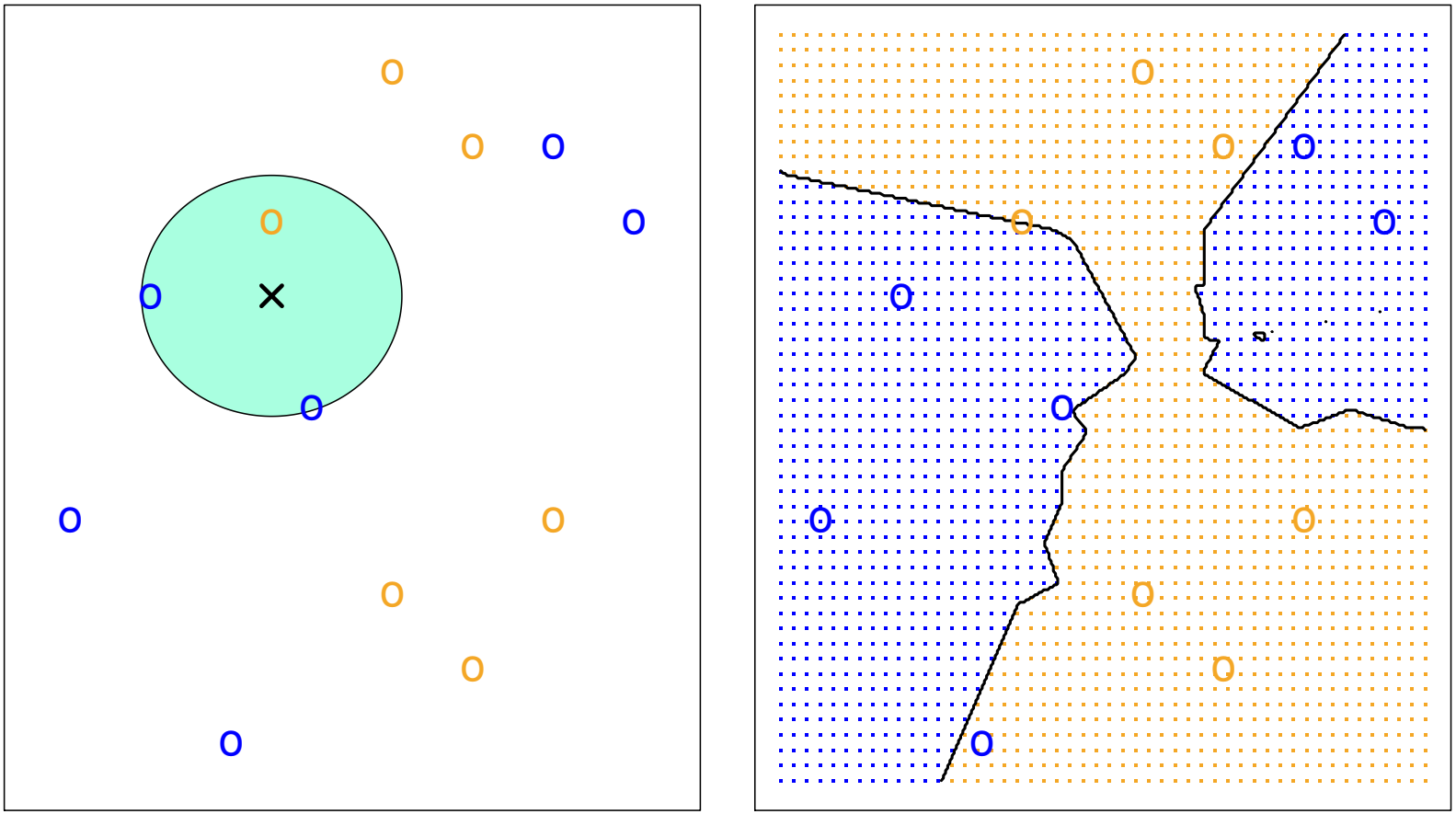
$$P(Y = j|X = x_0) = \frac{1}{K} \sum_{i \in N_0} I(y_i = j). \tag{13}$$

where N_0 is the set of K points closest to x_0 in the training set.

In other words, KNN estimates the probability by empirical results of K most similar cases.

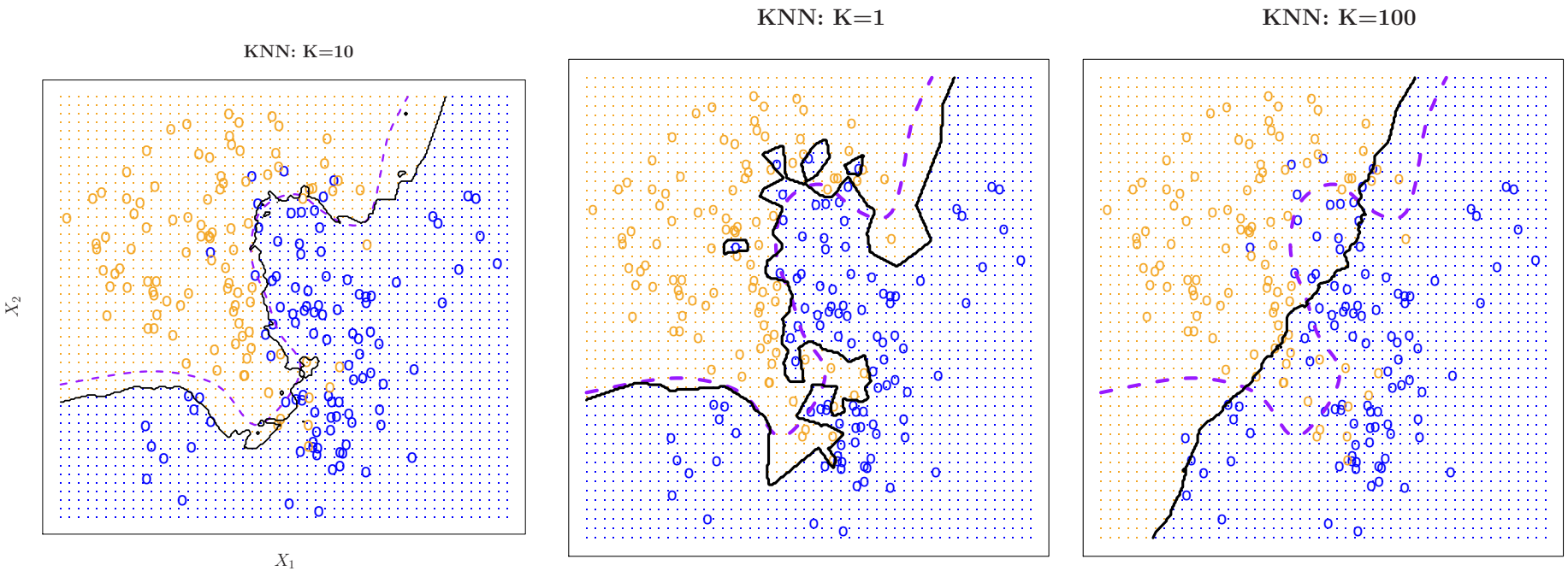
2.2.3 Classification setting

Fig 2-14: An example of KNN ($K = 3$). Left: At the cross, estimates are $P(Y = blue) = 2/3$ and $P(Y = orange) = 1/3$. Right: Bayesian classifier by this KNN.



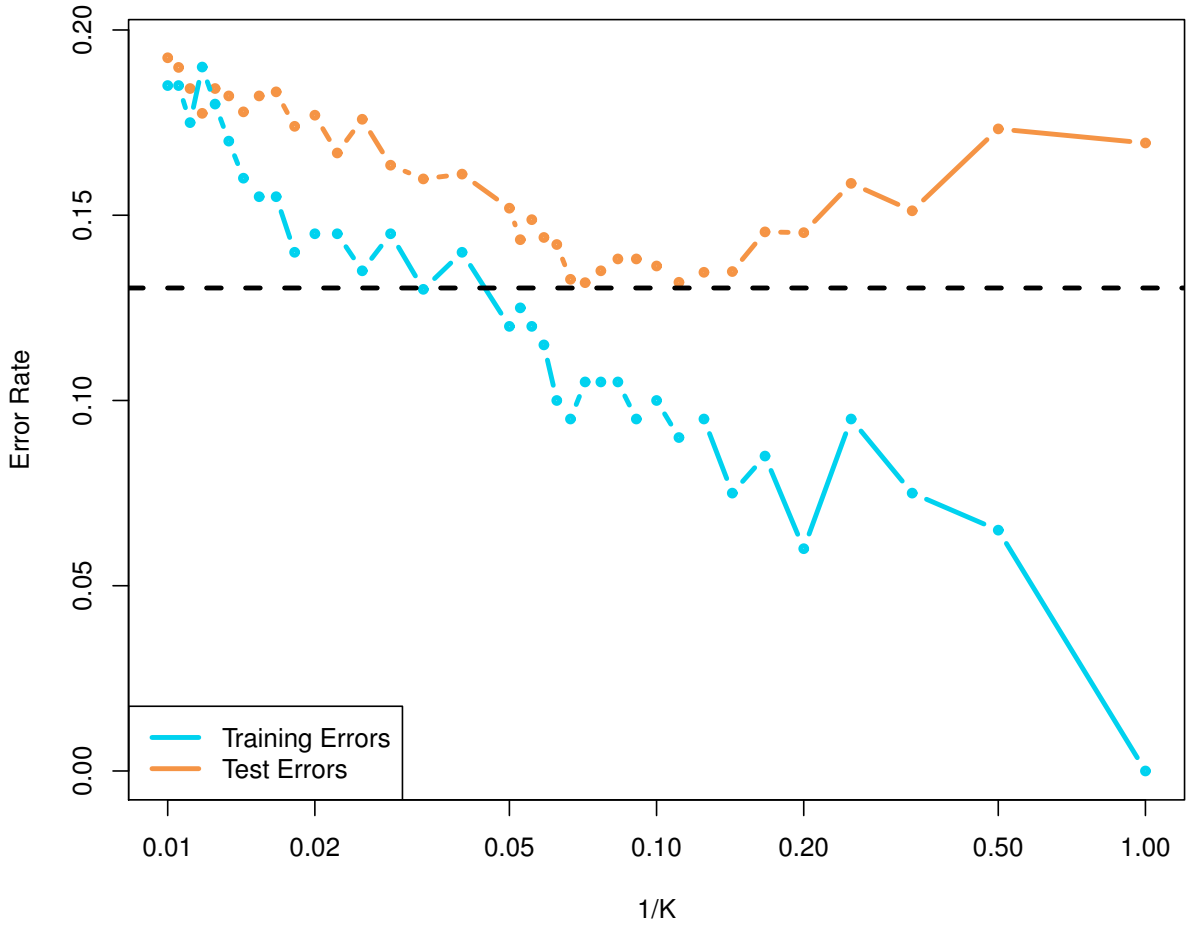
2.2.3 Classification setting

Fig 2-15&16: The result of KNN depends on K ($\mathbf{10, 1, 100}$). A small K is more detailed; a large K is more robust. Solid lines: KNN. Purple dashed lines: Bayesian boundary by a benchmark method (no details in book).



2.2.3 Classification setting

Fig 2-17: The error rate by ***K*** in KNN. There is an optimal ***K*** to minimize the (expected) test error.



2.2.3 Classification setting

Pros and cons of a small K and a large K is similar to bias-variance trade-off in regression problems.

- A small K: small bias, large variance.
- A large K: large bias, small variance.

Memo