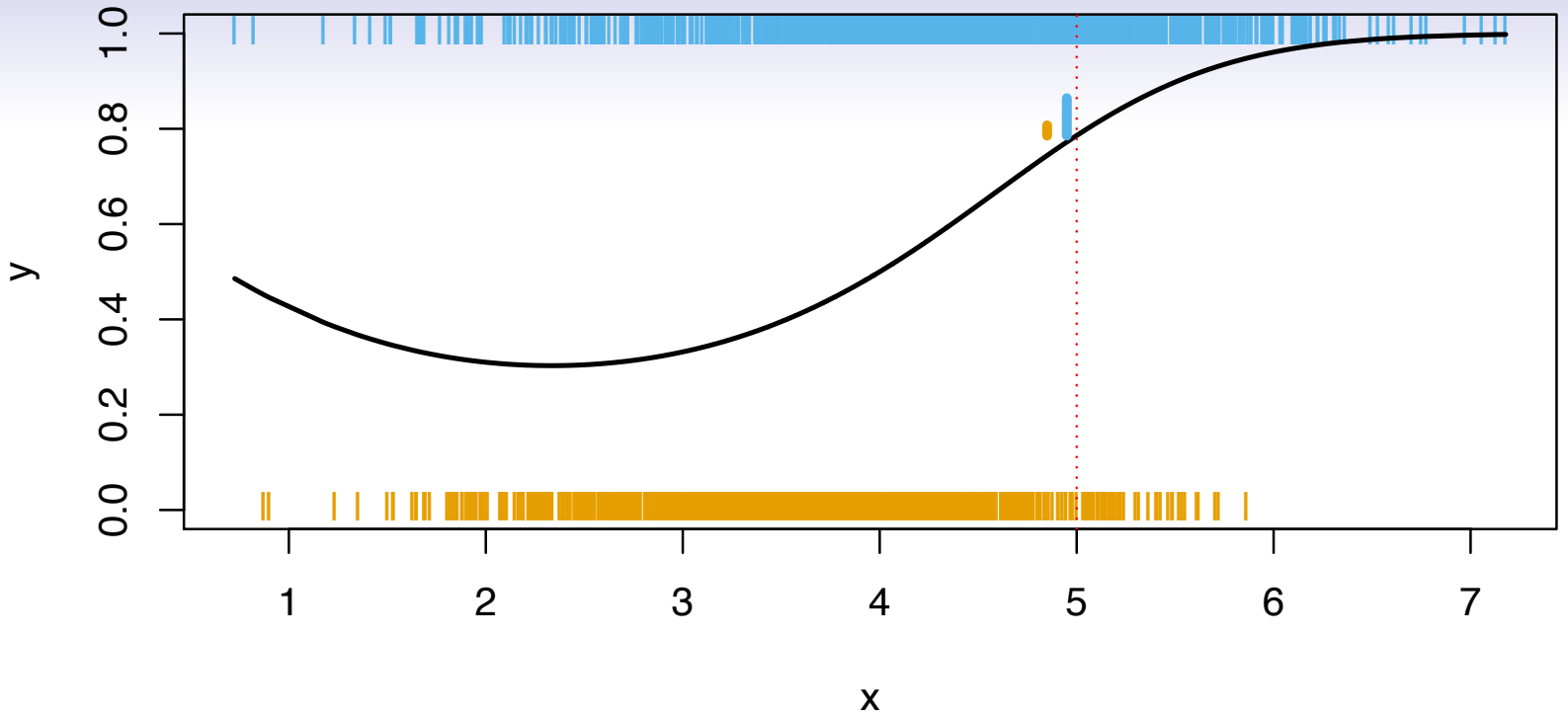
Classification Problems

Here the response variable Y is *qualitative* — e.g. email is one of $\mathcal{C} = (\texttt{spam}, \texttt{ham})$ ($\texttt{ham}$=good email), digit class is one of $\mathcal{C} = \{\texttt{0}, \texttt{1}, \ldots, \texttt{9}\}$. Our goals are to:

- Build a classifier $C(X)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation X.
- Assess the uncertainty in each classification
- Understand the roles of the different predictors among $X = (X_1, X_2, \ldots, X_p)$.

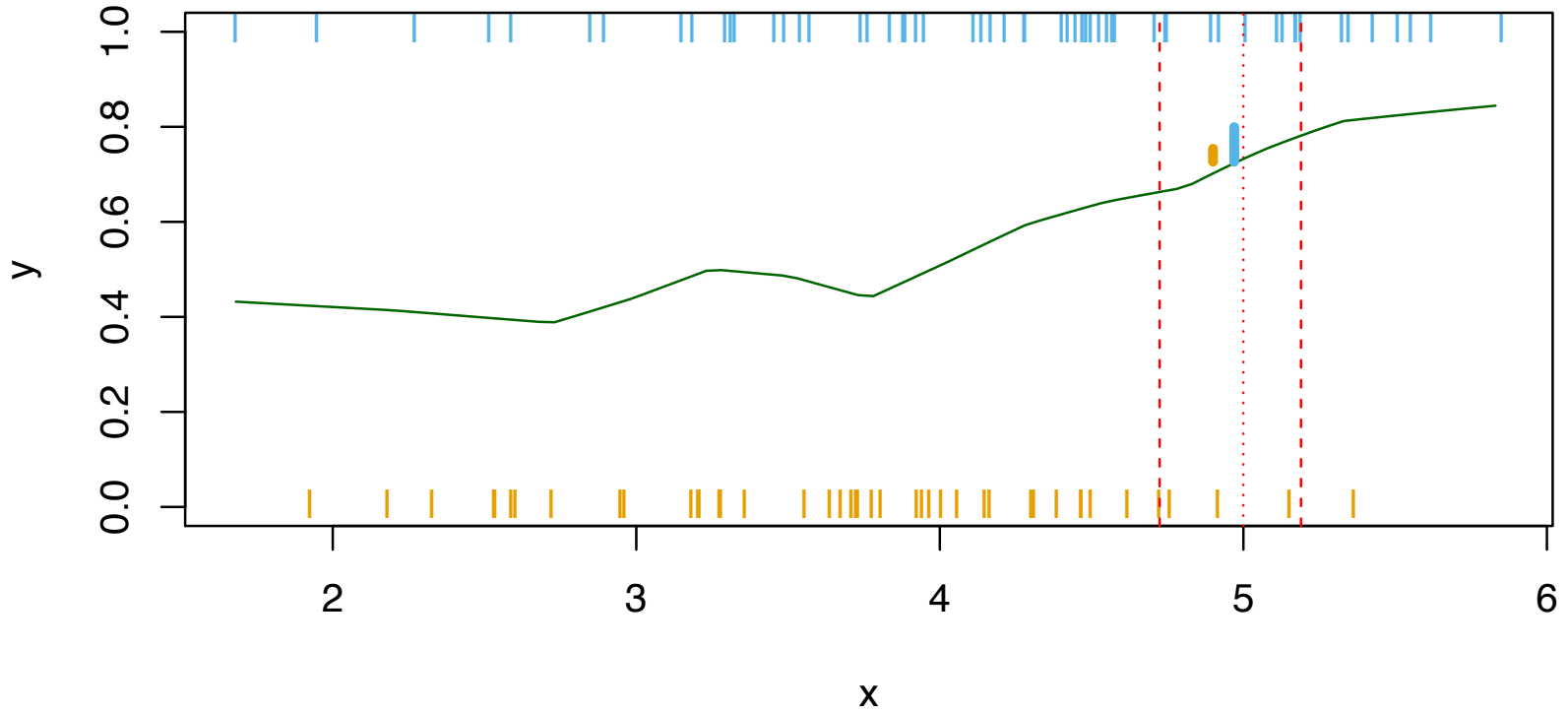


Is there an ideal $C(X)$? Suppose the K elements in $\mathcal{C}$ are numbered $1, 2, \ldots, K$. Let

$$p_k(x) = \Pr(Y = k|X = x), \ k = 1, 2, \ldots, K.$$

These are the *conditional class probabilities* at x; e.g. see little barplot at $x = 5$. Then the *Bayes optimal* classifier at x is

$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), p_2(x), \ldots, p_K(x)\}$$



Nearest-neighbor averaging can be used as before.

Also breaks down as dimension grows. However, the impact on $\hat{C}(x)$ is less than on $\hat{p}_k(x),\ k = 1, \ldots, K$.

Classification: some details

- Typically we measure the performance of $\hat{C}(x)$ using the misclassification error rate:

$$\text{Err}_{\mathsf{Te}} = \text{Ave}_{i \in \mathsf{Te}} I[y_i \neq \hat{C}(x_i)]$$

- The Bayes classifier (using the true $p_k(x)$) has smallest error (in the population).

Classification: some details

- Typically we measure the performance of $\hat{C}(x)$ using the misclassification error rate:

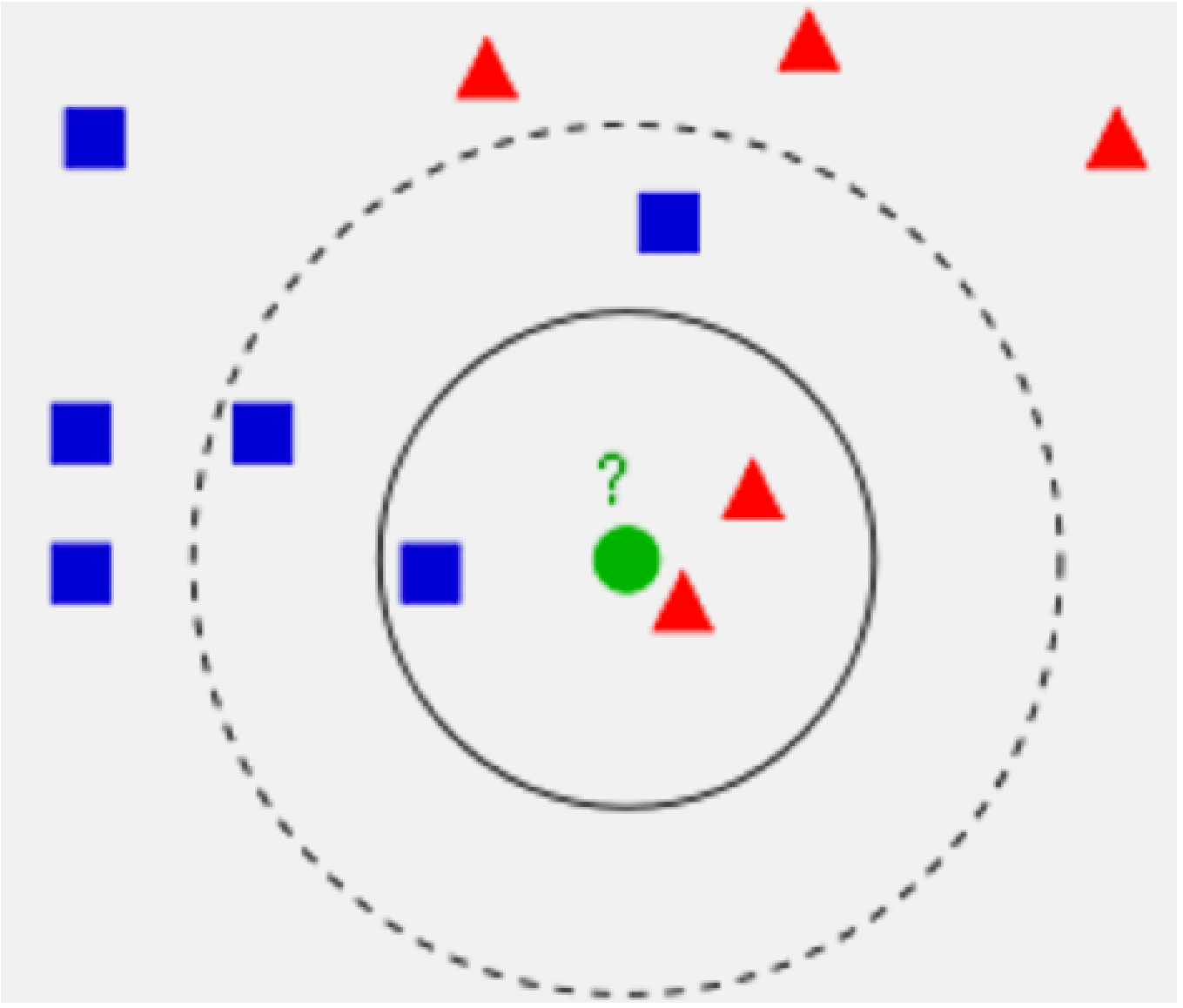
$$\text{Err}_{\mathsf{Te}} = \text{Ave}_{i \in \mathsf{Te}} I[y_i \neq \hat{C}(x_i)]$$

- The Bayes classifier (using the true $p_k(x)$) has smallest error (in the population).

- Support-vector machines build structured models for $C(x)$.
- We will also build structured models for representing the $p_k(x)$. e.g. Logistic regression, generalized additive models.

KNN for classification

KNN for classification



KNN for classification

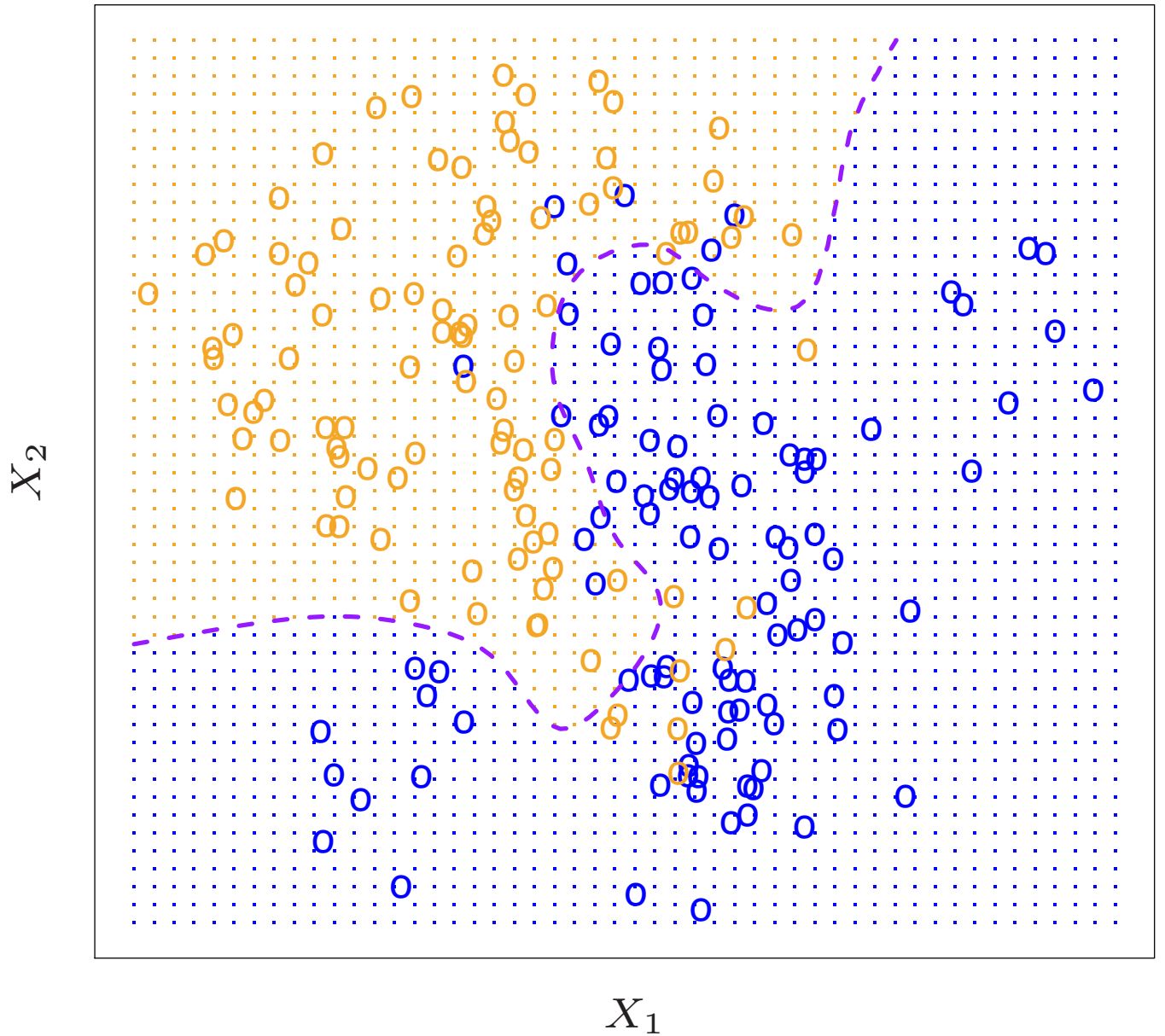
向量X1 耐酸时间（秒）	向量X2 压强(公斤/平方米)	品质Y
7	7	坏
7	4	坏
3	4	好
1	4	好

KNN for classification

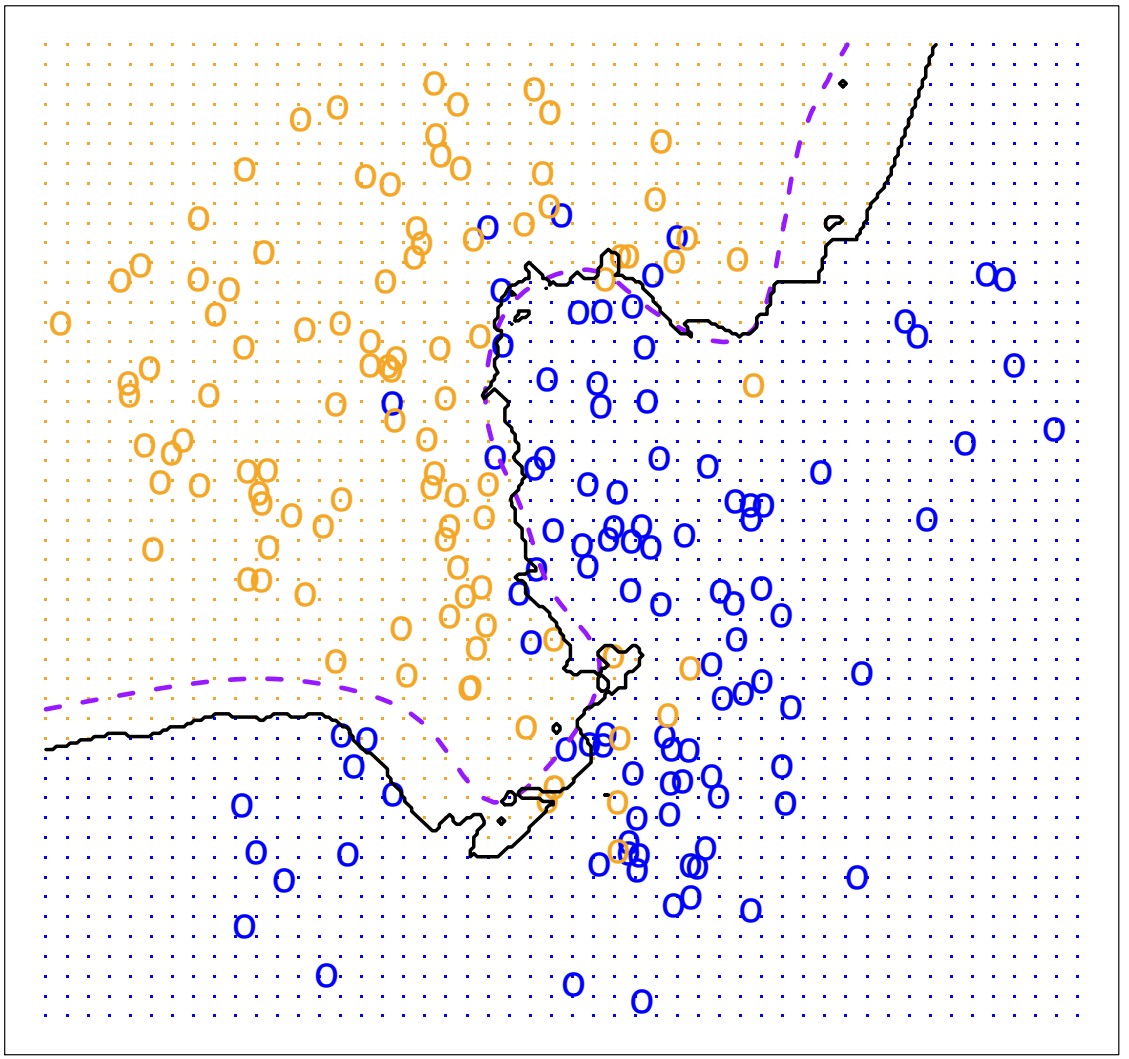
K=3 X0=(3,7)

向量X1 耐酸时间（秒）	向量X2 压强(公斤/平方米)	计算到 (3, 7)的距离	向量Y
7	7	$(7-3)^2+(7-7)^2=16$	坏
7	4	$(7-3)^2+(4-7)^2=25$	N/A
3	4	$(3-3)^2+(4-7)^2=9$	好
1	4	$(1-3)^2+(4-7)^2=13$	好

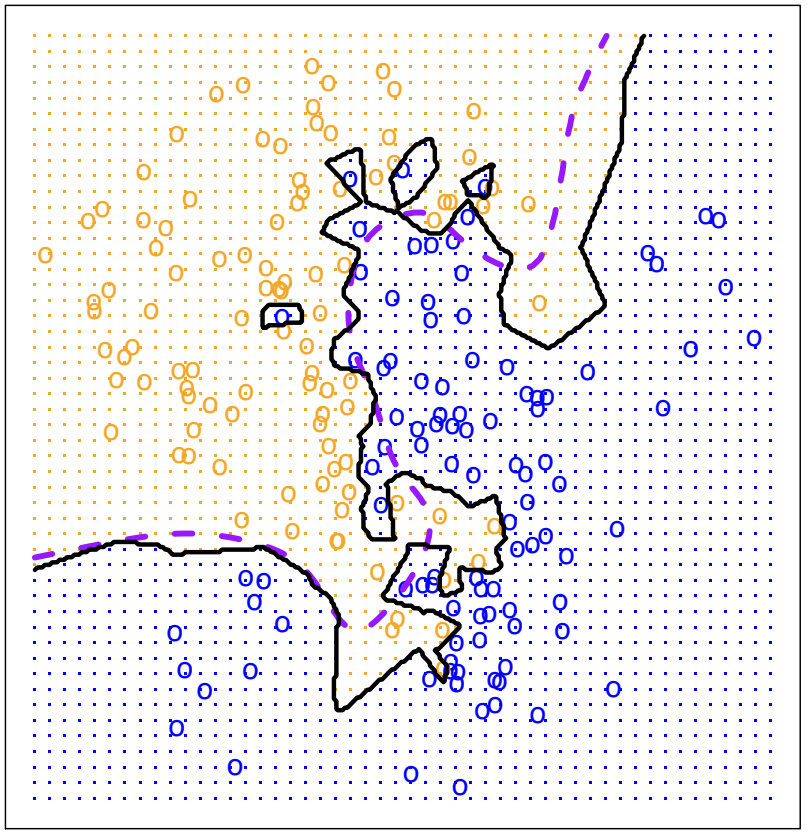
Example: K-nearest neighbors in two dimensions



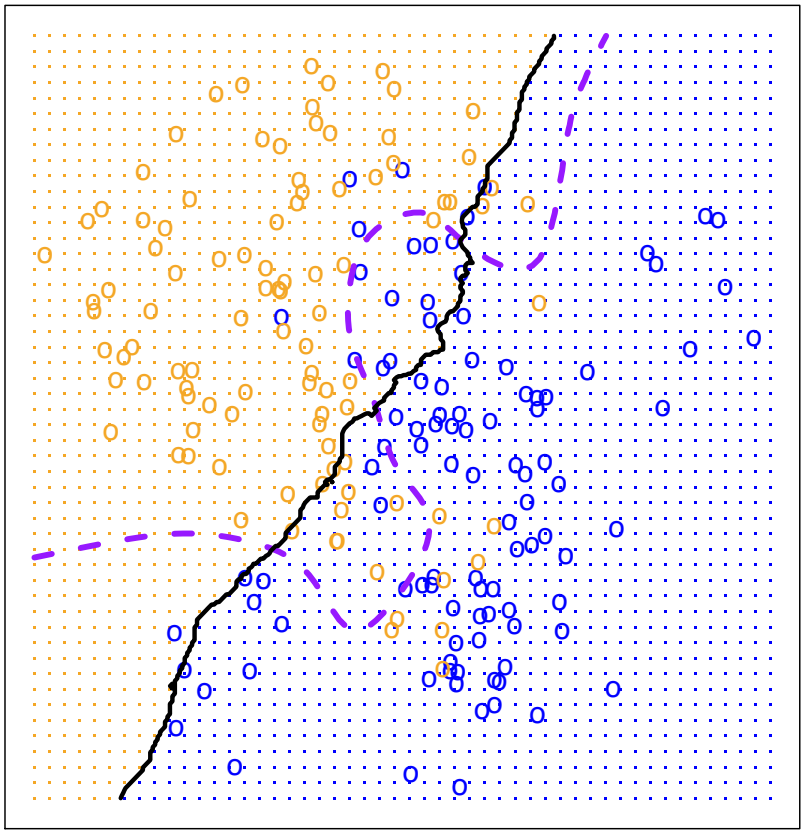
KNN: K=10

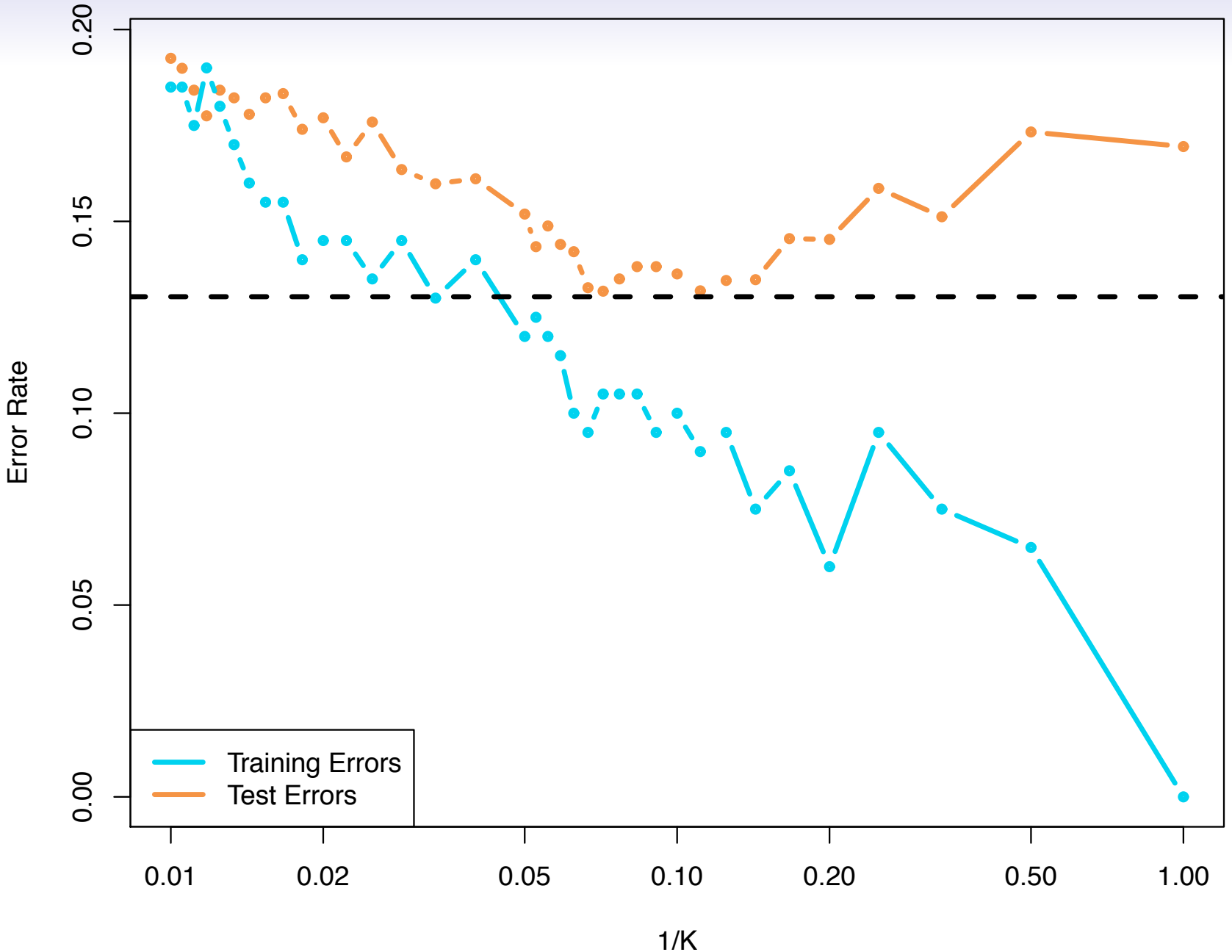


KNN: K=1



KNN: K=100



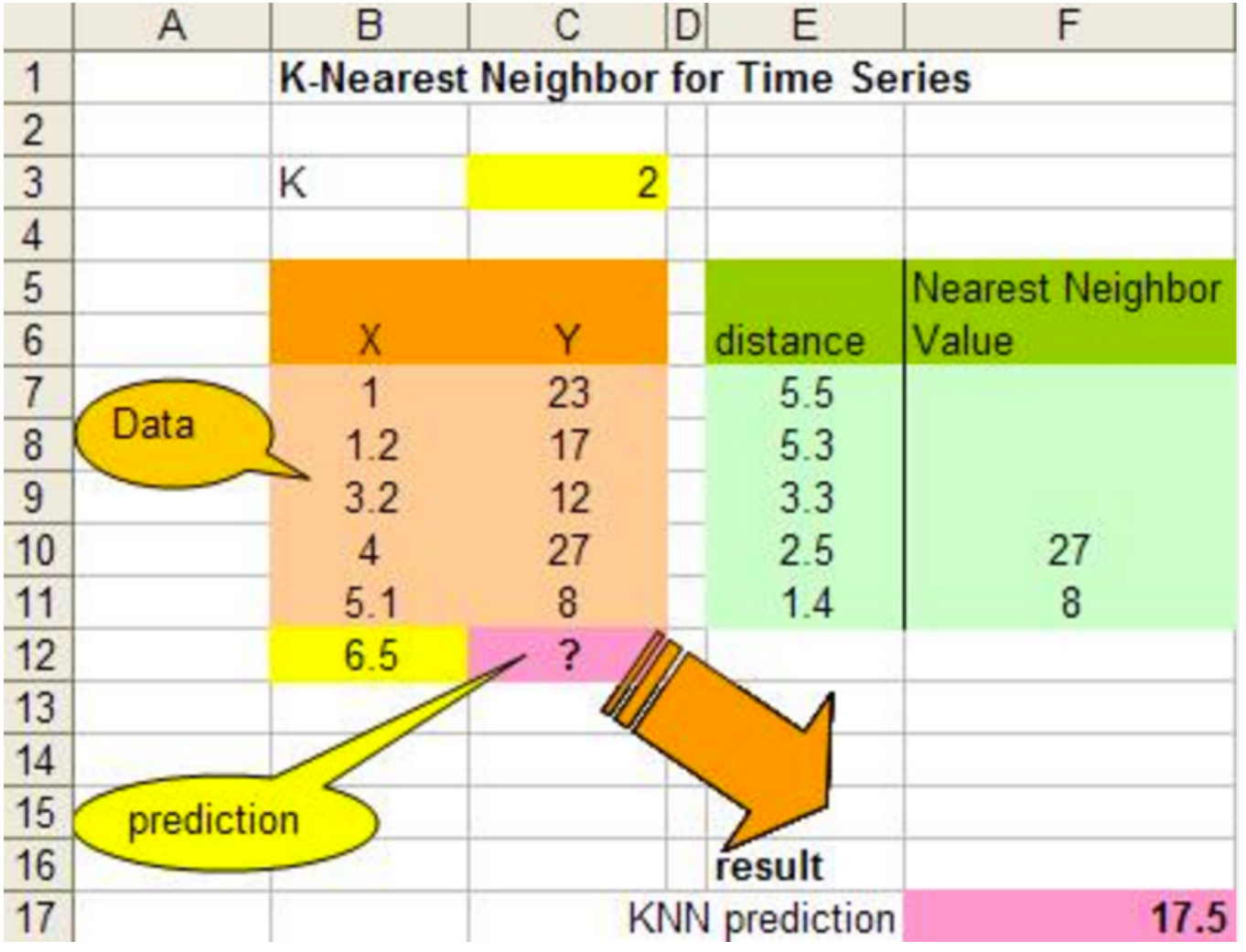

Error Rate
0.00
0.05
0.10
0.15
0.20
0.01
0.02
0.05
0.10
0.20
0.50
1.00
1/K
Training Errors
Test Errors

Linear regression VS KNN regression

- linear regression (parametric method) has some advantages: easy to fit, easy to interpret; it also has some advantage: strong assumption about the form of $f(X)$, specification error?
- Nonparametric methods don't assume a parametric form for $f(X)$.
- The simplest nonparametric method, K-nearest neighbors regression (KNN regression).
- Given a value for K and a prediction point x_0, KNN regression first identifies the K training observations that are closest to x_0, represented by $\mathcal{N}_0$.

$$\hat{f}(x_0) = \frac{1}{K} \sum_{x_i \in \mathcal{N}_0} y_i$$

KNN for regression



Linear regression VS KNN regression

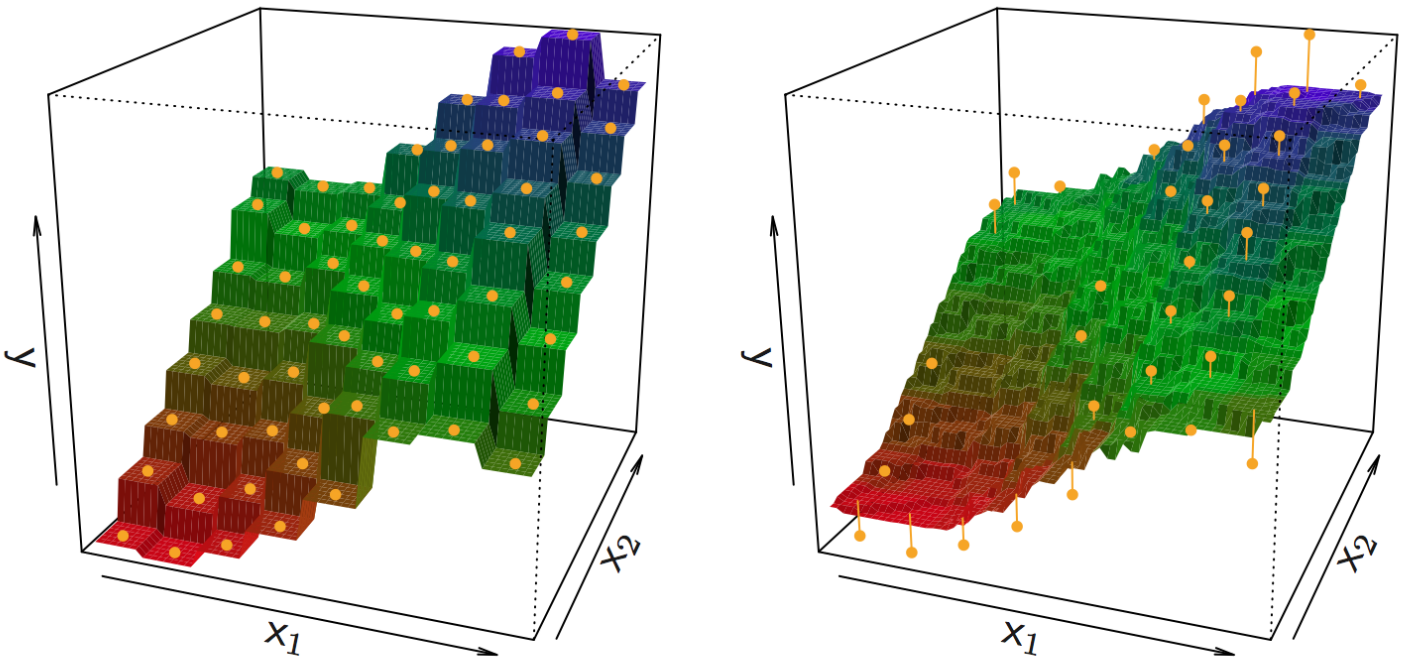


FIGURE 3.16. *Plots of $\hat{f}(X)$ using KNN regression on a two-dimensional data set with 64 observations (orange dots).* Left: $K = 1$ *results in a rough step function fit.* Right: $K = 9$ *produces a much smoother fit.*

Linear regression VS KNN regression

- The parametric approach will outperform the non-parametric approach if the parametric form that has been selected is close to the true form of f.

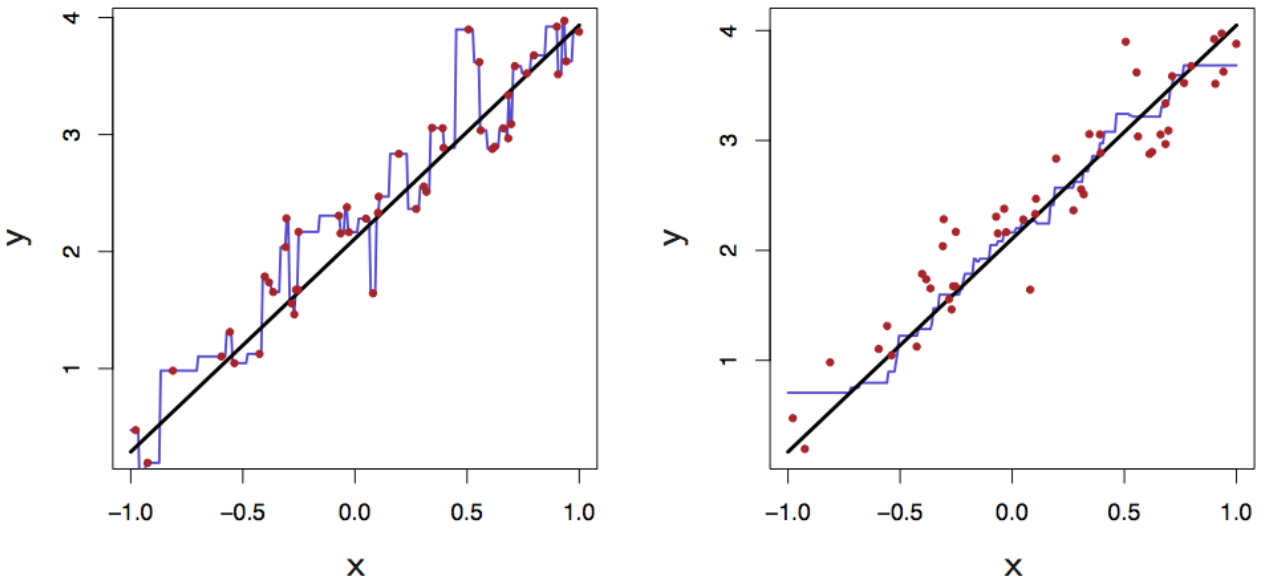


FIGURE 3.17. *Plots of $\hat{f}(X)$ using KNN regression on a one-dimensional data set with 100 observations. The true relationship is given by the black solid line.* Left: *The blue curve corresponds to $K = 1$ and interpolates (i.e. passes directly through) the training data.* Right: *The blue curve corresponds to $K = 9$, and represents a smoother fit.*

Linear regression VS KNN regression

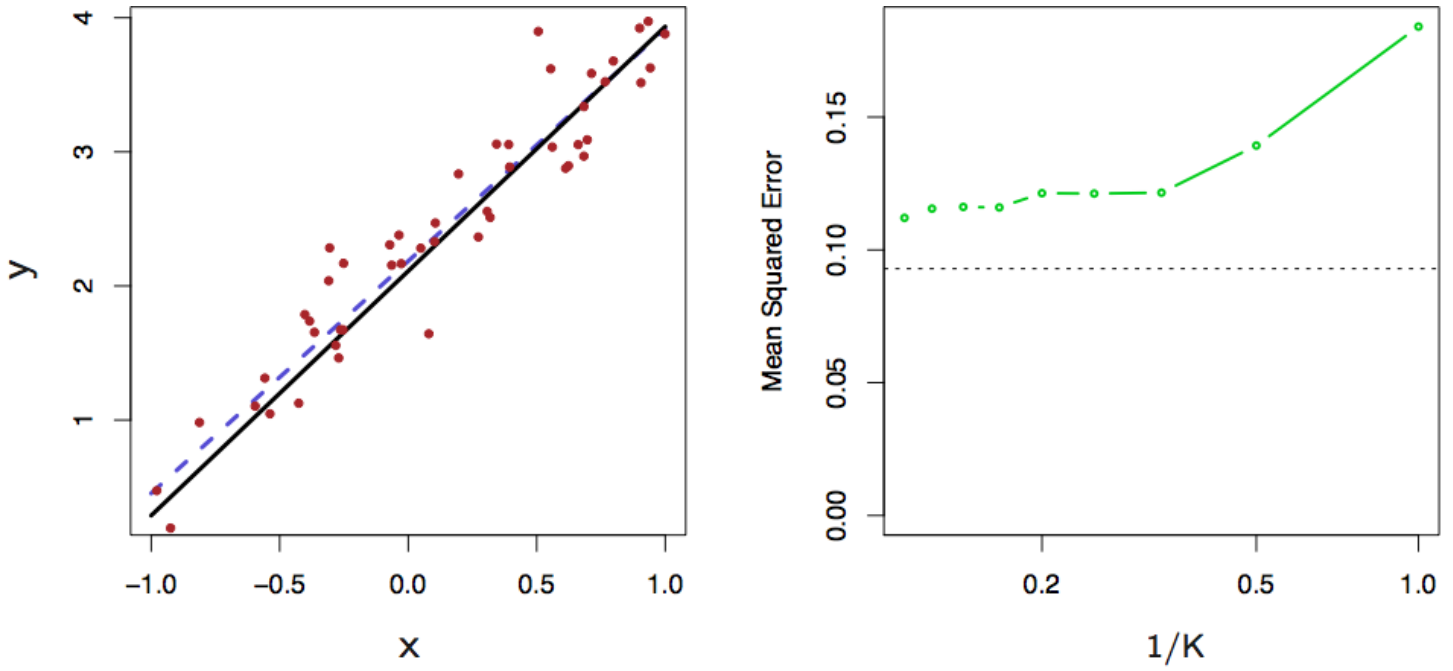
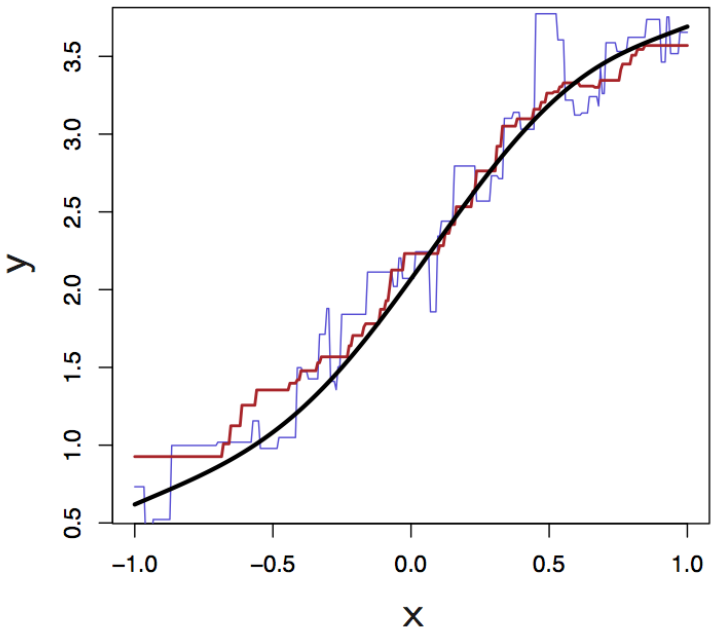
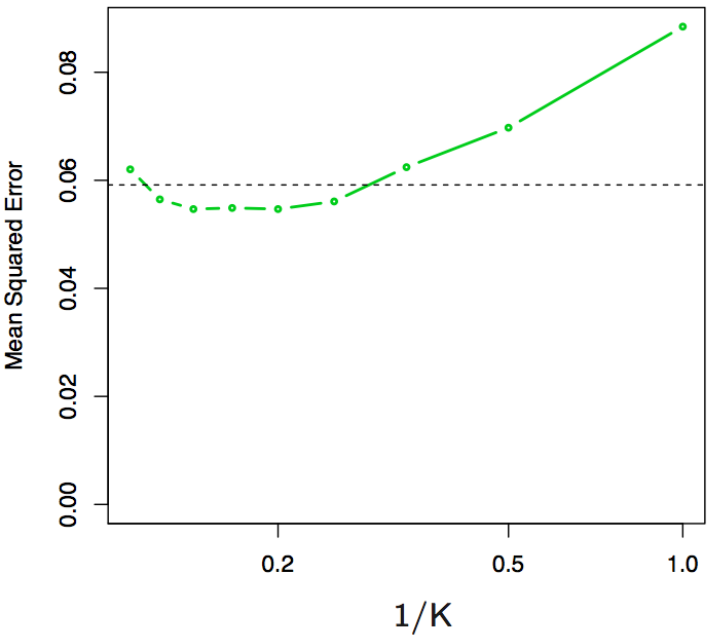


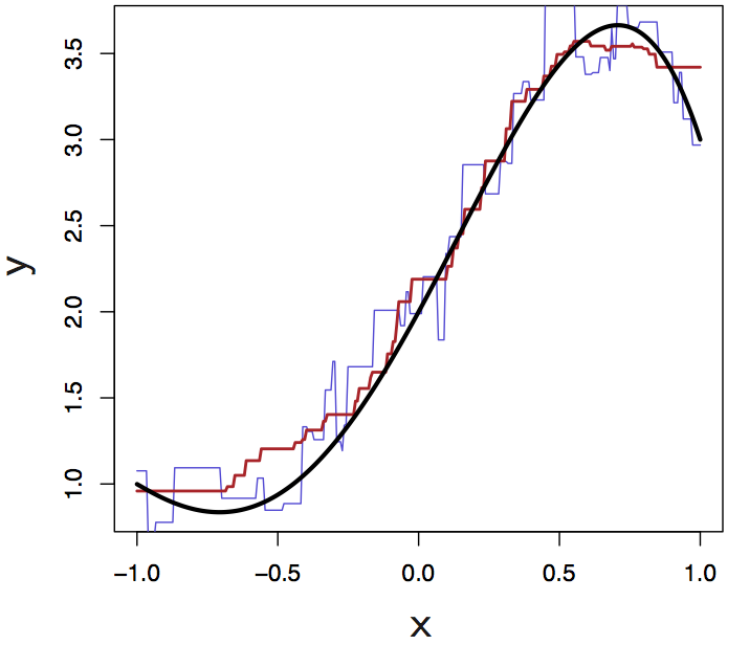
FIGURE 3.18. *The same data set shown in Figure 3.17 is investigated further.* Left: *The blue dashed line is the least squares fit to the data. Since* $f(X)$ *is in fact linear (displayed as the black line), the least squares regression line provides a very good estimate of* $f(X)$. Right: *The dashed horizontal line represents the least squares test set MSE, while the green solid line corresponds to the MSE for KNN as a function of* $1/K$ *(on the log scale). Linear regression achieves a lower test MSE than does KNN regression, since* $f(X)$ *is in fact linear. For KNN regression, the best results occur with a very large value of* K*, corresponding to a small value of* $1/K$.

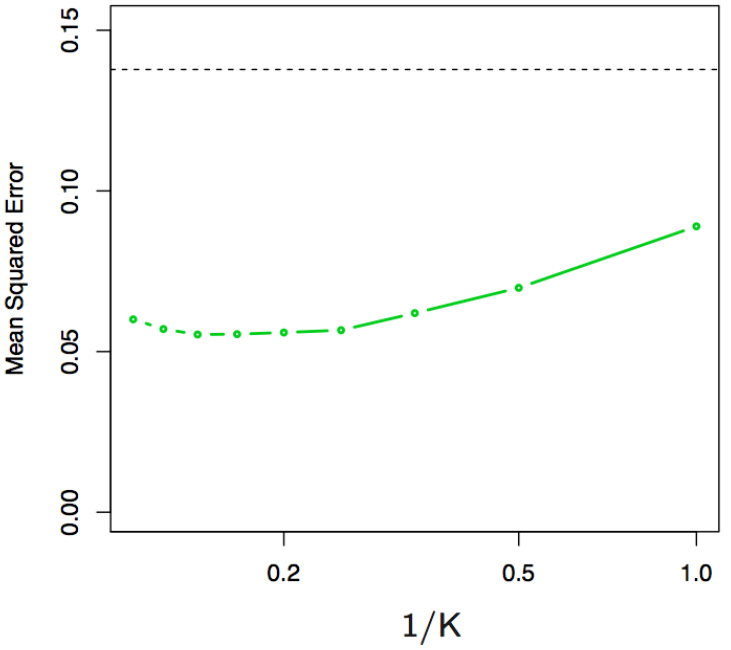
Linear regression VS KNN regression





Linear regression VS KNN regression





Linear regression VS KNN regression