Introduction to Data Science

Wenjing Liao

School of Mathematics, Georgia Institute of Technology

Math 4803
Fall 2019

Textbook: James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani, An introduction to statistical learning. Vol. 112, New York: Springer, 2013

Data sets to explore:

- Data sets in textbook
- MNIST digit data
- Characters
- Textures
- Flowers
- Face data and more face data

More advanced data:

- ImageNet
- UCI Machine Learning Repository
- Visual Geometry Group

Outline

Task I: Numerical prediction

Wage data

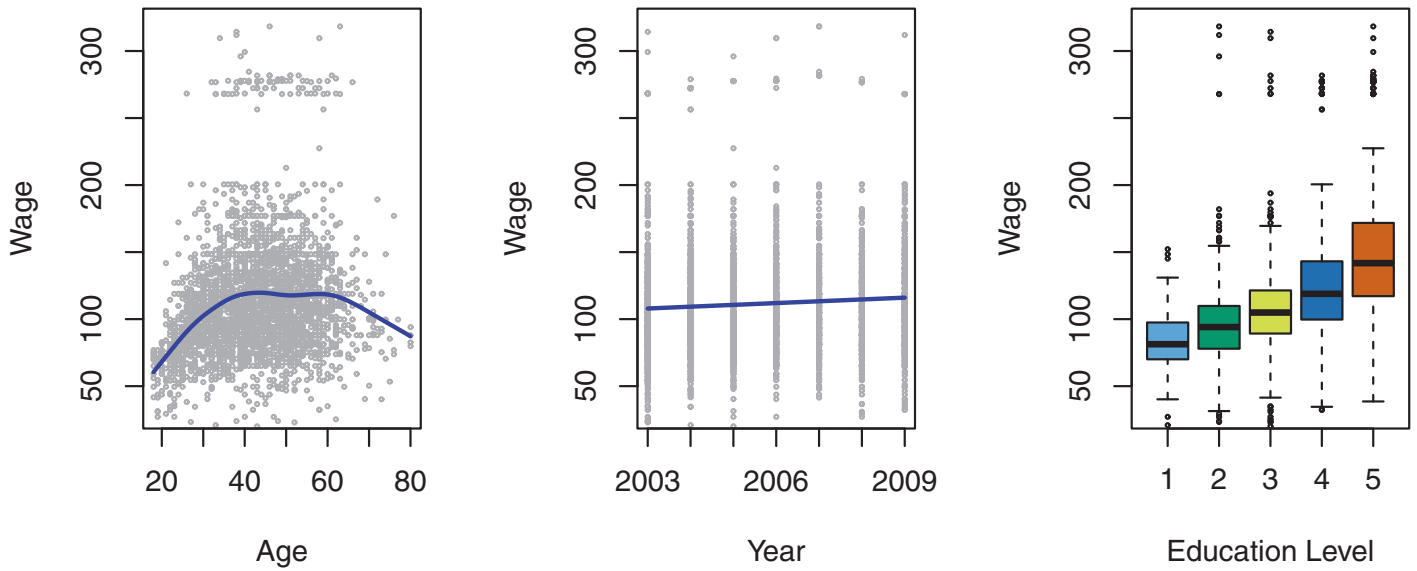


FIGURE 1.1. `Wage` *data, which contains income survey information for males from the central Atlantic region of the United States.* Left: `wage` *as a function of* `age`*. On average,* `wage` *increases with* `age` *until about* 60 *years of age, at which point it begins to decline.* Center: `wage` *as a function of* `year`*. There is a slow but steady increase of approximately* $10,000 *in the average* `wage` *between* 2003 *and* 2009*.* Right: *Boxplots displaying* `wage` *as a function of* `education`*, with* 1 *indicating the lowest level (no high school diploma) and* 5 *the highest level (an advanced graduate degree). On average,* `wage` *increases with the level of education.*

Numerical prediction

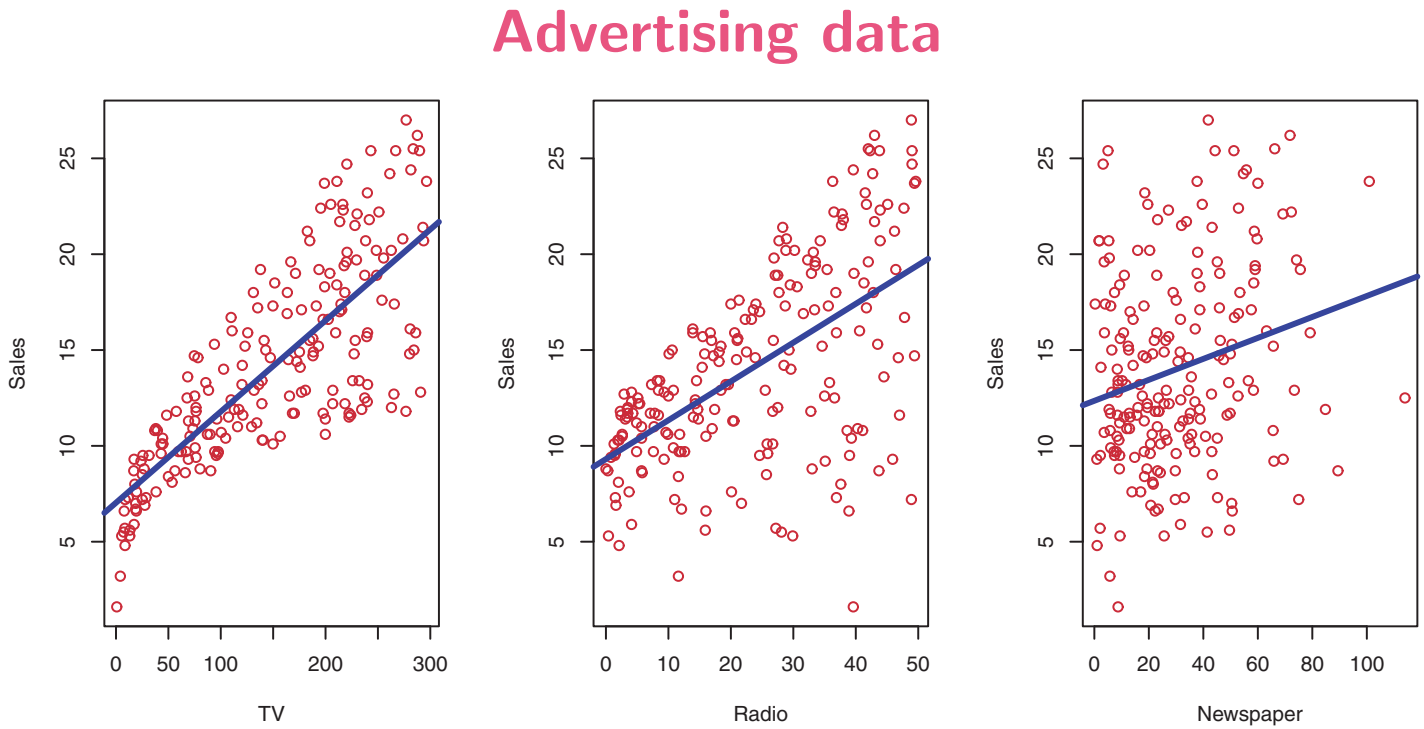


FIGURE 2.1. *The* `Advertising` *data set. The plot displays* `sales`*, in thousands of units, as a function of* `TV`*,* `radio`*, and* `newspaper` *budgets, in thousands of dollars, for* 200 *different markets. In each plot we show the simple least squares fit of* `sales` *to that variable, as described 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.*

Task II: Categorical prediction

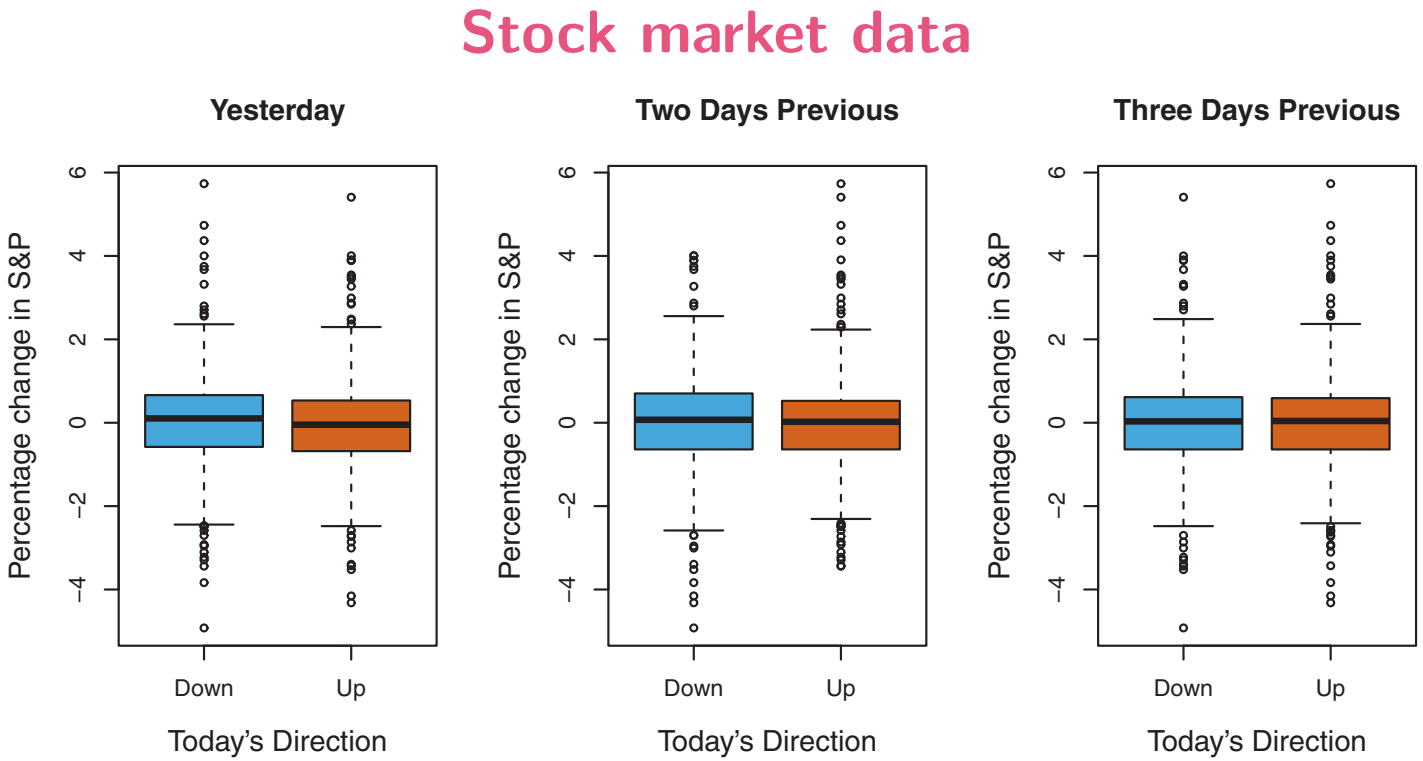


FIGURE 1.2. Left: *Boxplots of the previous day's percentage change in the S&P index for the days for which the market increased or decreased, obtained from the* `Smarket` *data.* Center and Right: *Same as left panel, but the percentage changes for 2 and 3 days previous are shown.*

Categorical prediction

- Gender: Male or Female?
- Email spam: Yes or No?
- Disease: Yes or No?
- Digit: $0, 1, 2, \ldots, 9$?
- Images: Cat or Dog?

Task III: Clustering

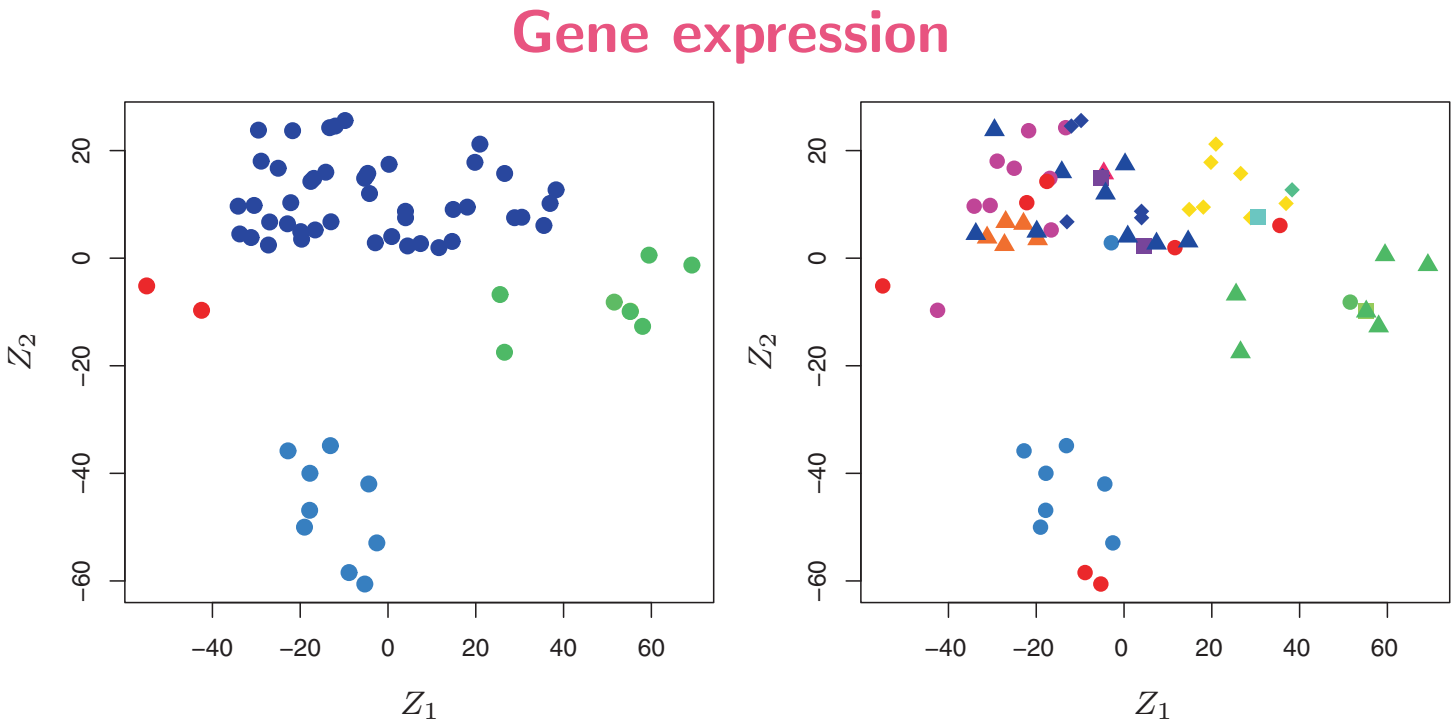
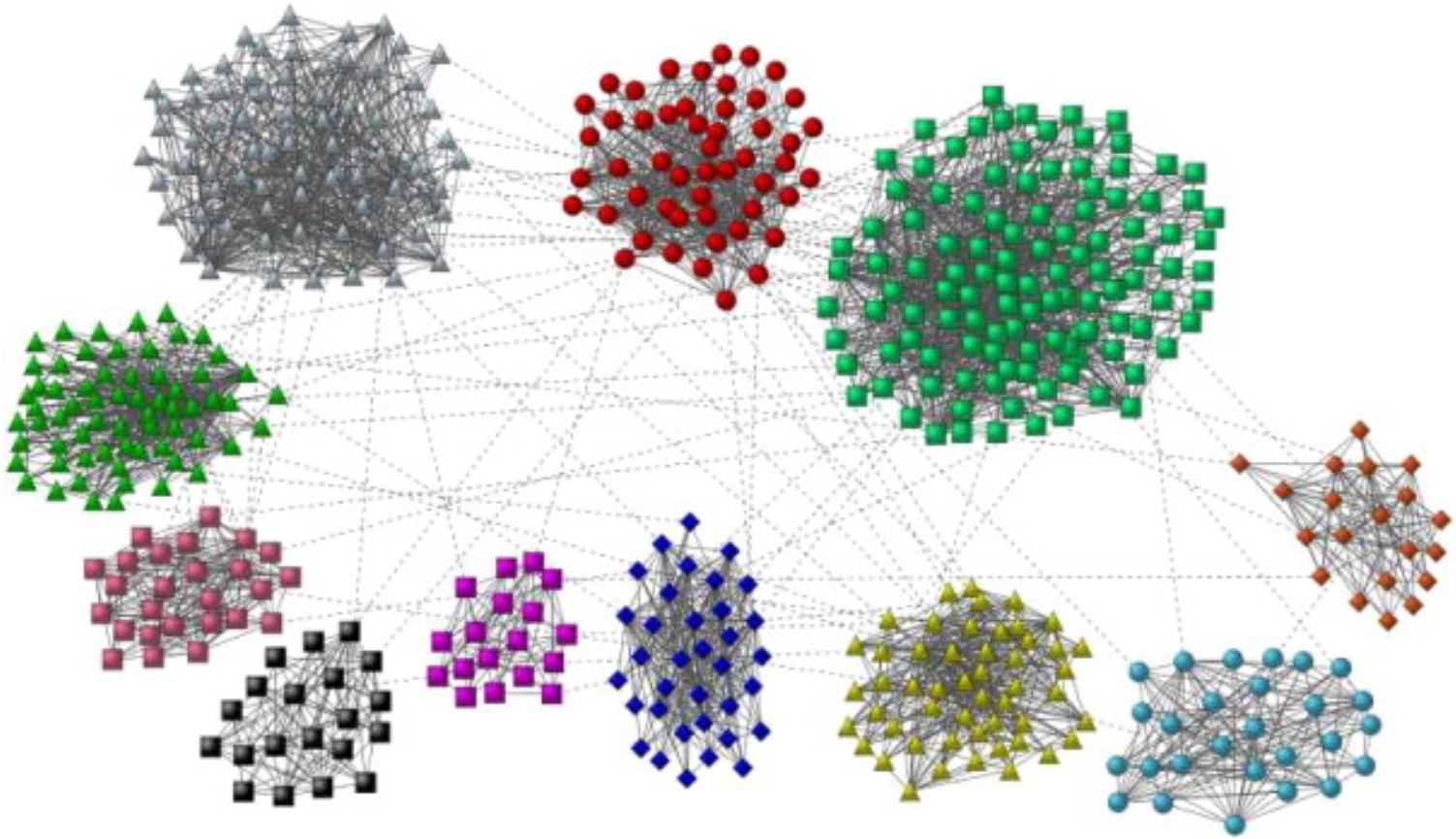


FIGURE 1.4. Left: *Representation of the* NCI60 *gene expression data set in a two-dimensional space,* Z_1 *and* Z_2*. Each point corresponds to one of the* 64 *cell lines. There appear to be four groups of cell lines, which we have represented using different colors.* Right: *Same as left panel except that we have represented each of the* 14 *different types of cancer using a different colored symbol. Cell lines corresponding to the same cancer type tend to be nearby in the two-dimensional space.*

Community detection



Outline

Regression

Model:

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

where $f : \mathbb{R}^p \to \mathbb{R}$

Given data: $\{(X_i, Y_i)\}_{i=1,\ldots,n}$

Estimator:

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

Theory: study $\|f - \hat{f}\|$

Some terminology

1. Supervised versus Unsupervised Learning
   - **Supervised:** For each data X_i, there is an associated response or label y_i.
   - **Unsupervised:** The X_i's do not have associated responses or labels.
   - **Semi-supervised:**
2. Regression versus Classification

Example

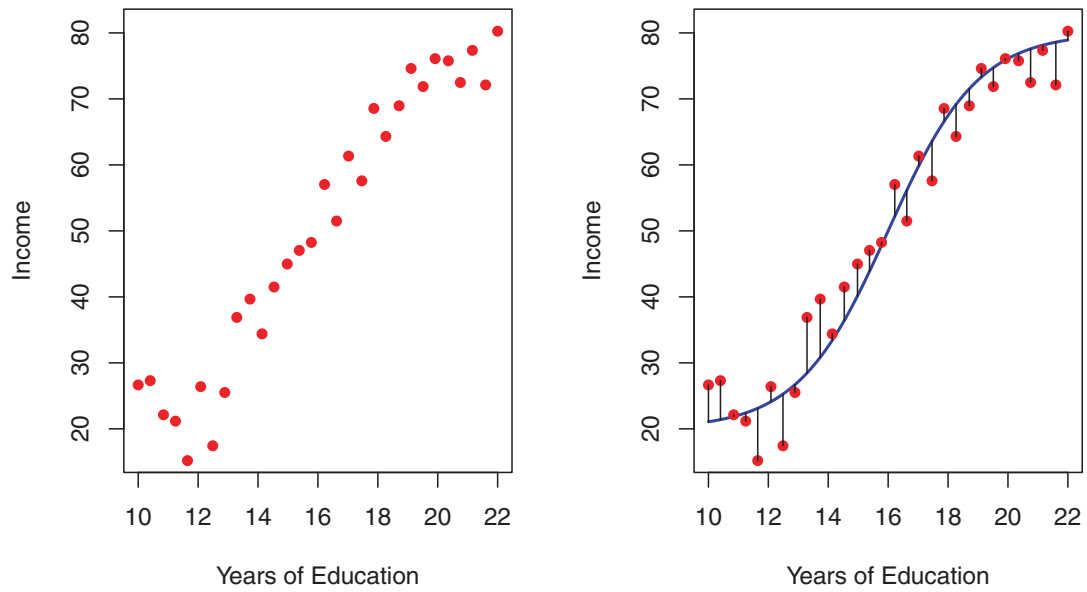


FIGURE 2.2. *The* `Income` *data set.* Left: *The red dots are the observed values of* `income` *(in tens of thousands of dollars) and* `years of education` *for* 30 *individuals.* Right: *The blue curve represents the true underlying relationship between* `income` *and* `years of education`*, which is generally unknown (but is known in this case because the data were simulated). The black lines represent the error associated with each observation. Note 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.*

Example

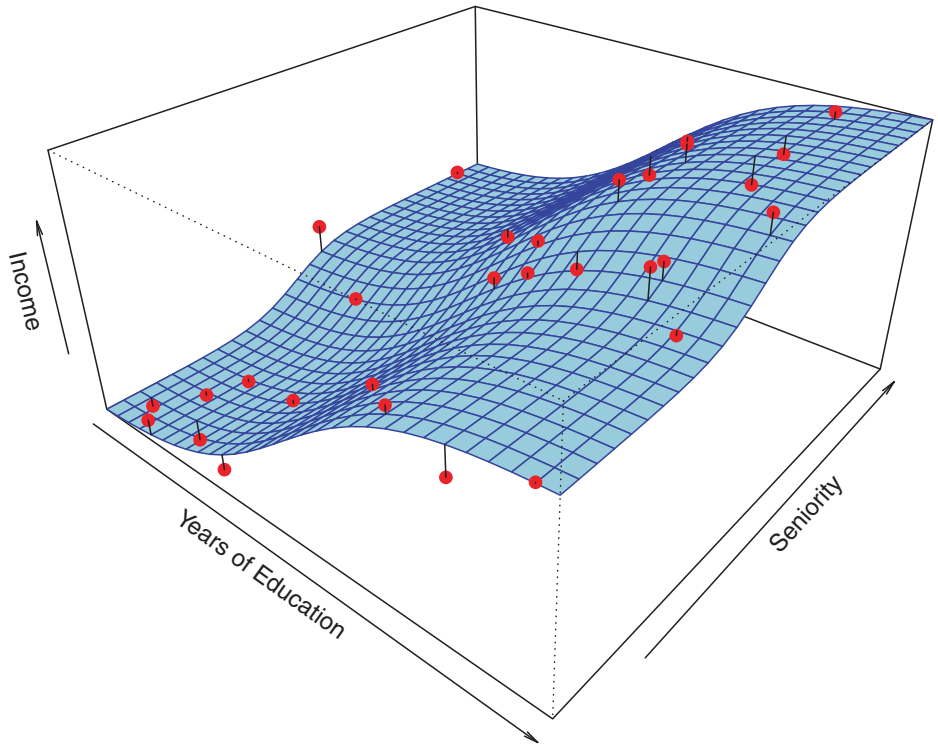


FIGURE 2.3. *The plot displays* `income` *as a function of* `years of education` *and* `seniority` *in the* `Income` *data set. The blue surface represents the true underlying relationship between* `income` *and* `years of education` *and* `seniority`*, which is known since the data are simulated. The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

Interesting questions

Write $X = (X_1, \ldots, X_p)^T$. Then $f(X) = f(X_1, \ldots, X_p)$.

- Which X_i is associated with the response?
- What is the relation between $f(X)$ and each X_i?
- Can the relation between $f(X)$ and X be adequately summarized using a linear equation, or is the relation more complicated?

How to estimate f?

Parametric methods – linear regression:

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

Find $\{\beta_0, \beta_1, \ldots, \beta_p\}$ such that

$$Y_i \approx \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}$$

More generally: $f(X)$ is a linear combination of L basis functions $\{\phi_l(X)\}_{l=1,\ldots,L}$:

$$f(x) = \beta_1 \phi_1(X) + \ldots + \beta_L \phi_L(X)$$

Find $\{\beta_1, \ldots, \beta_L\}$ such that

$$Y_i \approx \beta_1 \phi_1(X_i) + \ldots + \beta_L \phi_L(X_i)$$

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

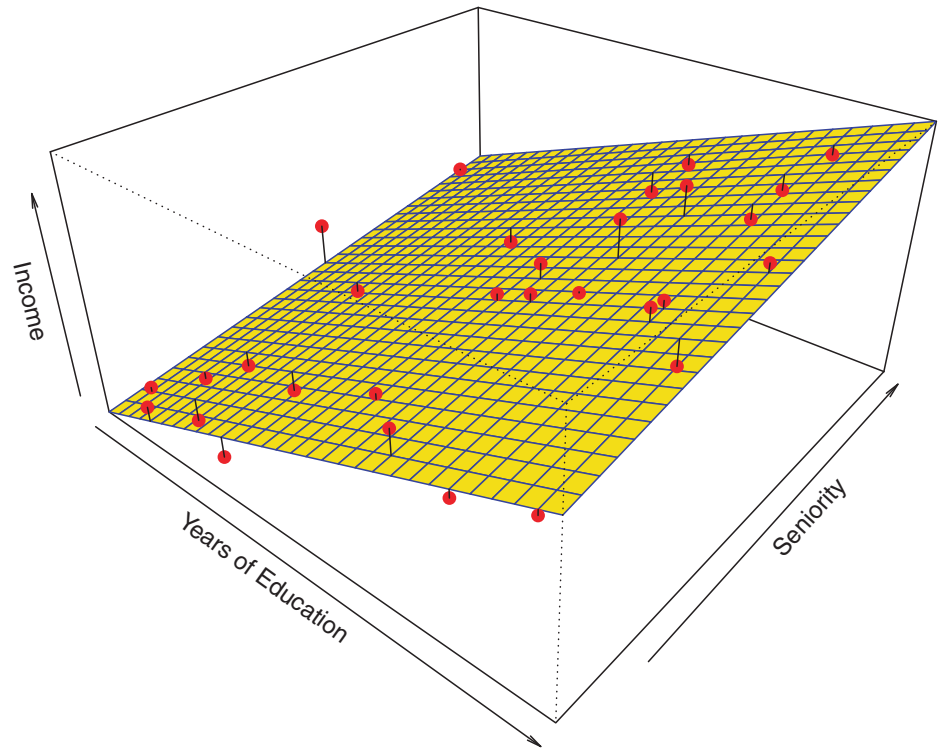


FIGURE 2.4. *A linear model fit by least squares to the* `Income` *data from Figure 2.3. The observations are shown in red, and the yellow plane indicates the least squares fit to the data.*

Non-parametric methods

- Nearest neighbors
- Kernel regression
- Local linear/polynomial regression
- Partitioning estimates

Error measurements

Training data: $\{(X_i, Y_i)\}_{i=1,\ldots,n_{\text{train}}} \rightarrow$ estimator $\hat{f}$

$$\text{Training Error} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \left(Y_i - \hat{f}(X_i) \right)^2$$

A test data point: (X_0, Y_0)

$$\text{Prediction error } = \left| Y_0 - \hat{f}(X_0) \right|$$

Test data set: $\{(X_j, Y_j)\}_{j=1,\ldots,n_{\text{test}}}$

$$\text{Mean Squared Error } (MSE) = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} \left(Y_j - \hat{f}(X_j) \right)^2$$

The classification setting

Training data: $\{(X_i, Y_i)\}_{i=1,\ldots,n_{\text{train}}}$ where Y_i are qualitative

$$\text{Training Error} = \frac{1}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} I(Y_i \neq \hat{Y}_i)$$

A test data point: (X_0, Y_0)

$$\text{Prediction error } = I(Y_0 \neq \hat{Y}_0)$$

Test data set: $\{(X_j, Y_j)\}_{j=1,\ldots,n_{\text{test}}}$

$$\text{Mean Squared Error } (MSE) = \frac{1}{n_{\text{test}}} \sum_{j=1}^{n_{\text{test}}} I(Y_j \neq \hat{Y}_j)$$

Model complexity/flexibility

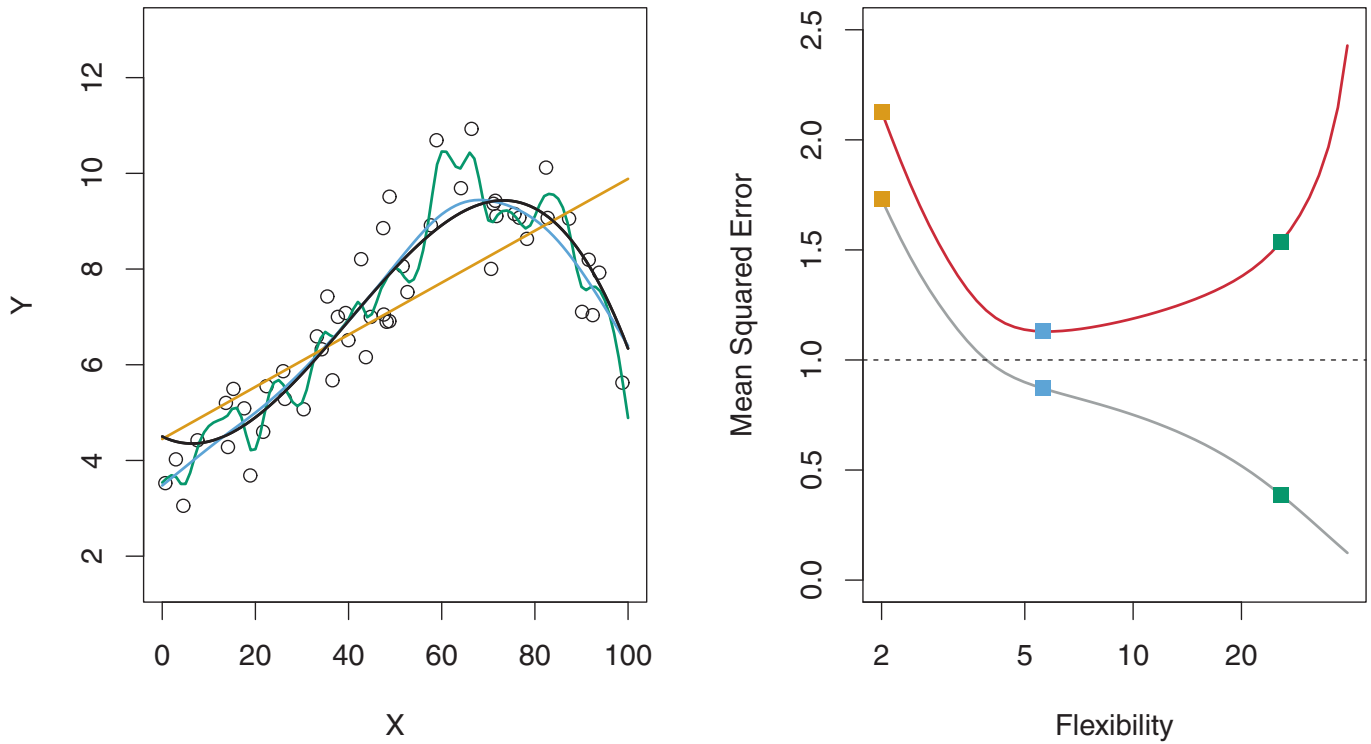


FIGURE 2.9. Left: *Data simulated from f, shown in black. Three estimates of f are shown: the linear regression line (orange curve), and two smoothing spline fits (blue and green curves).* Right: *Training MSE (grey curve), test MSE (red curve), and minimum possible test MSE over all methods (dashed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

Model complexity/flexibility

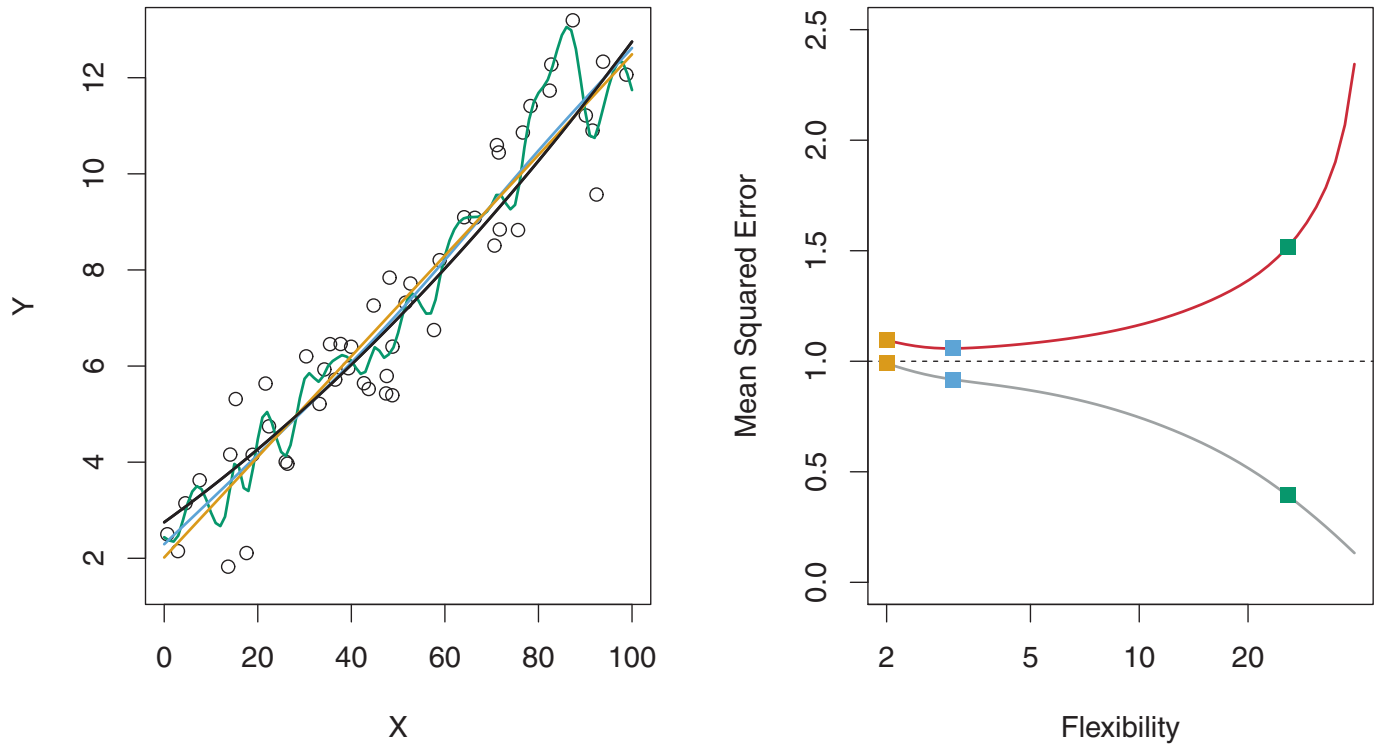


FIGURE 2.10. *Details are as in Figure 2.9, using a different true f that is much closer to linear. In this setting, linear regression provides a very good fit to the data.*

How to choose a good parameter? – Cross validation

K-Nearest Neighbors (KNN)

Training data: $\{(X_i, Y_i)\}_{i=1,\ldots,n}$

Regression: at X_0,

$$\hat{Y}_0 = \frac{1}{K} \sum_{i \in \mathcal{N}_0} Y_i$$

where $\mathcal{N}_0$ contains the K points in the training data that are closest to X_0

Classification: Assume $Y \in \{1, 2, \ldots\}$. At the test obervation X_0,

$$\mathbb{P}(Y = j | X = X_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(Y_i = j).$$

KNN classifies X_0 to the class with the largest probability.

KNN Example

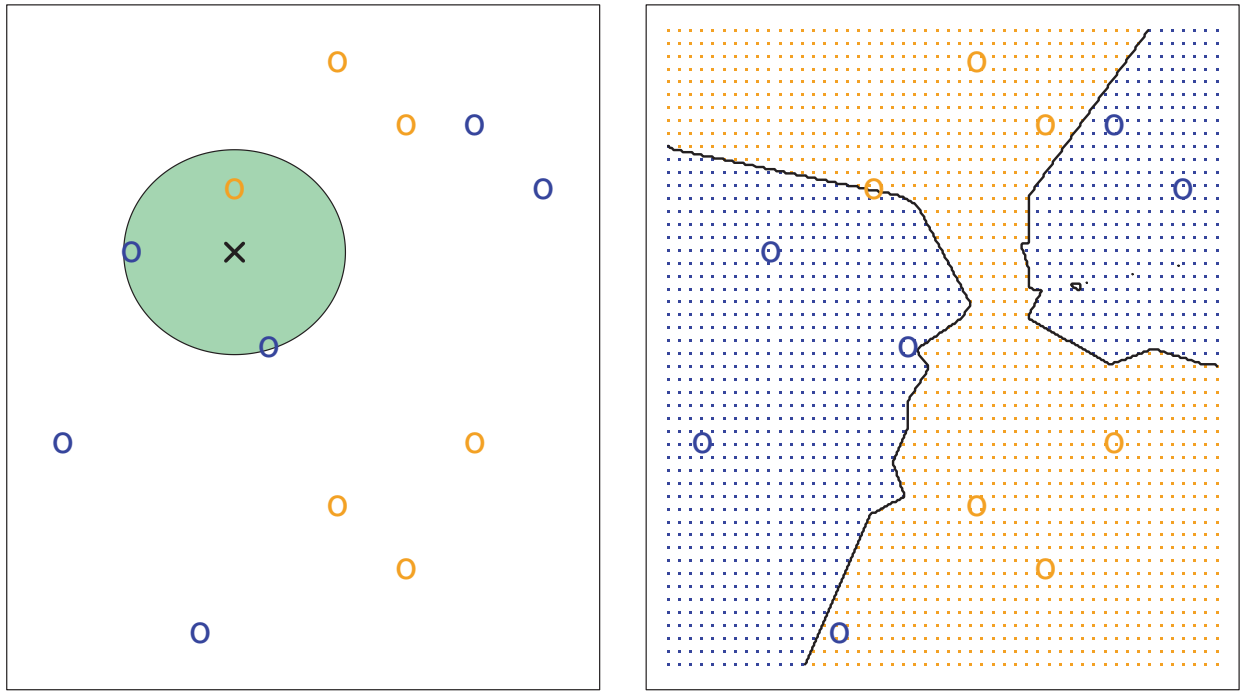


FIGURE 2.14. *The KNN approach, using $K = 3$, is illustrated in a simple situation with six blue observations and six orange observations.* Left: *a test observation at which a predicted class label is desired is shown as a black cross. The three closest points to the test observation are identified,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

Too small or too large K

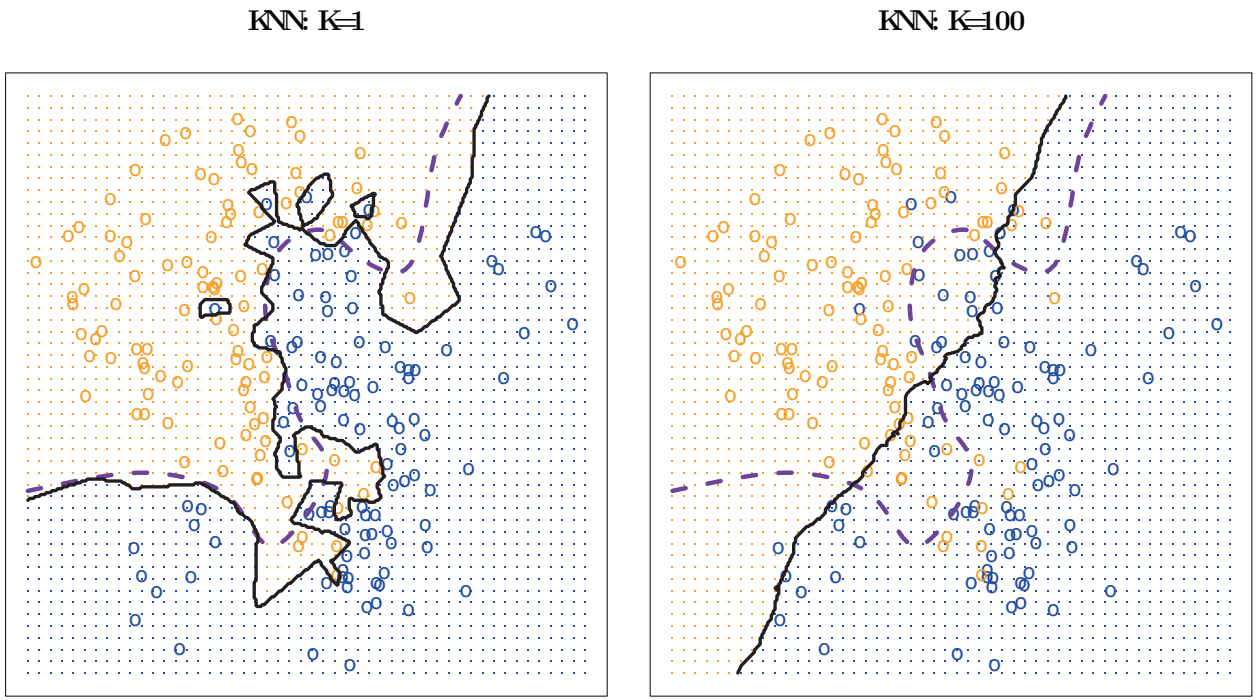


FIGURE 2.16. *A comparison of the KNN decision boundaries (solid black curves) obtained using $K = 1$ and $K = 100$ on the data from Figure 2.13. With $K = 1$, the decision boundary is overly flexible, while with $K = 100$ it is not sufficiently flexible. The Bayes decision boundary is shown as a purple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

- $K = 1$: training error 0, test error 0.1695
- $K = 100$: test error 0.1925

Proper $K = 10$

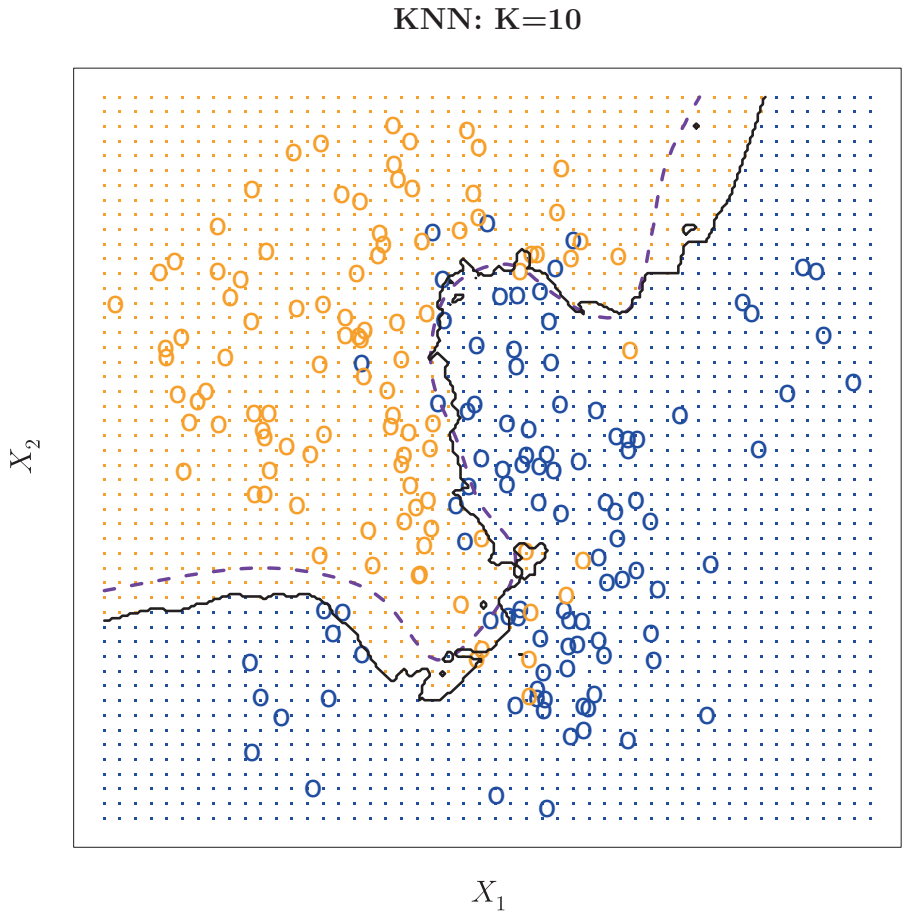


FIGURE 2.15. *The black curve indicates the KNN decision boundary on the data from Figure 2.13, using $K = 10$. The Bayes decision boundary is shown as a purple dashed line. The KNN and Bayes decision boundaries are very similar.*

It is important to choose a proper K

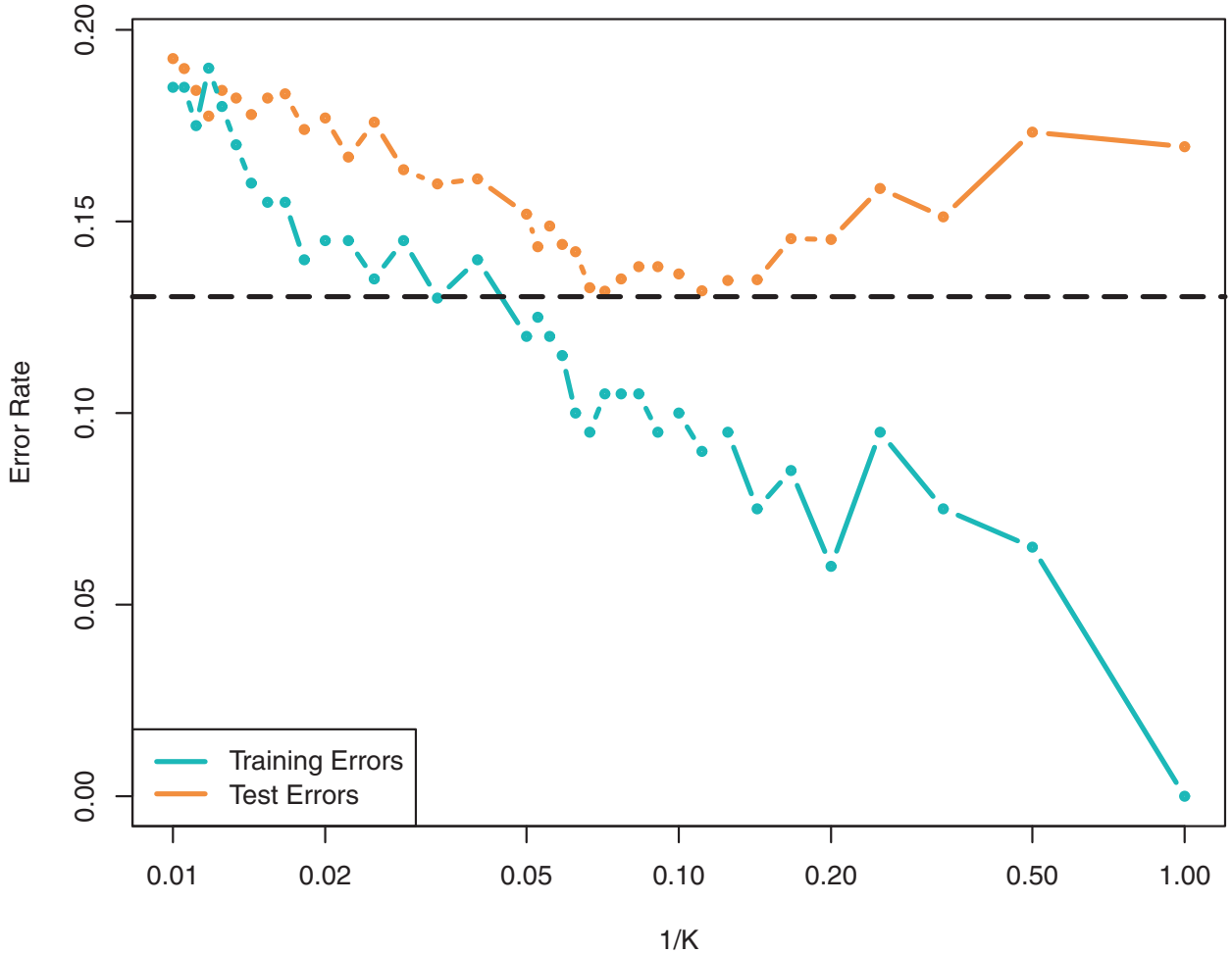


FIGURE 2.17. *The KNN training error rate (blue, 200 observations) and test error rate (orange, 5,000 observations) on the data from Figure 2.13, as the level of flexibility (assessed using $1/K$) increases, or equivalently as the number of neighbors K decreases. The black dashed line indicates the Bayes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