STATISTICAL LEARNING - CONCEPTS

Part II – Assessing Model Accuracy

Outline

- Measuring the quality of fit
- The bias-variance trade-off
- The classification setting

- No one method dominates all others over *all* possible data sets
- Decide for a specific data set which method produces the best results

Measuring the quality of fit

- The general model

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

- Training data

$$\{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$$

- Apply a statistical learning method to the training data to obtain the estimate $\hat{f}$

- The prediction rule

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

- Prediction error

$$\left(Y - \hat{Y}\right)^2$$

Mean squared error (MSE)

- Training MSE

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

Test MSE

- Test data $\{(x_0, y_0)\}$
  - Previously unseen observations not used to train the statistical learning method
- Test MSE

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

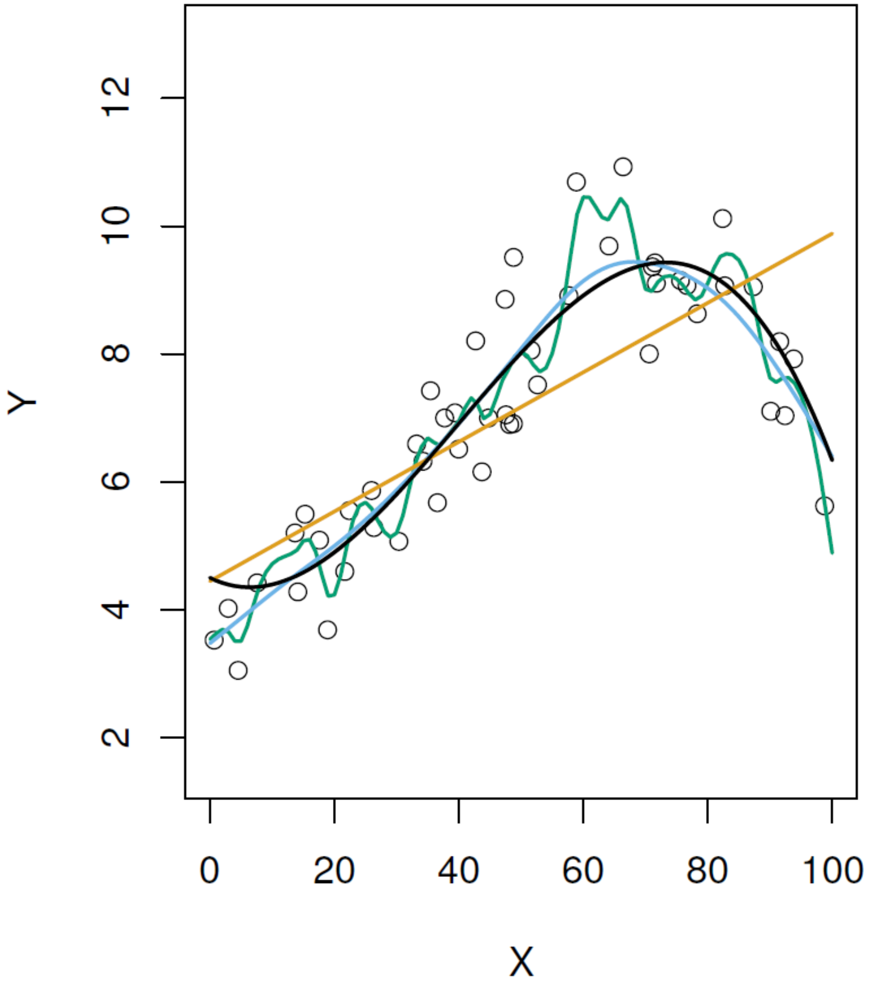
- We care about how well the method works on the test data
- How can we go about trying to select a method that minimizes the test MSE?

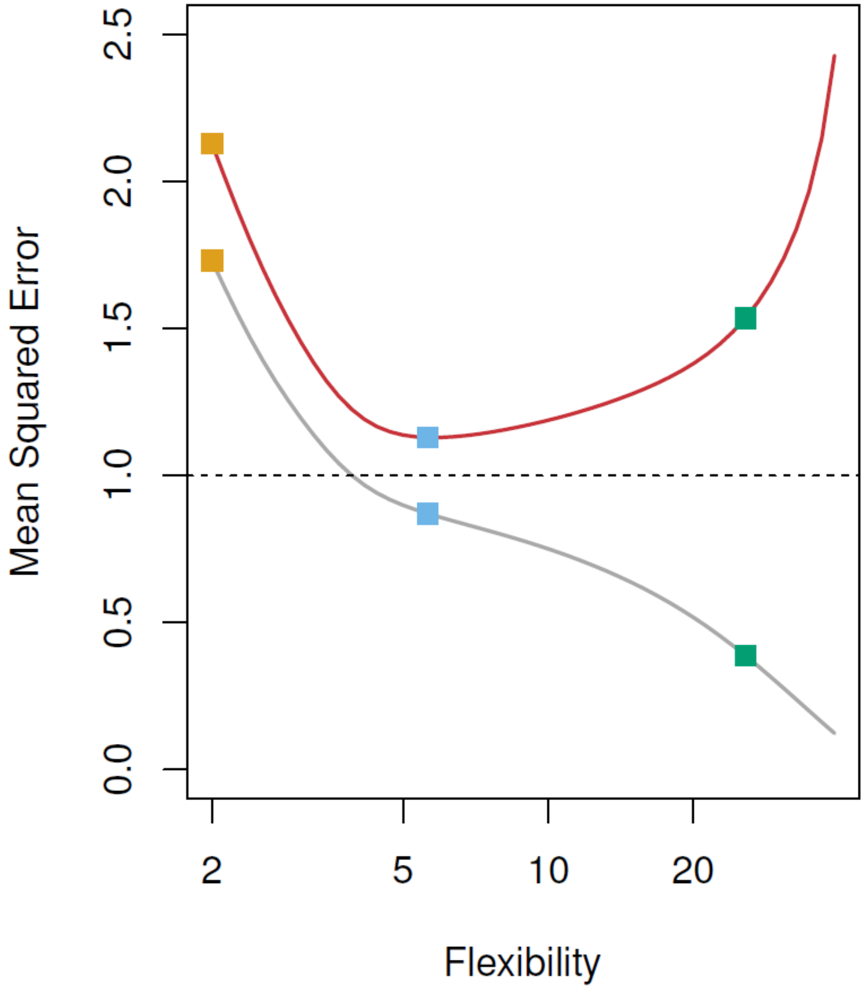
Simulated examples

- Example 1: f is non-linear
- Example 2: f is approximately linear
- Example 3: f is highly non-linear

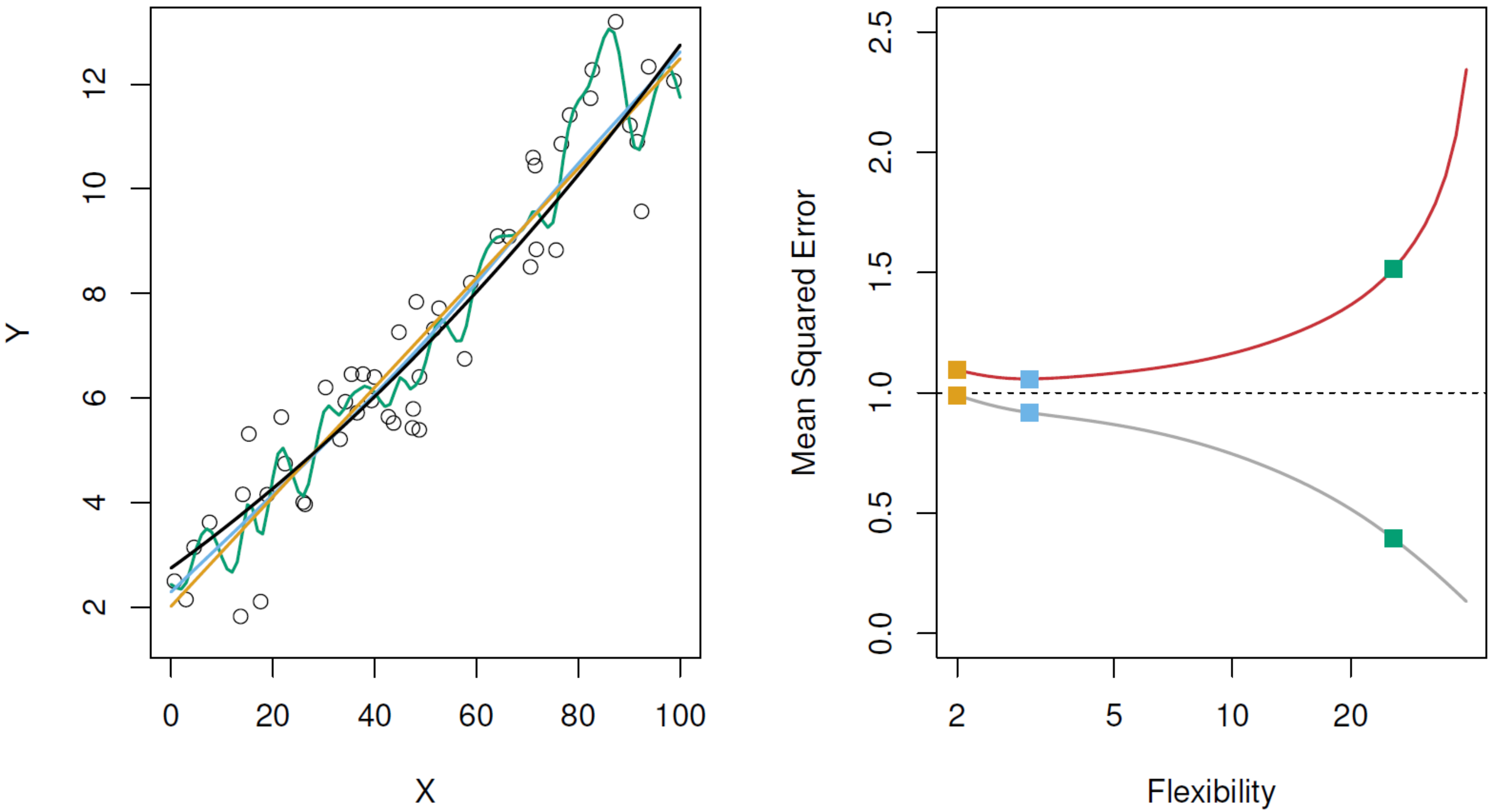
- Three methods for estimating f with increasing levels of flexibility
  - Linear regression
  - Smoothing spline
  - Smoothing spline (more flexible)
- Compute the test MSE over a very large test set

Example 1: f is non-linear

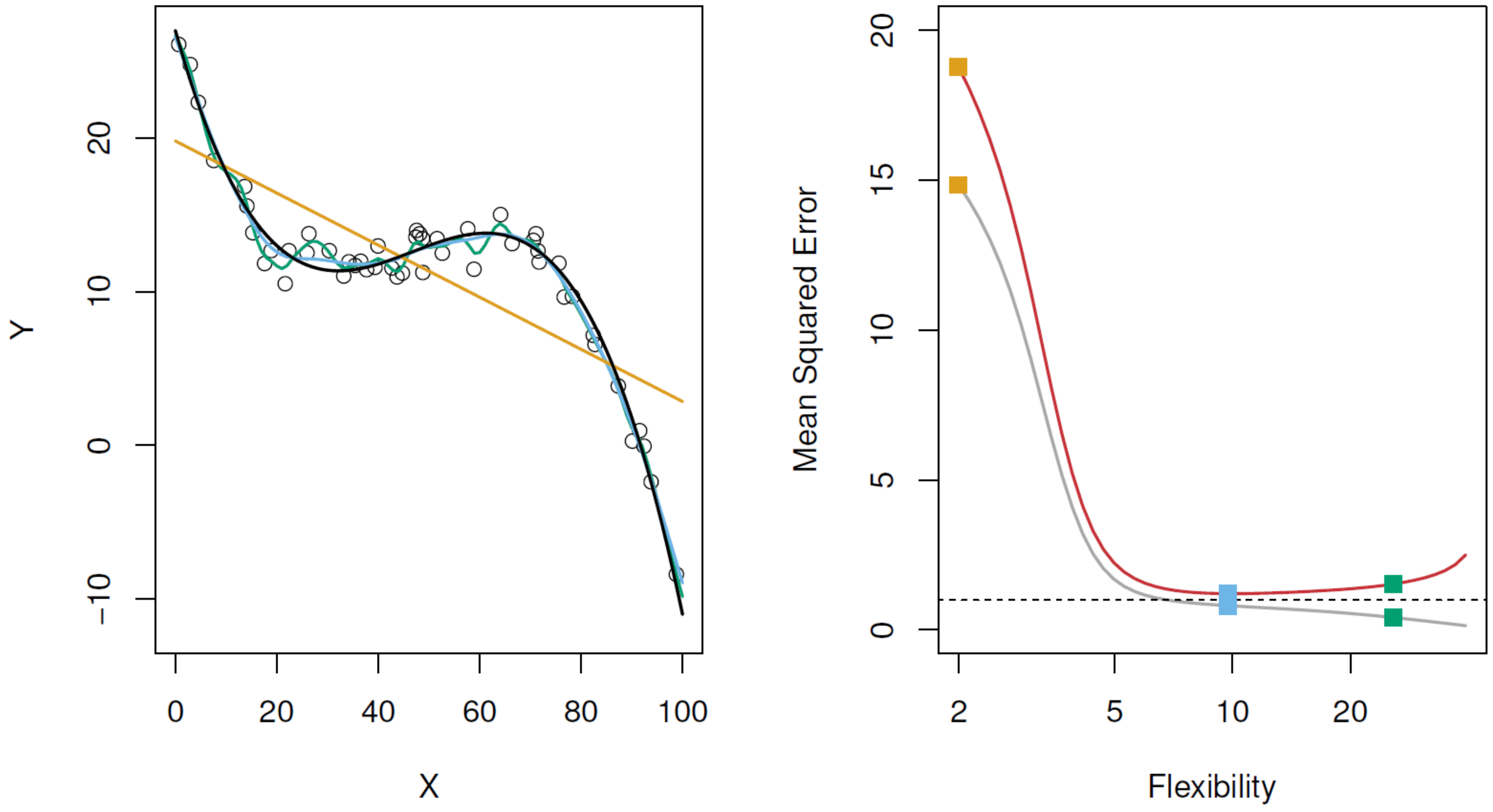




Example 2: f is approximately linear



Example 3: f is highly non-linear



Observations

- A monotone decrease in the training MSE
- A *U-shape* in the test MSE
- The flexibility level corresponding to the minimal test MSE can vary considerably

- There is no guarantee that the method with the lowest training MSE will also have the lowest test MSE
  - How can we select a method that minimizes the test MSE?
  - How can we compute the test MSE when no test data are available?

The bias-variance trade-off

Expected test MSE

- Test MSE

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

- *Expected* test MSE

$$E\{Y - \hat{f}(X)\}^2$$

The bias-variance decomposition

- Show that

$$E\left[\{Y-\hat{f}(X)\}^2|X\right] = Var_{Train}\{\hat{f}(X)\} + \left[E_{Train}\{\hat{f}(X)\} - f(X)\right]^2 + Var(\epsilon)$$

$$= Variance(X) + Bias^2(X) + Irreducible\ Error$$

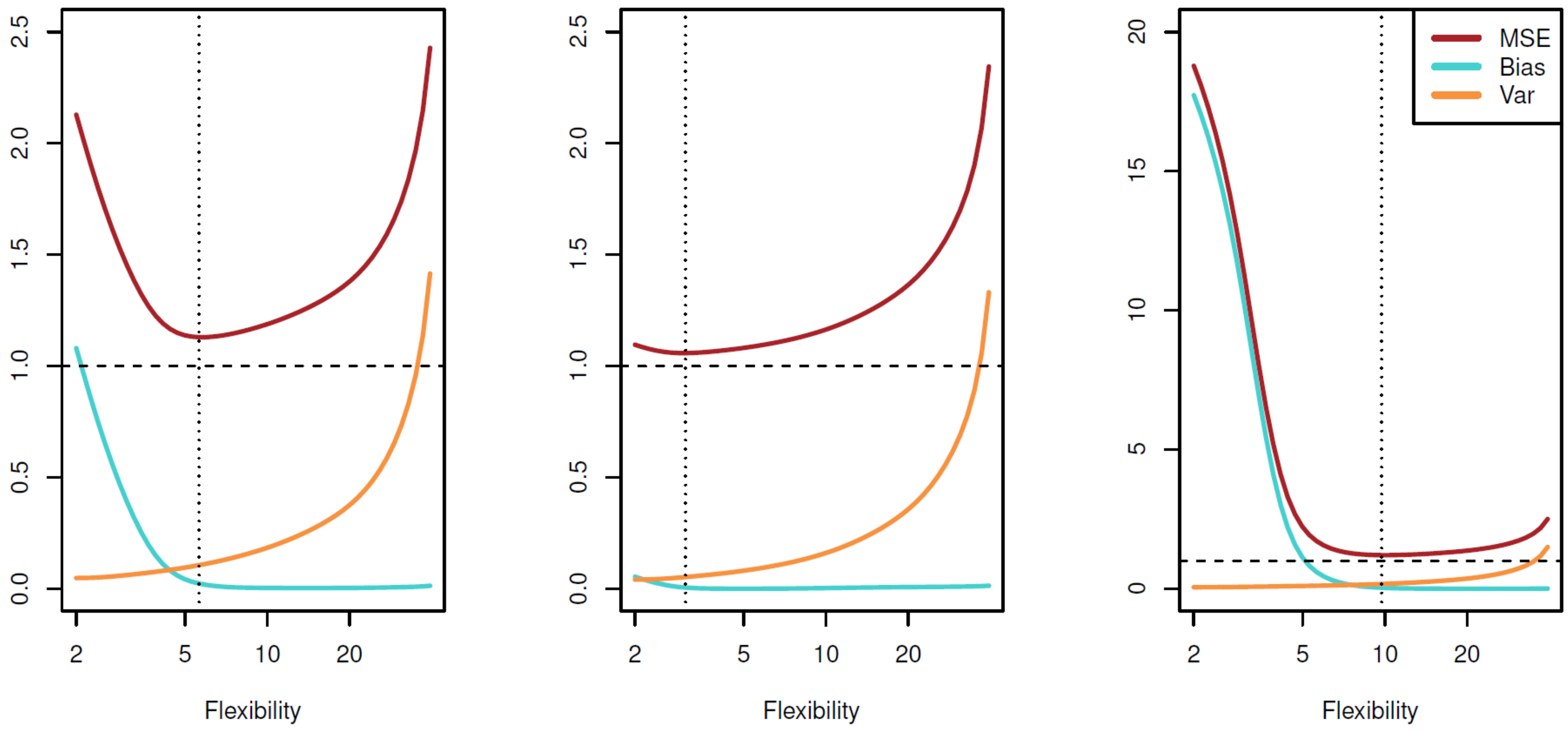
The bias-variance decomposition

- Show that

$$E\left[\{Y-\hat{f}(X)\}^2|X\right] = Var_{Train}\{\hat{f}(X)\} + [E_{Train}\{\hat{f}(X)\} - f(X)]^2 + Var(\epsilon)$$
$$= Variance(X) + Bias^2(X) + Irreducible\ Error$$

- Hints:

$$E\left[\{Y-\hat{f}(X)\}^2|X\right] = E_{Train}\left(E_Y\left[\{Y-\hat{f}(X)\}^2|X\right]\right)$$
$$E_Y\left[\{Y-\hat{f}(X)\}^2|X\right] = Var(\epsilon) + \{\hat{f}(X) - f(X)\}^2$$
$$E_{Train}\{\hat{f}(X) - f(X)\}^2 = Var_{Train}\{\hat{f}(X)\} + [E_{Train}\{\hat{f}(X)\} - f(X)]^2$$



Expected test MSE, bias, and variance for the three data sets in Examples 1-3

The trade-off

- As the flexibility increases, the variance will increase, and the bias will decrease
- The flexibility level corresponding to the minimal test MSE can vary considerably

The classification setting

Error rates

- Training error rate

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

- Test error rate

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

The Bayes classifier

- Also known as the Bayes rule
- Assign an observation to the most likely class, given the its predictor values

$$\Pr(Y = j | X = x_0)$$

  - Class conditional probabilities

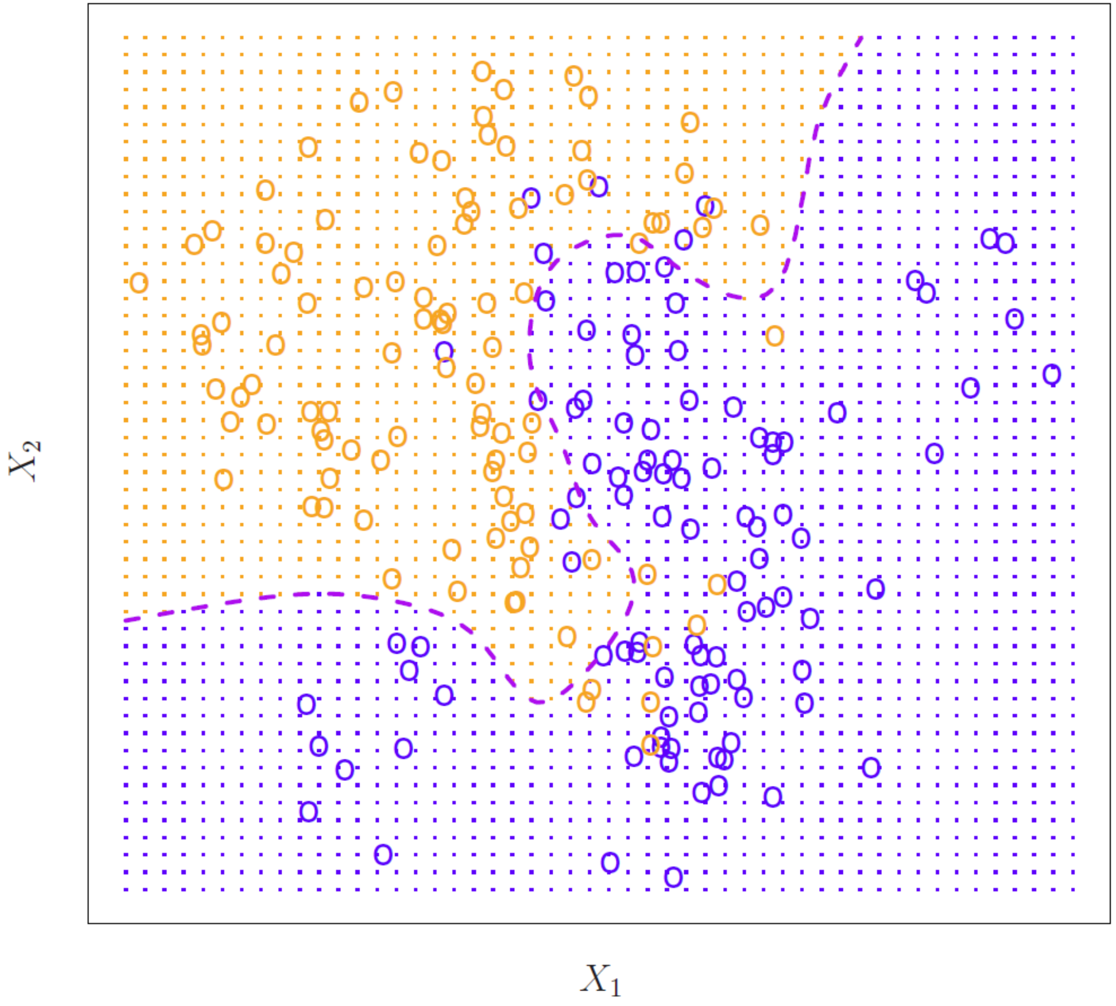
The Bayes error rate

- The lowest test error rate (exercise)

$$1 - \max_j \Pr(Y = j | X = x_0)$$

- Expected test error rate (irreducible error)

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



A simulated data set consisting of 100 observations in each of two groups. The Bayes error rate is 0.1304

K-nearest neighbors (KNN) classifier

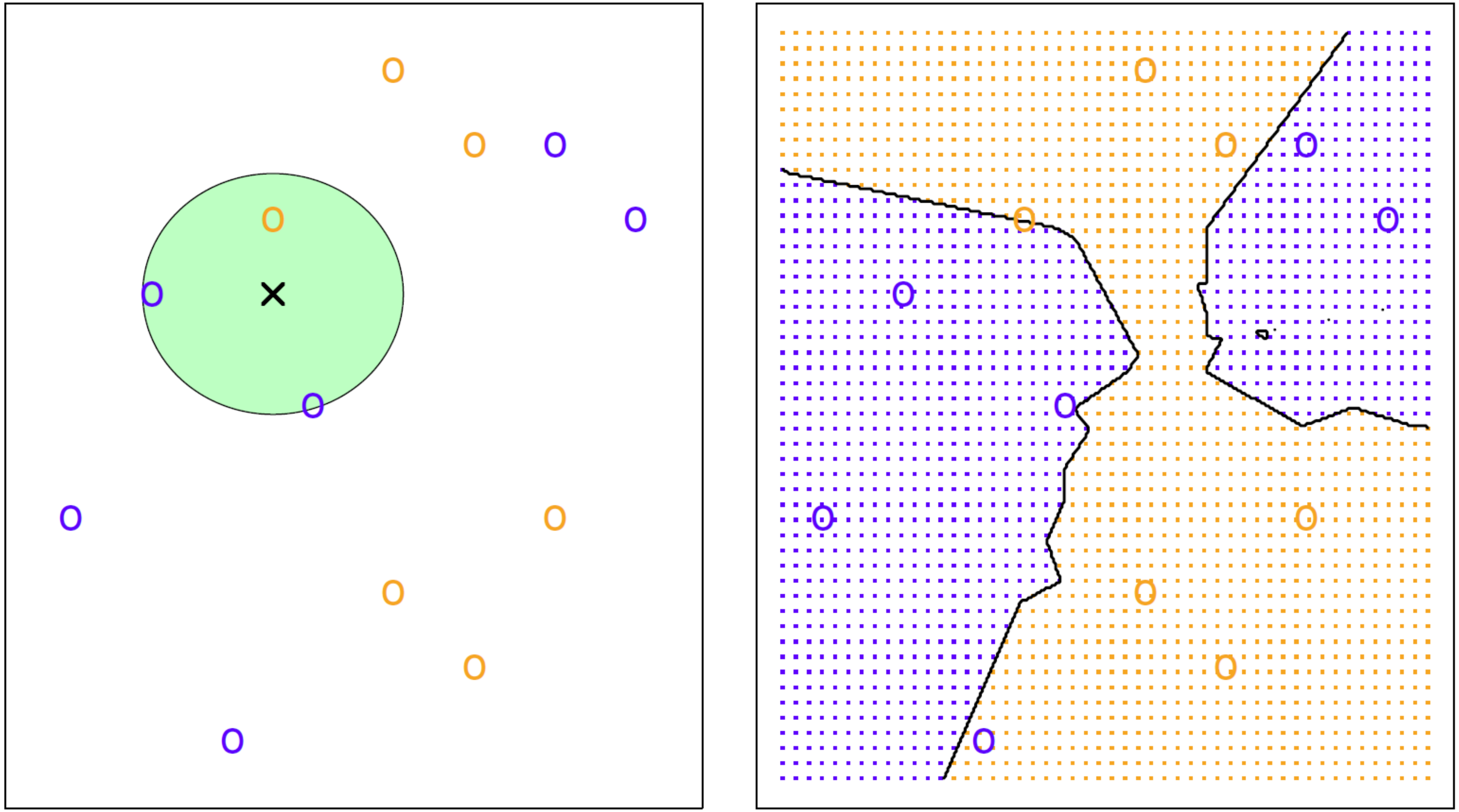
① Specify a positive integer K

② Identify the K points in the training data that are closest to x_0, represented by $\mathcal{N}_0$

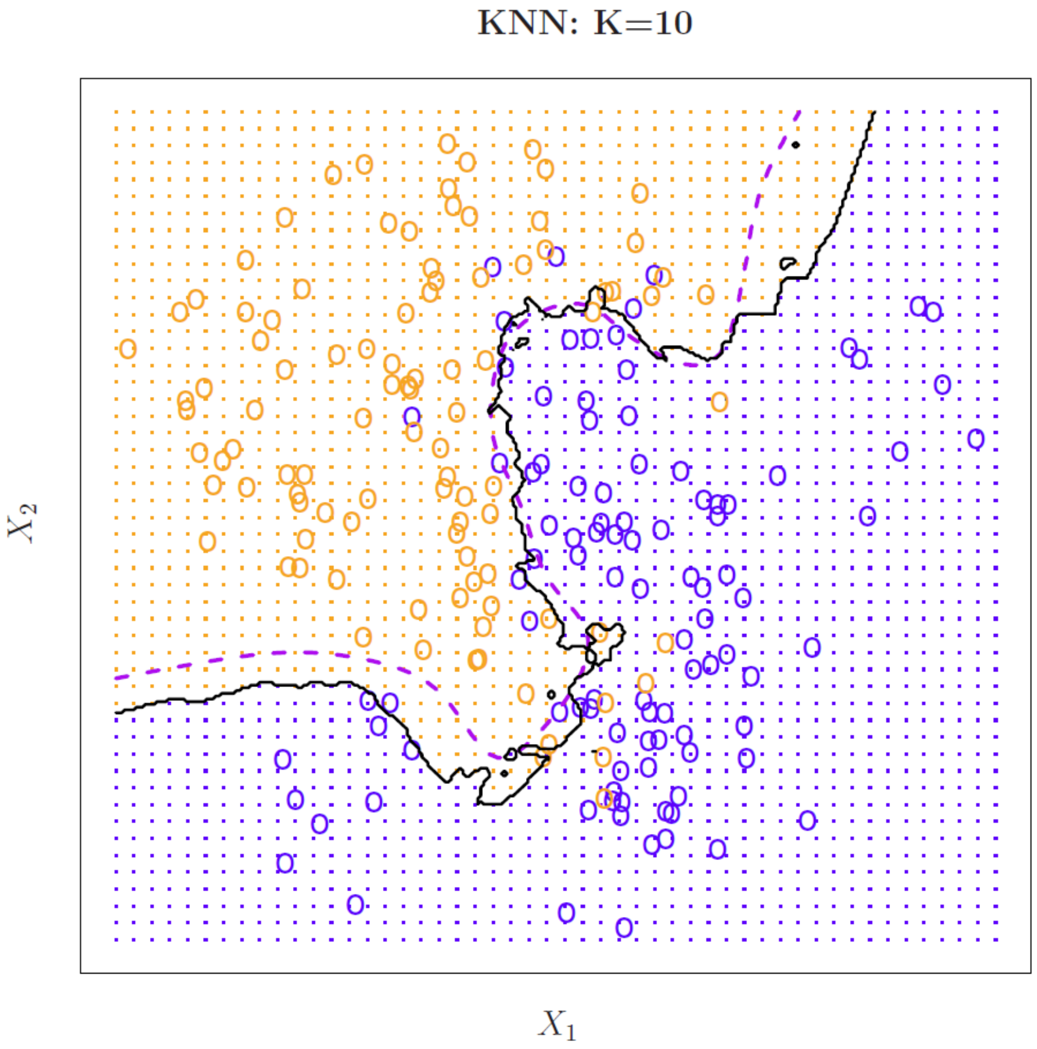
③ Estimate the conditional probabilities

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

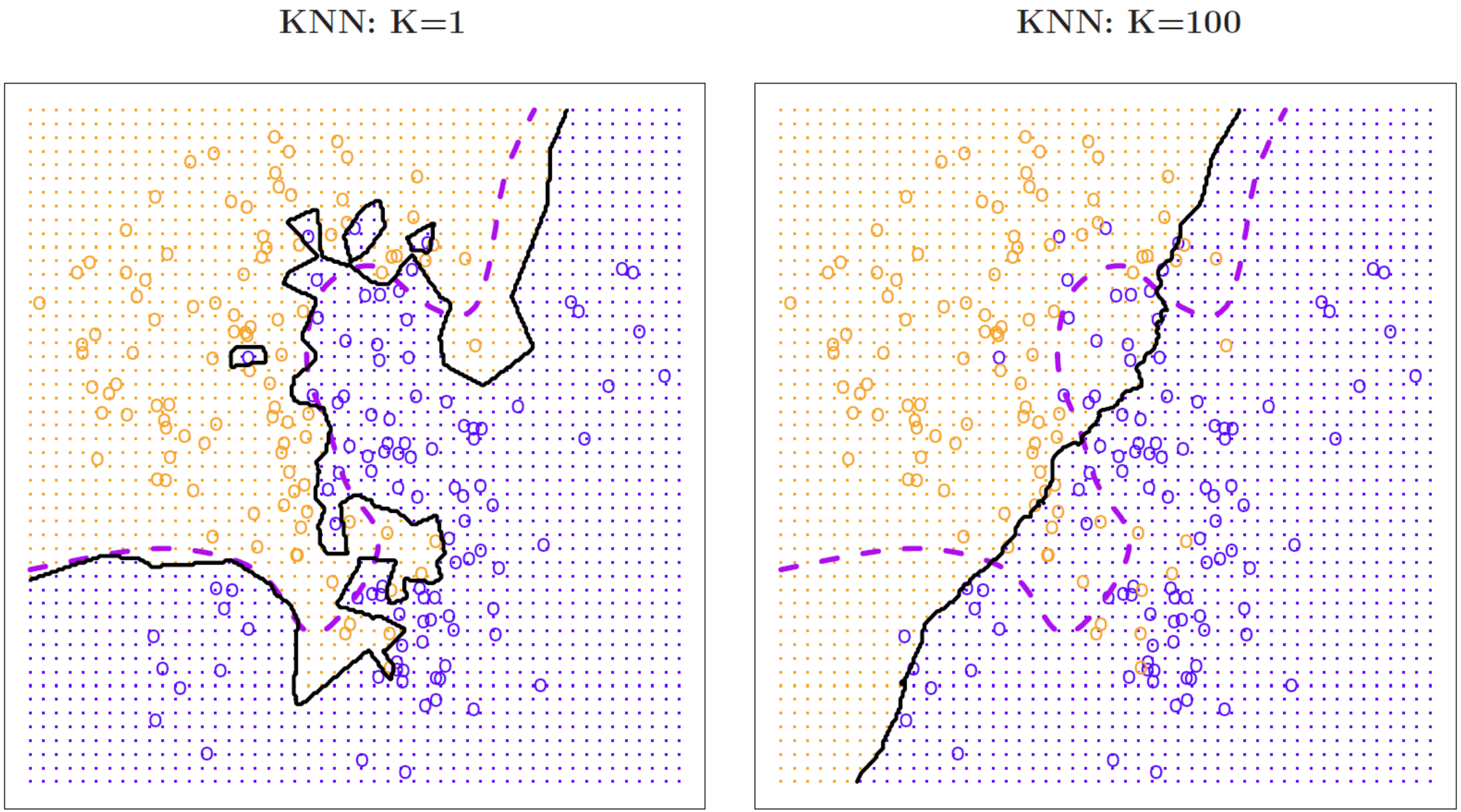
④ Apply the Bayes rule

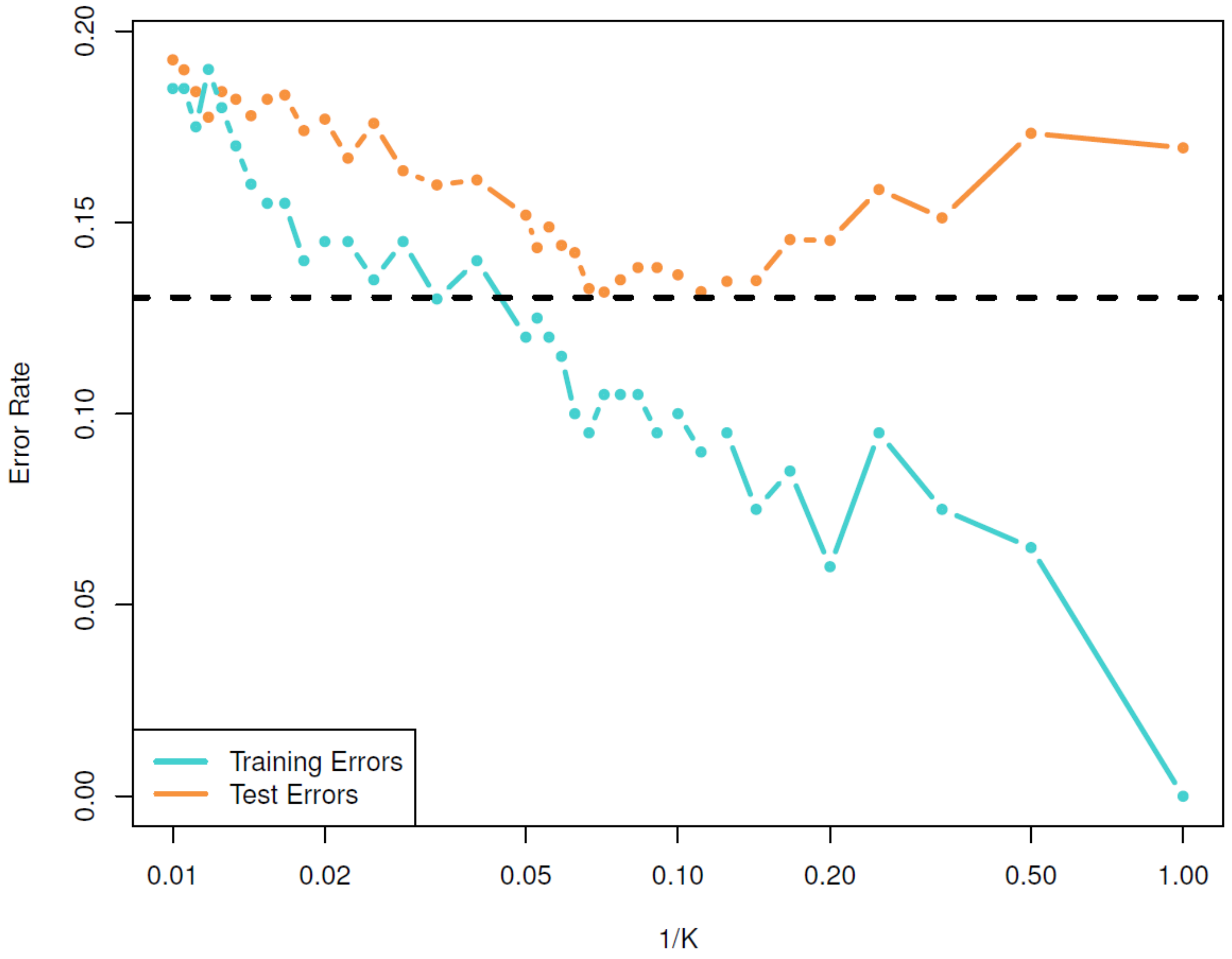


An illustrative example



The KNN and Bayes decision boundaries. The test error rate using KNN is 0.1363


KNN: K=1
KNN: K=100



The KNN training error rate (blue, 200 observations) and test error rate (orange, 5000 observations)